Exploring Factors Associated with Voucher Program for Speech Language Therapy for the Preschoolers of Parents with Communication Disorder using Weighted Random Forests

Haewon Byeon[1]
Department of Speech Language Pathology, Honam University
Gwangju, Republic of Korea

Sulki Cha[2]
Dept. of Rehabilitation
Graduate School, Honam University
Gwangju, Republic of Korea

KyoungYuel Lim*[3]
Department of Speech Language Pathology, Honam University
Gwangju, Republic of Korea

***Abstract*—It is necessary to identify the demand level of consumers and recognize the support target priority based on it in order to provide efficient services with a limited budget. This study provided baseline data for spreading the use of consumer-oriented voucher service by exploring factors associated with the demand of the Voucher Program for Speech Language Therapy for preschool children. This study were analyzed 212 guardians living with children (≤5 years old) who resided in Seoul from Aug 11 to Oct 9, 2015. The outcome variable was defined as the demand (i.e., required and not required) of the Voucher Program for Speech Language Therapy. The results of the developed prediction model were compared with the results of a decision tree based on classification and regression tree (CART). The prediction performance of the developed model was evaluated using a confusion matrix. Among the 212 subjects, 112 (52.8%) responded that the Voucher Program for Speech Language Therapy was necessary. The weighted random forest-based model predicted five variables (i.e., whether preschooler caregiving services were used or not, economic activity after childbirth, the awareness of Seoul's welfare counselor operation, mean monthly living expenses, and whether welfare related information was obtained) as the variables associated with the demand of the Voucher Program for Speech Language Therapy and the accuracy was 72.1%. It is needed to develop systematic policies to expand consumer-oriented language therapy services based on the developed prediction model for the Voucher Program for Speech Language Therapy.**

Keywords*—Weighted random forests; CART; speech language therapy; prediction model; voucher program

I. Introduction

The number of people with disabilities is increasing in South Korea. Korean Act on Welfare of Persons with Disabilities divides disabilities into 15 types including physical disability and hearing impairment. As of 2017, the population of people with disabilities is estimated as 2,660,000 people, which is an increase of more than 20% compared to 2005 (2,140,000 people) [1]. The most common disability is a physical disability, followed by visual impairment, hearing impairment, low intelligence, and autism [2]. Among them, hearing impairment, low intelligence, and autism affect language development to lead communication problems [3].

When people with communication disorders get married and have a child, the language development of the child is more likely to be delayed even if the child does not have a disability [4]. It is because the communication restriction of the parents with language impairment limits the language development support needed for the child. Therefore, Korean Ministry of Health and Welfare has implemented "the Voucher Program for Speech Language Therapy" for children without disabilities raised by parents with disabilities since 2009 in order to support the successful language development of children and strengthen the competence of families with disabilities [5]. The key components of the program are to provide language rehabilitation services such as language development and aural development to children (<12 years old) without disabilities and having at least one parent with a disability such as hearing impairment or language impairment. It is supported in the form of a voucher. The target amount is 220,000 KRW per month as of 2015, and there may be copayment according to the income of the target family based on the mean nationwide household income.

A voucher means a subsidy that is provided for somebody and it limits purchasing power by allowing a user to select goods or services within a limited range [6]. It is defined as "tied demand side subsidy" [5]. Although the South Korean government actively supports the voucher system, the Voucher Program for Speech Language Therapy, providing vouchers for users who have a desire for the service instead of supporting social welfare service institutes, experiences restrictions on access to services (e.g., reduced project support personnel) due to budgetary deficit [7,8].

It is necessary to identify the demand level of consumers and recognize the support target priority based on it in order to provide efficient services with a limited budget. Nevertheless, previous studies on the Voucher Program for Speech Language Therapy have mainly focused on how to understand voucher, a new policy, and user satisfaction of it [9,10,11,12]. Moreover, their research methods mostly aimed to conduct a factual survey and identify the characteristics of consumers [8]. The

*Corresponding Authors

fact that previous studies have rarely discussed voucher with taking into account the unique characteristics of social welfare services indicates that there is insufficient baseline data, which is needed to run a voucher system effectively, and necessary discussions have not been made yet. As far as we are aware of, there is no study evaluating the factors affecting the demand of the Voucher Program for Speech Language Therapy with considering socio-demographic factors, knowledge of voucher services, methods of obtaining welfare-related information, and education policy satisfaction using data mining techniques.

It is necessary to analyze the demand level of consumers in-depth in order to operate "the Voucher Program for Speech Language Therapy" efficiently and the analysis results will be useful to suggest ways to improve the special education service support in the future. This study provided baseline data for spreading the use of consumer-oriented voucher service by exploring factors associated with the demand of the Voucher Program for Speech Language Therapy for preschool children. The composition of this study is as follows. Chapter 2 will explain the algorithm and model development procedure of weighted random forests as well as study subjects and variable measurement. Chapter 3 will compare the results of the developed prediction model with those of CART model. Lastly, chapter 4 will present conclusions and future research directions.

II. Methods and Materials

A. Target Subjects

This study selected children (≤5 years old) and guardians residing with them who conducted Seoul Welfare Study, which targeted the local population who resided in Seoul from Aug 11 to Oct 9, 2015. The population of this study was households living in Seoul at the time of census among target households of Statistics Korea's 2010 Population and Housing Census (complete enumeration). Systematic sampling was used. Three thousand household were planned to be sampled so that the maximum limit of error could be approximately 1.8% at the 95% confidence level. Sample households per plot were ten households and 300 sampling plots were selected. The computer-assisted personal interviewing method was used and, for this method, interviewers visited the target households in person and input the responses to the structured questionnaires to the portable computer directly. Fifty-three interviewers were trained from Aug 10 to Aug 13, 2015, prior to the survey. When it was hard to conduct an in-person interview due to speech impairment, hearing impairment, and other difficulties, the spouse of the target was surveyed. Among 3,019 target subjects (households), 2,807 subjects were excluded from the analysis because they did not live with a preschooler (≤5 years old). This study analyzed 212 subjects.

B. Measurements

The outcome variable was defined as the demand (i.e., required and not required) of the Voucher Program for Speech Language Therapy. When a target subject responded that he or she did not know "the Voucher Program for Speech Language Therapy, an interviewer explained it ("Voucher Program for Speech Language Therapy is a program to provide language development services for children who are raised by parents with hearing impairment or language impairment") and identified the demand.

The explanatory variables included age, number of children, mean monthly living expenses, whether a household was eligible for National Basic Living Security (yes or no), whether a subject use child care services (yes or no), family life satisfaction (dissatisfied, okay, or satisfied), economic activity after giving birth (yes or no), whether a subject knew the welfare counselor service of Seoul (don't know or know), welfare-related information acquisition (none, inquiring to a welfare facility in person, inquiring to a community service center in person, inquiring to a local resident, call center, internet search, and friend/relative/friend), satisfaction of education policy (satisfaction, okay, or dissatisfaction).

III. Analysis Methods

A. Model Development and Evaluation

Data were divided into training data (70%) and test data (30%) in order to develop a model to predict the demand of the Voucher Program for Speech Language Therapy. The development of the prediction model was based on weighted random forests algorithms. The results of the developed prediction model were compared with the results of a decision tree based on classification and regression tree (CART). The prediction performance of the developed model was evaluated using a confusion matrix. Moreover, the importance of variables and major risk factors were compared.

B. Bagging Tree

Bootstrap aggregating (Bagging) is an ensemble technique that combines multiple bootstrap samples and predicts outcome variable. It is mainly used for models with a small bias and a large variance [13]. Bootstrap means sampling with replacement with having the same sample size for various data. The bth bootstrap sample can be calculated as shown in Equation (1).

$$Z^{(b)} = \left(z_1^{(b)}, ..., z_N^{(b)}\right), where\ z_i^{(b)} = \left(x_i^{(y)}, y_i^{(b)}\right),\ i = 1, ..., N. \quad (1)$$

Bagging tree uses a decision tree model for each bootstrap sample [14]. The decision tree model has a large variance because a tree has a completely different structure according to the first divided variable (j) and division point (s) [15]. Therefore, it is possible to reduce the variance of an unstable tree model by obtaining the mean after constructing multiple tree models through bagging. Each tree model using bootstrap uses a tree model without pruning to minimize bias.

C. Random Forests

Random forests are one of the ensemble techniques that makes a tree model using bootstrap samples and predicts by integrating all models [16]. Random forests conduct division by randomly selecting m-dimension, which is smaller than p-dimension, explanatory variables rather than p-dimension explanatory variables. Random forests have the advantage of using out of bag (OOB) samples because they use bootstrap samples [17,18]. The importance variables score can be calculated easily through permutation [19], and the mean square error (MSE) of the OOB sample is calculated using the regression tree model generated from the bootstrap samples.

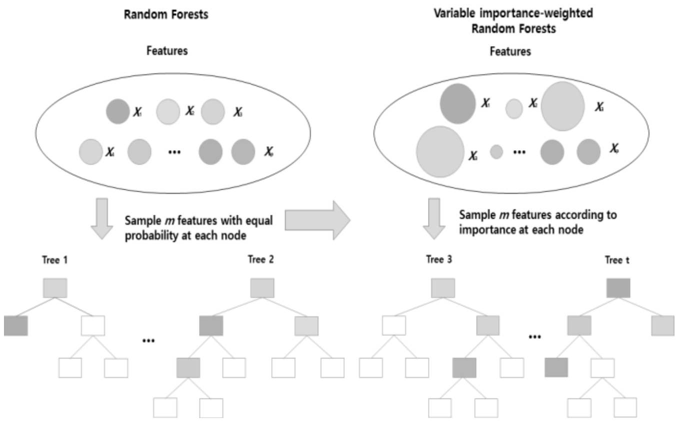

Fig. 1. Concepts of Weighted Random Forests Algorithm.

D. *Weighted Random Forests*

Random forests are one of the ensemble techniques and it conducts model averaging by applying the same weight to each tree model. It is possible that random forests generated by bootstrap have good models and bad models [20]. If model averaging is carried out by applying higher weights to good tree models, it can have better prediction power than the classical random forests giving the same weight to all tree models [21]. The weighted random forests algorithm is developed based on this concept (Fig. 1).

The weighted random forests also use OOB samples. When b = 1, . . . , B and MSE e(b) of OOB sample (O(b)) was calculated using a tree model generated by the bth booststrap sample (Tr(fb)), it is assumed that a large e(b) means a bad tree model and a small e(b) means a good tree model. The weighted random forests are defined as a model averaging technique using the weight, which is given to each tree model (Tr(fb)) using the calculated e(b). In this model, Akaike weights were used for AIC model selection [22].

IV. Results

A. *General Characteristics of Subjects According to the Demand of Voucher Program for Speech Language Therapy*

Table 1 shows the general characteristics of the subjects according to the demand of the Voucher Program for Speech Language Therapy. Among the 212 subjects, 112 (52.8%) responded that the Voucher Program for Speech Language Therapy was necessary. The results of chi-square test revealed that there a significant ($p<0.05$) difference in welfare-related information acquisition between the subjects who responded that the Voucher Program for Speech Language Therapy was needed and those who responded that it was not needed. The demand for the Voucher Program for Speech Language Therapy was higher in the group (53.7%) that obtained welfare related information from local residents.

TABLE I. General Characteristics of the Subjects based on Demand of the Voucher Program for Speech Language Therapy, n (%)

Variables	Demand of the Voucher Program for Speech Language Therapy (n=212)		p
	Not required (n=100)	Required (n=112)	
Age, mean±SD	38.9±8.1	38.1±7.3	0.446
Number of children, mean±SD	1.6±0.8	1.5±0.8	0.962
Mean monthly living expenses (KRW), mean±SD	372.3±980.1	244.3±103.4	0.171
Whether a household was eligible for National Basic Living Security			0.370
Yes	1 (25.0)	3 (75.0)	
No	99 (47.6)	109 (52.4)	
Whether a subject use child care services			0.461
Yes	63 (49.2)	65 (50.8)	
No	37 (44.0)	47 (56.0)	
Family life satisfaction			0.533
Dissatisfied,	4 (36.4)	7 (63.6)	
Okay	45 (51.1)	43 (48.9)	
Satisfied	51 (45.1)	62 (54.9)	
Economic activity after giving birth			0.317
Yes	33 (42.9)	44 (57.1)	
No	67 (50.0)	67 (50.0)	
Whether a subject knew the welfare counselor service of Seoul			0.128
Know	7 (31.8)	15 (68.2)	
Don't know	93 (48.9)	97 (51.1)	
Welfare-related information acquisition			0.011
None	39 (60.9)	25 (39.1)	
Inquiring to a welfare facility in person	4 (66.7)	2 (33.3)	
Inquiring to a community service center in person	16 (43.2)	21 (56.8)	
Inquiring to a local resident	0 (0.0)	2 (100.0)	
Call center	5 (71.4)	2 (28.6)	
Internet search	29 (34.1)	56 (65.9)	
Friend/relative/friend	7 (63.6)	4 (36.4)	
Satisfaction of education policy			0.718
Satisfaction	39 (44.8)	48 (55.2)	
Okay	37 (48.1)	40 (51.9)	
Dissatisfaction	18 (52.9)	16 (47.1)	

B. Results of Weighted Random Forests Model Development

The model to predict the demand of the Voucher Program for Speech Language Therapy was developed through the weighted random forests and the predictive power was compared with the results of CART (Table 2). Weighted random forests had higher classification accuracy than CART in both training and test data. The analysis results of training data showed that the classification accuracy was 72.5% for weighted random forests and 71.2% for CART. For test data, it was 72.1% for weighted random forests and 70.8% for CART.

C. Comparison of Language-Related Factors by Model

Table 3 shows the results of constructing prediction models based on CART and weighted random forests using 10 explanatory variables for predicting the demand of the Voucher Program for Speech Language Therapy. In this study, the weighted random forests model estimated the key variables using the decrease of the GINI coefficients [23]. In the CART-based model, four variables (i.e., economic activity after childbirth, the awareness of Seoul's welfare counselor operation, mean monthly living expenses, and whether welfare related information was obtained) were predicted as the factors associated with the demand of the Voucher Program for Speech Language Therapy and the accuracy was 70.8%. The weighted random forest-based model predicted five variables (i.e., whether preschooler caregiving services were used or not, economic activity after childbirth, the awareness of Seoul's welfare counselor operation, mean monthly living expenses, and whether welfare related information was obtained) as the variables associated with the demand of the Voucher Program for Speech Language Therapy and the accuracy was 72.1%.

TABLE II. THE PREDICTION PERFORMANCE OF THE DEVELOPED MODEL

Data	Model	Accuracy (%)
Training data	Classification and regression tree	71.2
	Weighted random forests	72.5
Test data	Classification and regression tree	70.8
	Weighted random forests	72.1

TABLE III. COMPARISON OF LANGUAGE-RELATED FACTORS BY MODEL

Model	Factors	Characteristics
Classification and regression tree	4	Economic activity after childbirth, the awareness of Seoul's welfare counselor operation, mean monthly living expenses, and whether welfare related information was obtained
Weighted random forests	5	Whether preschooler caregiving services were used or not, economic activity after childbirth, the awareness of Seoul's welfare counselor operation, mean monthly living expenses, and whether welfare related information was obtained

V. DISCUSSION

The establishment and expansion of Voucher Programs for Speech Language Therapy are very important in the aspect that it can enhance the language development of children in the high-risk communication disorder group and the quality of family's life. This study developed a model to predict the demand for language therapy service targeting preschooler without a disability and under parents with language or hearing impairment using the weighted random forest algorithm.

The weighted random forest-based prediction model showed that mean monthly living expenses (reflecting the mean household income) and whether welfare-related information was obtained or not were important factors to predict the demand for language therapy service. On the other hand, it was confirmed that whether receiving the National Basic Living Security or not, reflecting the low-income status of a household, was not a key factor. These results posed two meanings. First, it is necessary to choose the support target based on the actual demand survey rather than prioritizing the low-income class to expand the language therapy service. Additionally, it is needed to increase the budget and alleviate the income criteria for application. Korean Ministry of Health and Welfare allocated 83 billion KRW for development and rehabilitation service projects in 2019, which is 6.7 billion KRW increase from 2018 budget and predicted that it would support 57,094 children [24]. However, selection does not guarantee that all expenses for developmental rehabilitation services will be covered, and the grade will be determined according to the income level and the subsidy varies accordingly. It is because the demand for language therapy services far exceeds the supply. It is mainly because the government did not accurately estimate the demand for the project. In other words, the government tends to make a rough estimate from allocating budget and makes it similar to the previous year's, and the number of target subjects always exceeds the actual demand to exhaust the budget and malfunction the service continuously. Additionally, although each municipal receives application every year, it is already full and the existing applicants are given priority. Therefore, the entry barriers are too big for the new applicants and it requires a counterplan. In the case of Gwangju metropolitan city, the budget for language therapy of it is 4,751,074 USD in 2019, which means that 140 applicants will not be supported when they receive graded payment based on the number of registered people with disabilities [25]. Therefore, actual demand should be surveyed based on mean household income, not on low-income households.

Second, it is necessary to expand the language therapy service institutions and actively advertise the system in order to successfully expand the service. Under the existing unequal service supply system, relatively unfavorable groups should be set as service priority subjects and service institutions should be secured not to exclude them from services. Since it is possible that the access to services may not be guaranteed due to a serious information gap [26], supplying institutions and local governments, service and administrative agencies, should develop active promotion strategies not to make the information gap lead to inequality in social welfare services.

Another finding of this study was that the accuracy and prediction power of weighted random forests was higher than those of CART. It is believed that the weighted random forests had higher accuracy than CART because the former is based on the bagging algorithm that generates diverse decision trees from 500 bootstrap samples [27,28]. CART can be used for both regression and classification problems, and it is widely used because it is simple yet has strong prediction power [29]. The decision tree has a small bias, but the variance of the model is large because the structure of a tree model varies greatly according to the first divided variable. When N-1 divisions are performed for data (n=N), each area contains only one datum and a large tree model generated by it generally has an overfitting problem. Pruning is performed to prevent it, and a suitable size tree model is selected as a final model to conduct a prediction. However, the decision tree is still highly likely to have a overfitting problem.

Hothorn & Lausen (2003) [14] proposed the bagging tree to overcome the overfitting problem in this prediction model. The bagging tree has the advantage of minimizing the bias using each tree model created by bootstrap samples and effectively reducing the variance of the model at the same time [14]. However, bootstrap samples are positively correlated because they are sampled with replacement from the same data. In other words, there is a positive correlation between tree models, so the prediction value of the bagging tree has a higher variance. To complement this, Breiman (2001) [30] proposed random forests that can reduce the correlation for each tree model created by bootstrap samples. Although random forests is also an ensemble technique that predicts result variables by generating many tree models using bootstrap samples, just like the bagging tree. However, the algorithm for building the tree model of it is different from that of the bagging tree. Unlike the general tree model, which starts from dividing all the explanatory variables in the p-dimension in the growth process, the tree models constituting random forests use only randomly selected m ($\leqslant$ p) variables in each division process. The key idea of random forests is to give randomness in the growth process of the tree model. It can reduce the correlation for each tree model through it to make it have better prediction power than the bagging tree [31]. Therefore, the results of this study suggest using weighted random forests for developing a highly accurate prediction model.

VI. Conclusion

The weighted random forests had higher accuracy than the decision-making trees because they maintained the bias of trees and reduce the variance. Random forests that extract many training datasets, generate trees, and predict a target variable, are suitable to construct a prediction model using data containing many variables such as the Voucher Program for Speech Language Therapy or big data. Particularly, weighted random forests that give higher weights to better-performing tree models would have better prediction power than existing random forests granting the same weight to all tree models. Future studies need to seek ways to enhance the performance of weighted random forest models. It is needed to develop systematic policies to expand consumer-oriented language therapy services based on the developed prediction model for the Voucher Program for Speech Language Therapy.

Acknowledgment

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

References

[1] Ministry of Health & Welfare, Disability registration status, Korea Institute for Health and Social Affairs, Sejong, 2017.

[2] S. J. Kwon, Health care and health status of people with disabilities: policy issues. Health and Welfare Policy Forum, vol. 263, pp. 21-33, 2018.

[3] R. Paul and C. F. Norbury, Language disorders from infancy through adolescence, Elsevier Health Sciences, Oxford, 2012.

[4] D. V. Bishop, and L. Leonard, (Eds.), Speech and language impairments in children: causes, characteristics, intervention and outcome. Psychology press, Hove, 2014.

[5] H. Kang, Achievement of social service voucher. Health·welfare Issue & Focus , vol. 171, pp. 1-8, 2013.

[6] R. J. Daniels, and M. J. Trebilcock, Rethinking the welfare state: Government by voucher, Routledge, Abingdon-on-Thames, 2013.

[7] H. S. Bae, The meaning and assignments of introducing voucher system in social welfare services of Korea. Social Welfare Policy, vol. 31, pp. 319-342, 2007.

[8] Y. I. Jang, A study on the childcare voucher as a new system of state's subsidy for childcare: in the perspective of children's right. The Korea Association of Child Care and Education, vol. 58, pp. 189-217, 2009.

[9] J. A. Son, Rethinking about Publicness of Social Welfare. Journal of Critical Social Welfare, vol. 62, pp. 131-155, 2019.

[10] H. S. Jo, Prospects for 2019 Health and Welfare Policy. Health and Welfare Policy Forum, vol. 267, pp. 2-5, 2018.

[11] S. J. Yu, Qualitative research on paenting stress & homeostatic factors of parents with disabilities in vocational rehabilitation service(voucher) of disabled children in community welfare centers – grounded theory approach. Journal of Regional Studies, vol. 27, no. 1, pp. 123-143, 2019.

[12] S. Kang and J, Moon, Differences between sociodemographic characteristics, instrumental activities of daily living, and healthcare needs in disabled persons with and without language. Journal of The Korean Society of Integrative Medicine, vol. 7, no. 1, pp. 37-45, 2019.

[13] L. Breiman, Bagging predictors. Machine learning, vol. 24, no. 2, pp. 123-140, 1996.

[14] T. Hothorn, and B. Lausen, Bagging tree classifiers for laser scanning images: a data-and simulation-based strategy. Artificial Intelligence in Medicine, vol. 27, no. 1, pp. 65-79, 2003.

[15] T. G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, vol. 40, no. 2, pp. 139-157, 2000.

[16] H. Byeon, Developing a model for predicting the speech intelligibility of South Korean children with cochlear implantation using a random forest algorithm. International Journal of Advanced Computer Science & Applications, vol. 9, no. 11, pp. 88-93, 2018.

[17] C. Strobl, A. L. Boulesteix, A, Zeileis, and T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics, vol. 8, no. 25, pp. 1-21, 2007.

[18] R. Genuer, J. M. Poggi, and C. Tuleau-Malot, Variable selection using random forests. Pattern Recognition Letters, vol. 31, no. 14, pp. 2225-2236, 2010.

[19] 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 & Applications, vol. 9, no. 9, pp. 494-499, 2018.

[20] S. J. Winham, R. R. Freimuth, and J. M. Biernacka, A weighted random forests approach to improve predictive performance. Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 6, no. 6, pp. 496-505, 2013.

[21] A. Booth, E. Gerding, and F. Mcgroarty, Automated trading with performance weighted random forests and seasonality. Expert Systems with Applications, vol. 41, no. 8, pp. 3651-3661, 2014.

[22] E. J. Wagenmakers, and S. Farrell, AIC model selection using akaike weights. Psychonomic Bulletin & Review, vol. 11, no. 1, pp. 192-196, 2004.

[23] H. Pham, and S. Olafsson, On Cesaro averages for weighted trees in the random forest. Journal of Classification, pp. 1-14, 2019.

[24] Ministry of Health & Welfare, Developmental Rehabilitation Service, Korea Institute for Health and Social Affairs, Sejong, 2019.

[25] Gwangju City Hall, Budget scale of Gwangju metropolitan city, Gwangju City Hall, Gwangju, 2019

[26] H. Kang, Issues in Social Services Policy. Health and Welfare Policy Forum, vol. 125, pp. 6-22, 2007.

[27] H. Byeon, A prediction model for mild cognitive impairment using random forests. International Journal of Advanced Computer Science & Applications, vol. 6, no. 12, pp. 8-12, 2015.

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

[29] G. De'ath, and K. E. Fabricius, Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, vol. 81, no. 11, pp. 3178-3192, 2000.

[30] L. Breiman, Random forests. Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[31] R. O'Brien, and H. Ishwaran, A random forests quantile classifier for class imbalanced data. Pattern Recognition, vol. 90, pp. 232-249, 2019.