IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 7, 2020, accepted September 4, 2020, date of publication September 21, 2020, date of current version October 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3025553

Prediction Model of Dementia Risk Based on XGBoost Using Derived Variable Extraction and Hyper Parameter Optimization

SEONG-EUN RYU[1], **DONG-HOON SHIN**[2], **AND KYUNGYONG CHUNG**[1]

[1]Division of Computer Science and Engineering, Kyonggi University, Suwon 16227, South Korea
[2]Department of Computer Science, Kyonggi University, Suwon 16227, South Korea

Corresponding author: Kyungyong Chung (dragonhci@gmail.com)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1F1A1058651).

ABSTRACT With the development of healthcare technologies, the elderly population has grown and therefore populating ageing has emerged as a social issue. It is a cause of rise in patients with geriatric disorders, among which dementia is very fatal to the elderly's activities of daily living. In the studies on dementia risk prediction, a method using deep learning was proposed. It requires a lot of image data and much time to learn. Therefore, this study proposes a prediction model of dementia risk based on XGBoost using derived variable extraction from numericalized dementia data and hyper-parameters optimization. The proposed method extracts variable importance from typical independent variables with the use of gradient boosting and then generates derived variables. The generated derived variables are applied to variable importance analysis and thereby a Top-N group is created. Then, for achieving optimal performance in line with the data characteristics of each Top-N group, hyper-parameter tuning is conducted. With the optimized groups, XGBoost model based performance is evaluated. In addition, for the performance evaluation of the proposed model, goodness-of-fit for machine learning classification models is evaluated. According to the Top-N group performance evaluation with different numbers of derived variables, Top-20 model showed the best performance, and the optimized hyper-parameter values were eta = 0.10, gamma = 0, max_depth = 4, and min_child_weight = 1. As a result, the accuracy of the XGBoost model proposed in this study was 85.61%, and its F1-score was 79.28%. When the proposed model is compared with Decision Tree, Random Forest, SVM, and k-NN models, it has the best performance.

INDEX TERMS Healthcare, machine learning, dementia, extreme gradient boosting, hyper-parameter optimization, grid search, classification, parallel processing, risk prediction.

I. INTRODUCTION

With the development of medical technologies, average life expectancy has increased, birthrate has decreased, and the elderly population has been on the rapid rise annually in the world [1], [2]. According to the UN, in an ageing society, the elderly population aged sixty-five years and over accounts for over 14% of the total population, and in a super-ageing society, it amounts to over 20% of the total population. According to Statistics Korea, the elderly population aged and over accounts for 14.9% of the Korean total population [3]. The rate is 9% higher than the average rate of the global elderly population [4]. As such, the Korean society becomes a super-ageing society the fastest in the world, beyond an ageing society. In the circumstance of the rapid population ageing in Korea, the number of patients suffering from three major geriatric disorders-dementia, Parkinson's disease, and cerebral stroke-has on the rapid increase [5]. According to the present condition of domestic dementia centers, of 7,380,000 elderly persons aged 65 and over, 750,000 are estimated to suffer from dementia. and the total prevalence rate of dementia is estimated to be 10.16% [6]. The rapidly growing elderly population leads to a sharp rise in the number of dementia patients, which raises a severe social issue [7]. As a clinical syndrome, dementia causes one's lowering cognitive function in multiple areas, such

The associate editor coordinating the review of this manuscript and approving it for publication was Jiri Mekyska.

as memory, language, and judgment and thereby prevents his or her activities of daily living [8]. There are different types of dementia, such as Alzheimer's disease dementia, vascular dementia, and Parkinson's disease dementia. Causes of dementia are not clear, but complex. An elderly dementia patient has difficult recognizing the fact that he or she has dementia. Up to now, there are not appropriate treatments for dementia [9]. For these reasons, detecting dementia in its early stage is far more important than treating the disease. Therefore, many studies on the early prediction of dementia risk have been conducted. Van de Vorst *et al.* [10] proposed the death rate prediction model by using the cohort data of elderly dementia patients on the basis of logistics regression analysis. The elderly dementia patient cohort was completed with the uses of hospital discharge record, national death cause record, and population record. The proposed method can predict individual patients' risk in their activities of daily living on the basis of different kinds of data, and is easily applicable to clinical treatments. However, the applied model was not verified externally, and its logistic regression model had limited performance. Miled *et al.* [11] proposed the dementia prediction model by using Electronic Medical Record (EMR) data on the basis of Random Forest (RF) and SVM ML model. EMR means a variety of health information that includes cognitive test data, MRI image data, and neuropsychiatric test data. The accuracy of each model using RF and SVM was 73% and 76%, respectively. However, in the research, model parameter optimization was not sufficiently applied to data sets. Therefore, in order to increase the performance of dementia risk prediction, this study proposes a XGBoost (eXtreme Gradient Boost) based dementia risk prediction model using the extraction of derived variables and hyper-parameter optimization. The data set, which is used for the XGBoost based modeling of dementia risk prediction, is collected from Open Access Series of Imaging Studies (OASIS) and then is preprocessed. XGBoost is Classification and Regression Trees (CART) ensemble model based on Gradient Boosting Machine (GBM). It solves the problems of typical GBM which are slow execution time and over-fitting, and therefore it features fast learning and prediction [12]. In order to give a positive influence on the prediction of Clinical Dementia Rating (CDR) as a dependent variable, the proposed method extracts new derived variables from typical independent variables. The extraction process is aimed at creating significant new indexes and adding them to independent variables, positively influencing the performance of XGBoost model. In the process, a model is improved with parallel processing technique and hyper-parameter optimization, and thereby an optimized dementia risk prediction model with high prediction rate is offered.

II. RELATED WORK

A. XGBOOST (EXTREME GRADIENT BOOSTING) CLASSIFICATION TECHNOLOGY IN ENSEMBLE

As one of machine learning techniques, XGBoost is an ensemble model. It utilizes the boosting technique to improve weak classifier models sequentially and making them as strong classifier ones [13]. In addition, as a CART (Classification and Regression Trees) based model, it is applicable to predict both categorical and continuous variables. As a decision tree based model, XGBoost makes residual learned sequentially in Residual Fitting way so as to increase accuracy of data classification. In particular, it applies parallel operation processing at the time of model learning so that it is possible to learn fast [14]. Since there are many hyper-parameters at the time of model learning, it is possible to learn different kinds of data sets flexibly. A XGBoost algorithm designs a prediction model of weak classifier and then evaluates performance with a training set. After that, with the use of Gradient descent, it learns in an ensemble with the new prediction model that has the gradient for better performance as an independent model. In other words, Gradient boosting technique sequentially generates a new model to predict the residual of previous tree models and thereby gradually increases performance [12]. As final prediction, an error is minimized in combination of all models. Equation (1) shows the XGBoost algorithm as a formula.

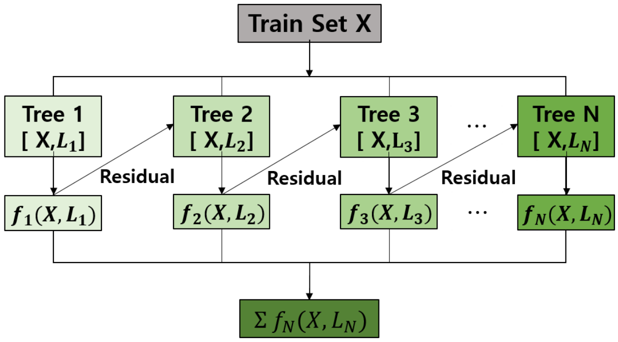


FIGURE 1. XGBoost model process based on gradient descent.

$$Obj=\sum_{k=1}^{j} L\left(y_k,\sum_{n=1}^{N} f_n\left(x_k\right)\right)+\sum_{n=1}^{N}\Omega(f_n),\quad f_n\in F \tag{1}$$

In the Equation (1), Object(Obj) represents a tree ensemble model. L means a loss function which is made with a difference between the actual value and the predicted value. represents the decision tree models generated through learning. It draws an outcome after each independent tree model's score and each leaf node's score are added together, and each added value is compared. N is the number of trees. F represents the set of CART which can be learned in a relevant algorithm. means the regularization term for, a parameter to define the complexity of each tree [15]. It overcomes the over-fitting problem that a typical GBM algorithm faces due to Regularization. Fig. 1 shows the XGBoost model process based on gradient descent for the training set X.

Fig. 1 shows the technique of reducing residual gradually and decreasing an error rate through the gradient descent based ensemble tree learning for CDR (Clinical Dementia

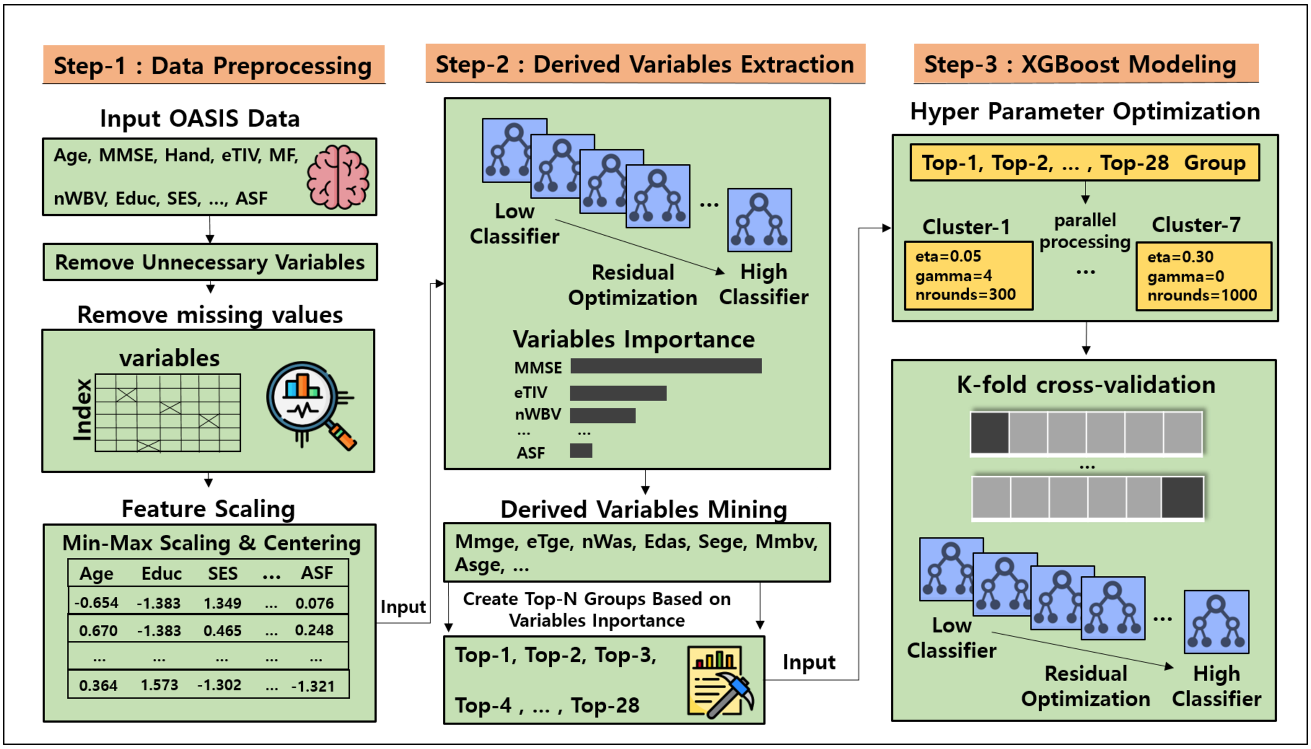
Rating) Classification. As a learning result, the tree model with minimized residual is drawn. To predict the fluctuation of stock price indexes, Hah *et al.* [16] developed the method of predicting the categorical data classification with the use of a XGBoost model. The developed method compared predicted outcomes in each time slot with the uses of time sliding technique and window-size. According to the comparison, the XGBoost model had relatively better performance than other classification models. However, it sets the hyper-parameters of XGBoost to the same values in each model so that a non-optimized model is applied to the learning data of each model. Therefore, this study utilizes Gradient boosting to extract optimal hyper-parameters for each model.

B. PREDICTION OF DEMENTIA RISK BASED ON MACHINE LEARNING

Dementia occurs due to many different causes, such as Alzheimer's disease, vascular disease, or Parkinson's disease, so that no accurate treatments are found up to now [17]. It means that the early detection of dementia is important for prevention. Therefore, many studies have been conducted to detect dementia in its early stage. Mitchell [18] proposed the test for cognitive disorder rating by using the simple cognitive tool Mini-Mental State Exam (MMSE) for dementia risk patients. The proposed method makes it possible to conduct a cognitive test relatively fast and is excellent at psychological cognitive disorder screening. However, in terms of reliability and effectiveness, it is hard to judge dementia, which occurs due to various causes, simply with MMSE index. To solve the problem, Ullah *et al.* [19] proposed the technique of detecting Alzheimer's disease from the image data of OASIS 3D MRI (Magnetic Resonance Imaging) by using Support Vector Machine (SVM) and Convolution Neural Network (CNN). The proposed method extracts a variety of dementia features from the 3D brain image data of a test participant, and applies deep learning on the basis of the difference between the white matter and gray matter of the brain. Inspired by Inception-V4 network, Islam and Zhang [20] proposed the CNN model with redesigned Softmax layer in order for the automated detection and prediction of Alzheimer's disease. A softmax layer has four different output classes, receives a MRI image as input data, and extracts layer-by-layer shape expressions from the first stem layer to the last dropout layer [20]. It is helpful to learn the features of various dementia causes. Unfortunately, a deep learning model has millions of parameters, and it takes a lot of time to train them before their use in production [21]. Moreover, to develop a strong deep learning neural network, it is necessary to collect a great deal of image data [22]. To solve the problem, Manandhar *et al.* [23] proposed the K-Nearest Neighbor based dementia risk detection. The proposed method utilizes the preprocessed OASIS MRI image data and their digitalized number data, and its accuracy is 81.13%. The accuracy of Artificial Neural Network (ANN) model used in their research is 69.81%. Tohka *et al.* [24] performed the Support Vector Machine based dementia feature selection with the use of the brain MRI image data obtained from Alzheimer's Disease Neuroimaging Initiative (ANDI). In MRI machine learning analysis, accuracy of dementia risk detection and stability of selected features are evaluated through SVM. Since image data is used input data, learning time and result output delay can occur [25]. Given the characteristics of SVM model, the more input data, the slower speed, the larger memory allocation, and the lower performance [26]. To solve the problem this study utilizes a XGBoost model as one of machine learning models, which can detect the risk of dementia with the uses of digitalized number data and relatively smaller data. The model uses the number data obtained from OASIS MRI brain image data so as to shorten a model learning time efficiently. In addition, it extracts new derived variables from existing independent variables to predict dementia risk and adds them in order to select more diverse features. This process is aimed at influencing dementia risk prediction positively. As such, many studies have steadily been conducted to detect dementia in its early stage.

III. PREDICTION MODEL OF DEMENTIA RISK BASED ON XGBOOST USING DERIVED VARIABLE EXTRACTION AND HYPER PARAMETER OPTIMIZATION

For dementia risk prediction, this study proposes the XGBoost based dementia risk prediction model using the extraction of derived variable and the optimization of hyper-parameters. The proposed method has a three-step process structure. Fig. 2 shows the process structure of the dementia risk prediction model. As shown in Fig. 2, in the first step, dementia data is collected and preprocessed. The data is OASIS-1 and OASIS-2 data offered by OASIS (Open Access Series of Imaging Studies) [27]. OASIS-1 and OASIS-2 data includes the gender, age, years of education, socioeconomic status, Mini-Mental State Examination (MMSE) result, and longitudinal/cross-sectional brain Magnetic Resonance Imaging (MRI) data of 566 test subjects. In the way of preprocessing, a missing value of data is removed. Since numbers are largely different between variables, Min-Max method is applied to training in order to perform Feature Scaling. In the second step, a Top-N group is extracted on the basis of variable importance of gradient boosting. A Top-N group is generated through the extraction of derived variables from existing independent variables and the analysis of variable importance. Top-N represents a new group of independent variables that will be used in a XGBoost model along with the top N number of variables in terms of variable importance. It positively influences the prediction of Clinical Dementia Rating (CDR) which represents the clinical scale of dementia. In the last step, hyper-parameter optimization is applied to each one of Top-N groups, in order to improve dementia risk prediction. To draw optimal parameter values efficiently in the process, Grid Search technique and parallel processing based on clusters are applied. Such methods are aimed at minimizing the time of model learning.



Age	Educ	SES	...	ASF
-0.654	-1.383	1.349	...	0.076
0.670	-1.383	0.465	...	0.248
...	...	...	...	...
0.364	1.573	-1.302	...	-1.321

FIGURE 2. Process structure of the dementia risk prediction model.

A. DATA COLLECTION AND PREPROCESSING FOR THE PREDICTION OF DEMENTIA

To establish a dementia risk prediction model, this study collects the dementia data (OASIS-1, OASIS-2) offered by OASIS [27]. The data includes the number data of brain structure which is generated with the longitudinal or cross-sectional brain MRI data of test subjects according to whether or not they have dementia. In addition, it includes the unique number, gender, handedness, age, years of education, socioeconomic status, MMSE result, and CDR index of the test subjects. OASIS-1 incorporates the cross-sectional brain MRI data of 416 test subjects aged 18 to 96. OASIS-2 includes the longitudinal brain MRI data of 150 test subjects aged 60 to 96. The collected subjects are a part of the participants in MRI research at University of Washington and the Alzheimer's Disease Research Center (ADRC) of the university. OASIS-1 and OASIS-2 have a lot of missing values so that preprocessing is performed. In the way of preprocessing, a volume of data is reduced. As a result, a lack of data occurs, which negatively influences dementia risk prediction. For this reason, OASIS-1 and OASIS-2 data sets are combined together. Table 1 shows the data variable structure of OASIS-1 and OASIS-2.

Shown in Table 1, the common variables of OASIS-1 and OASIS-2 are ID, M/F, Hand, Age, Educ, SES, MMSE, CDR, eTIV, nWBV, ASF, and Delay, each of which is used as an independent variable or dependent variable. A dementia patient can have a lowering intellectual level. By finding the causes of dementia in the way of analyzing the changes in the brain structure and in the particular regions of the brain, it is possible to detect the disease in its early stage. Accordingly, estimated Total Intracranial Volume (eTIV), normalized Whole-Brain Volume (nWBV), and Atlas Scaling Factor (ASF) are used as the longitudinal and cross-sectional brain number data. Fig. 3 shows an example of the cross-sectional and longitudinal brain MRI data of test subjects, which are offered by OASIS [28], [29].

TABLE 1. Configure data variables for oasis-1 and oasis-2.

	Common	Difference
OASIS-1	ID, M/F, Hand, Age, Educ, SES, MMSE, CDR, eTIV, nWBV, ASF, Delay	-
OASIS-2		MRI_ID, Group, Visit

From the cross-sectional and longitudinal MRI imaging data shown in Fig. 3, eTIV, nWBV, and ASF number data are extracted. eTIV means the total intracranial volume estimated by MRI examination. nWBV represents the normalized whole-brain volume measured with MRI imaging data. ASF means the total cranial area measured with MRI imaging data. Digital values are extracted from image data, and are used as input data of a XGBoost model. Gender, years of education, and socioeconomic status, as well as age, influence the brain structure change and ageing speed related to dementia [30]. Accordingly, the test subjects' M/F

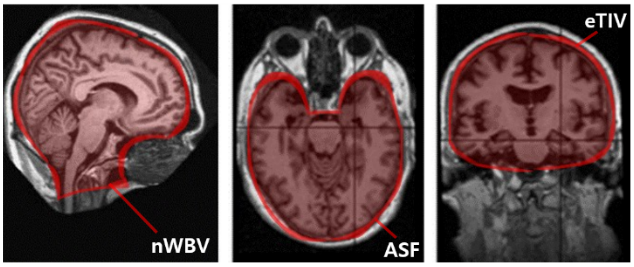


FIGURE 3. An example of the cross-sectional and longitudinal brain MRI data of test subjects, which are offered by OASIS.

(Gender; female:0, male:1), Age (test subject's age), Educ (Years of education), SES (Socioeconomic status; the highest status: 1, the lowest status: 5), and MMSE (Mini-Mental State Examination) data are used. MMSE is a standardized questionnaire survey to analyze a test subject's cognitive intelligence and dementia. In the examination, the full score is 30 points; the lower score, the higher dementia risk [31]. As a dependent variable to predict dementia risk, CDR is used in this study. The CDR developed by ADRC of University of Washington in the US is a scale to evaluate the severity of dementia. Dementia breaks out due to complex causes, not one cause. Therefore, the CDR is evaluated in six domains: Memory, Problem-Solving, Judgment, Orientation, Community Affairs, Home and Hobbies, and Personal Care. In the CDR score categories, '0' = 'Normal', 0.5 = 'Very Mild Dementia', 1 = 'Mild Dementia', 2 = 'Moderate Dementia', and 3 = 'Severe Dementia'. OASIS data include the variables unnecessary for learning, the variables not influential on response variable CDR, and missing values, each of which has a different number unit. For this reason, data preprocessing is applied appropriately in order to remove variables without discrimination and missing values, and feature scaling is performed. Fig. 4 shows dementia data preprocessing.

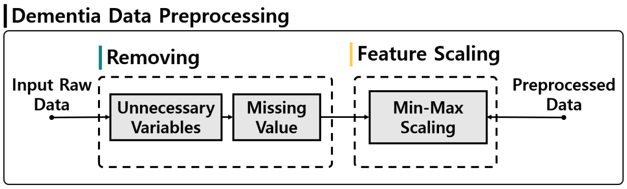


FIGURE 4. Dementia data preprocessing.

In Fig. 4, Input Raw Data means the combined data of OASIS-1 and OASIS-2. The two data sets include the data missed in statistics, or the values not saved in variables in the data collection step. These missing values can negatively influence dementia risk prediction so that the data row including missing values is removed. In this process, a data volume is reduced. Therefore, by combining the two data sets with the same format, it is possible to secure more data with a variety of situations. Among unnecessary and common variables, there are ID, Hand, and Delay. ID is a test subject's index number. The variable is removed from independent variables, since it does not influence the dependent variable CDR at all. All the test subjects are right-handed so that Hand, the variable representing handedness, is meaningless and removed. Delay means a MRI delay time. Since the variable does not influence CDR prediction at all, it is removed from independent variables. In Table 1, differences are MRI_ID, Group, and Visit variables. MRI_ID is a MRI serial number, which does not influence the dependent variable CDR at all. Group is the binary-type variable to represent whether there are any dementia symptoms. The variable is unnecessary, since it can be expressed with the result of CDR. Visit is the number of visits, which does not influence CDR at all. These three variables are removed, since they have no discrimination for CDR prediction. In the last step of preprocessing, variables have a different number unit. Therefore, Min-Max Scaling as a Feature Scaling method is applied for normalization. Equation (2) shows the formula of Min-Max scaling [32].

$$Min - MaxScaling\ X = \frac{\boldsymbol{x - min\,(x)}}{\boldsymbol{max\,(x) - min\,(x)}} \tag{2}$$

In the Equation (2), max(x) and min(x) represent the maximum value of x and minimum value of x, respectively. Data is adjusted in the range of 0 to 1, and is normalized to make the total standard deviation of a data set as '1'. In addition, the mean value of the total data set is normalized to '0' through centering technique. Scaling and centering makes all features set to the same scale, supports faster learning, and prevents over-fitting [33].

B. EXTRACTION OF DERIVATIVE VARIABLES USING GRADIENT BOOSTING

For the improvement in dementia risk prediction, derived variables are generated and then are added to existing independent variables. Through the creation of derived variables, significant and new indexes are generated with existing independent variables. In terms of risk prediction, learning through the creation of a significant ratio of independent variables, rather than simply using the values of existing independent variables, positively influences dependent variables. To create derived variables, gradient descent based Gradient Boosting is applied to extract Cover, Frequency, and Gain indexes that contribute to variable importance. Fig. 5 shows the relative values of independent variables according to Cover, Frequency, and Gain values.

In Fig. 5, Gain is the measured value of the contribution to each tree of an ensemble model. Cover is the relatively measured value of the observed value through the leaf node of each tree in the model. Frequency is the measured value as to how frequently each independent variable is used decisively in the model. In MMSE, the Frequency value is 12%, which is not a relatively high, but Cover and Gain values are 20% and 42%, respectively, which are relatively high. The Cover value of eTIV and of nWBV is 26% and 19%, respectively, and the Frequency value of eTIV and of nWBV is 27% and 23%, respectively. It means that the use frequency of eTIV is

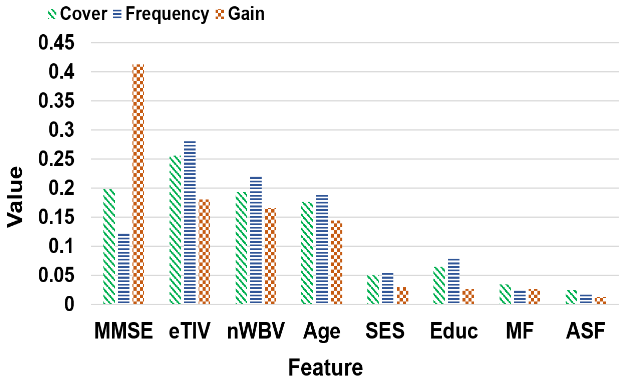


FIGURE 5. Relative values of independent variables according to Cover, Frequency, and Gain values.

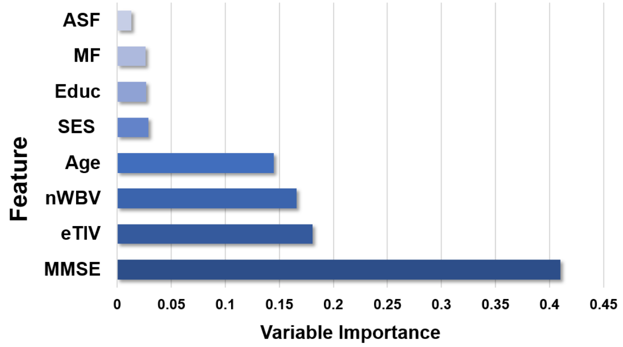


FIGURE 6. Results of variable importance extracted with gradient boost.

relatively higher than that of nWBV. The Gain value of eTIV and of nWBV is 17% and 16%, respectively, which are not much different. It means that the two independent variables similarly contribute to dementia risk prediction, relatively. In terms of Age, Cover, Frequency, and Gain values are 17%, 18%, and 14%. It means that age also has relatively large variable importance. Each variable importance is extracted according to indexes. Fig. 6 shows the results of the variable importance extracted with gradient boost.

In Fig. 6, the variable importance of MMSE is 0.43, which is the highest value. The variable importance of eTIV, of nWBV, and of Age is 0.18, 0.16, and 0.14, respectively, which are relatively high. Compared to the values, the variable importance of Educ, of MF, and of SES is relatively low. ASF has the lowest variable importance. It means that MMSE, eTIV, nWBV, and Age variables have relatively high importance to the prediction of the dependent variable CDR. In the Gradient Boosting process using ensemble technique, a small number of independent variables can lower accuracy and performance of prediction. Therefore, this study applies the extraction of derived variables in order to expand independent variables. Accordingly, seven independent variables except for the categorical independent variable MF are utilized so that a total of twenty-one derived variables are generated. In this way, all the cases that two different variables are selected among seven independent variables are generated. Table 2 shows the twenty-one derived variables created with existing independent variables and their description. It is necessary to make an objective judgment as to how positively the twenty-one derived variables shown in Table 2 influence the prediction of the dependent variable CDR. That is because a lot of variables whose contribution to CDR prediction is not measured can negatively influence CDR prediction. Therefore, derived variables and eight independent variables are used in gradient boosting to extract variable contribution values. In order to compare the contribution of each derived variable with the contribution of each independent variable, this study measured Gain, Cover, and Frequency values.

TABLE 2. Twenty-one derived variables created with existing independent variables and their description.

Derived Variables	Description
Mmge(MMSE/Age)	MMSE ratio for Age of each test subjects
Mmiv(MMSE/eTIV)	MMSE ratio for eTIV of each test subjects
Mmbv(MMSE/nWBV)	MMSE ratio for nWBV of each test subjects
Mmed(MMSE/Educ)	MMSE ratio for Educ of each test subjects
Mmas(MMSE/ASF)	MMSE ratio for ASF of each test subjects
Mmss(MMSE/SES)	MMSE ratio for SES of each test subjects
eTge(eTIV/Age)	eTIV ratio for Age of each test subjects
nWge(nWBV/Age)	nWBV ratio for Age of each test subjects
Edge(Educ/Age)	Educ ratio for Age of each test subjects
Asge(ASF/Age)	ASF ratio for Age of each test subjects
Sege(SES/Age)	SES ratio for Age of each test subjects
eTbv(eTIV/nWBV)	eTIV ratio for nWBV of each test subjects
eTed(eTIV/Educ)	eTIV ratio for Educ of each test subjects
eTas(eTIV/ASF)	eTIV ratio for ASF of each test subjects
eTss(eTIV/SES)	eTIV ratio for SES of each test subjects
nWed(nWBV/Educ)	nWBV ratio for Educ of each test subjects
nWas(nWBV/ASF)	nWBV ratio for ASF of each test subjects
nWss(nWBV/SES)	nWBV ratio for SES of each test subjects
Edas(Educ/ASF)	Educ ratio for ASF of each test subjects
Edss(Educ/SES)	Educ ratio for SES of each test subjects
Asss(ASF/SES)	ASF ratio for SES of each test subjects

Fig. 7 shows three kinds of contribution values of the generated derived variables and existing variables. In Fig. 7, the Gain and Cover values of MMSE are 46.5% and 15.8%, respectively, which are the highest values. A Gain value

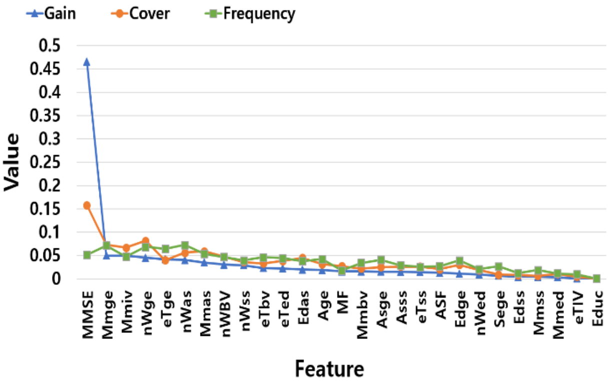


FIGURE 7. Three kinds of contribution values of the generated derived variables and existing variables.

decreases in the order of Mmge, Mmiv, nWge, eTge, nWas, and nWBV. It shows that other variables than MMSE have no big difference in relative contribution. The Cover value, which is a relative value of the observed value through the leaf node in each tree, is 15.8% in MMSE, which is the highest. Subsequently Cover value decreases with a small difference in the order of nWge, Mmge, Mmiv, Mmas, and nWas. Just like gain, the observed values of other variables than MMSE show a small difference. The Frequency value, which represents the frequency of independent variables in the classification of the dependent variable CDR, is 7.2% in nWas, which is the highest. And this decreases with a little difference in the order of Mmge, nWge, eTge, Mmas, and MMSE. It means that, unlike other two indexes, the use frequency of all variables including MMSE has a small difference. Among independent variables, ASF, eTIV, Educ, and SES had low values in terms of three indexes for CDR prediction. In case of SES and eTas, their Gain value is 0.001 or less. Since these variables are not significant, they are excluded in terms of contribution. Based on the indexes, the variable importance of each one of eight independent variables and twenty-one derived variables is extracted. Fig. 8 shows their variable importance from existing independent variables and derived variables using gradient boost.

In Fig. 8, as an independent variable, MMSE has about 45.67% variable importance, which is the highest. That is because compared to other variables, MMSE data can provide a variety of information for dementia risk measurement. Regarding the importance of other variables than MMSE, in between MMSE and nWBV, Mmge, eTge, nWas, and nWge derived variables show a relatively even difference, and an importance value decreases. In between nWBV and Age, an importance value constantly decreases in the order of Mmas and Mmiv. In other words, other variables than MMSE which has the highest variable importance keep a relatively constant difference all, and their importance value decreases. Given the result, on the basis of Fig. 7, it is difficult to set a reference as to what kinds of derived variables should be incorporated into independent variables in order to classify

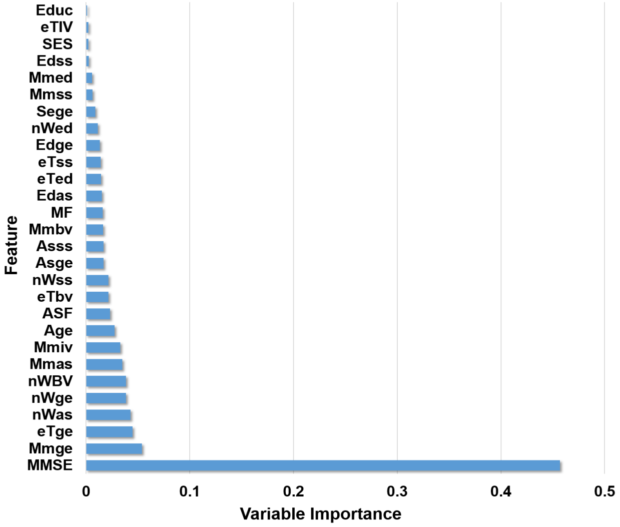


FIGURE 8. Extract the variable importance from existing independent variables and derived variables using gradient boost.

the dependent variable CDR. Therefore, in this study, a Top-N group is generated on the basis of the importance of top variables. The letter 'N' of Top-N means that the top N number of variables in terms of variable importance is used as independent variables for a learning model. At this time, eTas whose variable importance is 0.001 or less is excluded from Top-N groups. N ranges from 1 to 28. For example, Top-12 group has the top twelve variables (MMSE, Mmge, ..., nWss) in terms of variable importance. Algorithm 1 represents the algorithm of extracting derived variables with the use of gradient boosting.

In the Algorithm 1, Input values represent eight independent variables: MMSE, eTIV, nWBV, Age, SES, Educ, ASF, and MF. Output values represent twenty-eight Top-N models.

C. XGBOOST-BASED DEMENTIA RISK PREDICTION MODEL WITH HYPER-PARAMETER OPTIMIZATION

For the effective prediction modeling of dementia risk, the extraction of derived variables and the optimization of hyper-parameters are applied, and XGBoost is used to design a model. The model learns with the uses of various independent variables influencing dementia and the dependent variables CDR indexes. XGBoost uses gradient descent to reduce the residual of a classifier, showing good performance of classification model prediction [34]. By enabling a user to optimize hyper-parameters in line with data characteristics, it is possible to do modeling more flexibly [35]. A hyper-parameter is an important variable that a user needs to set up directly when a certain model is learned through machine learning [36]. If its value is not set up, learning is processed with its default value. Since each data has their own unique characteristic, it is necessary to adjust a hyper-parameter appropriately in order to consider these characteristics. A XGBoost model has over thirty

Algorithm 1 Derivative Extraction Algorithm Using Gradient Boosting

```
Input: Existing Independent Variable(MMSE, eTIV, nWBV, Age, SES, Educ, ASF, MF) → existVariable [1] to existVariable [8]
Output: Top [1] to Top [28] Group
Feature_Importance_list[k] ← NULL
for i is number of existing independent variables do
  result_1[i] ← Gradient_Boosting
      _Algorithm(existVariable[i])
  Cover ← Cover + Cover(result_1[i])
  Frequency ← Frequency + Frequency(result_1[i])
  Gain ← Gain + Gain(result_1[i])
end for
VI_1 ← extract_Variable_Importance(Cover, Frequency, Gain)
for i is existVariable[1] to existVariable [7] except MF do
  for j is (i+1) to existVariable [7] do
    derivedVariable[i, j] ← extract_Derived_Variable(i, j)
    result_2[i, j] ← Gradient_Boosting
        _Algorithm(derivedVariable[i, j])
    Cover ← Cover + Cover(result_2[i, j])
    Frequency ← Frequency + Frequency(result_2[i, j])
    Gain ← Gain + Gain(result_2[i, j])
  end for
end for
VI_2 ← extract_Variable_Importance(Cover, Frequency, Gain)
Feature_Importance_list[k] ← sum_of_Variable_Importance(VI_1, VI_2)
for i is length of Feature_Importance_list[k] do
  if Feature_Importance_list[i] < 0.001
    Except from Feature_Importance_list[k]
  //Arrange the Feature_Importance_list[k] in descending order
  Sort(Feature_Importance_list[k])
  //Create Top-N Group
  Top[i] ← Feature_Importance_list [1] to
          Feature_Importance_list[i]
return Top [1] to Top [28]
```

hyper-parameters. Performance highly relies on how to optimize hyper-parameters. For this reason, it is very important to tune up hyper-parameters. In this study, Gradient boosting was applied to draw the hyper-parameters optimized to each Top-N learning data. In this process, parallel processing based on Grid Search and Cluster was used to increase efficiency. Fig. 9 shows the Grid Search and parallel process applied to Top-N groups. As shown in Fig. 9, with the use of Grid Search, the grid of each hyper-parameter is generated for parameter tuning. Grid Search combines all possible parameters to be optimized, and then produces the value that supports the most improved performance. To minimize the time of parameter tuning, seven clusters that can be used most by the OS hardware in this study are generated for parallel processing. As a result, compared to single processing, the parallel processing speeds up training-set learning. In the above figure, eta means a learning rate. In the way of reducing a weight of each step in tree, model learning is more stabilized. If a learning rate is too high or too low, it is impossible to find the position of minimum loss function. Therefore, it is important to use an optimized value. Gamma specifies the loss reduction which is necessary to split tree nodes rightly in a loss function. In other words, It is a parameter that contributes to making an algorithm conservative. A different value can be specified depending on a loss function. It is significant to select an optimized value. max_depth is the maximum depth of a tree. The larger the max_depth value is, the more a model learns a very characteristic relation for a particular sample. The parameter is used to adjust over-fitting. min_child_weight is the parameter to adjust the minimum value of the sum of weights for all the observed values necessary to a particular child node. When the parameter value is relatively high, under-fitting occurs. For this reason, the parameter is also used to adjust over-fitting. The parameter Objective is set to multi: softprob in all models, since it is necessary to return the prediction probability of each class for multiple classification. As shown in Fig. 9, this study utilizes grid search and parallel processing to extract hyper-parameters which are used to minimize a value of mlogloss as an evaluation index of a training set.

IV. RESULT AND PERFORMANCE EVALUATION

The test hardware and software environment for implementing the proposed outlier detection model is as follows: Window10 Pro, AMD Ryzen 5 1600 Six-Core Processor, NVIDIA GeForce GTX 1070, and RAM 16GB. In the test, dementia risk prediction is performed with the use of OASIS (Open Access Series of Imaging Studies) OASIS-1 and OASIS-2 dementia data, and with the application of the extraction of derived variables and the optimization of hyper-parameters. The test subjects' longitudinal and cross-sectional MRI data are presented as numbers according to whether they have dementia. Unnecessary variables and missing values are removed from the combined data. Data is normalized by Min-Max Scaling and Centering technique. Table 3 shows the preprocessed data.

In the process of learning a risk prediction model, k-folds cross validation is utilized. k-folds cross validation is a data segmentation learning method to improve generalization ability. It segments a training set into k subsets with the same size. The method uses all regions of the whole data in order to validate a model. Since the validation method does not fix a validation set to one, it prevents over-fitting of a particular evaluation data set [37]. However, with a rise in the count of iteration, it can take long to do model training. A k value needs to be selected appropriately depending on the size of data in use. In this study, in order to increase generalization ability, 5-folds cross validation ($k = 5$) is applied in consideration of OASIS data set size. For the performance

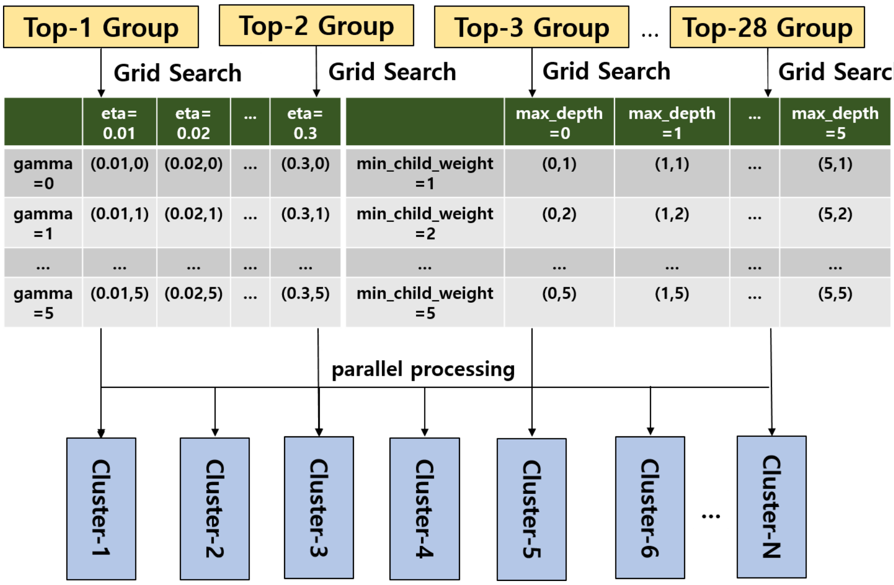


FIGURE 9. Grid Search and parallel process applied to Top-N groups.

TABLE 3. Preprocessed data with min-max scaling and centering techniques.

Idx	MF	Age	Educ	SES	MMSE	…	CDR
1	0	-0.654	-1.383	1.349	-1.204	…	0.5
2	1	0.670	-1.383	0.465	-0.096	…	0.5
3	0	-1.775	-0.890	-1.302	0.734	…	0
4	0	1.077	-1.383	1.349	-0.096	…	1
5	0	-1.062	-1.383	0.465	0.457	…	0
…	…	…	…	…	…	…	…
564	1	0.170	1.587	-1.304	-1.149	…	1
565	0	-0.229	1.258	-1.304	0.762	…	0
566	1	1.281	-1.547	1.349	-0.373	…	1
567	1	1.077	0.916	-1.302	-0.373	…	0.5
568	1	0.364	1.573	-1.302	-0.650	…	2
569	0	-1.266	0.423	-0.418	0.734	…	0
570	0	-1.062	0.423	-0.418	0.734	…	0

evaluation of dementia risk prediction, the performance of within the XGBoost models is evaluated, and the performance comparison with other models is evaluated. More specifically, for each one of Top-N groups generated on the basis of the variable importance of derived variables, the XGBoost model performance is evaluated. In addition, the XGBoost model that has the best performance is compared with other classification models, in terms of the goodness-of-fit evaluation. The XGBoost model without derived variables, and the XGBoost model without hyper-parameter optimization are also compared in terms of performance. As for the performance evaluation of classification, *Accuracy* and *F-measure* are used. Equation (3) shows the formula of *Accuracy*.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \tag{3}$$

In Equation (3), *TP* (True Positive) represents the case that the CDR whose actual dementia risk score is '1' is correctly classified as '1'. *FN* (False Negative) represents the case that the CDR whose actual dementia risk score is '1' is incorrectly classified as '0'. *TN* (True Negative) represents the case that the CDR whose actual dementia risk score is '0' is classified as '0'. *FP* (False Positive) represents the case that the CDR whose actual dementia risk score is '0' is incorrectly classified as '1'. *Accuracy* is an evaluation index to measure a performance in the most intuitive way. However, if data is imbalanced and the input volume of each class is different, data bias problems can occur. For this reason, *F-measure* is used as well. It is an index to evaluate performance in the trade-off integration of *Precision* and *Recall*. Equation (4) shows the formula of *F-measure*.

$$F-measure = \frac{\boldsymbol{2 \times Recall \times Precision}}{\boldsymbol{Recall + Precision}} \tag{4}$$

In Equation (4), *Precision* represents the case that CDR is an actual risk index, among all the cases that CDR is predicted to be dementia risk index. *Recall* represents the case that CDR is predicted to be a risk index, among all the cases that CDR is an actual risk index. With the harmonic mean of *Precision* and *Recall*, it is possible to present a *F-measure* value. In this way, the performance of a model with imbalanced data label can be evaluated effectively.

A. PERFORMANCE EVALUATION OF XGBOOST MODEL ACCORDING TO TOP-N MODEL

In this study, twenty-one derived variables are generated from seven independent variables. It is combined with existing

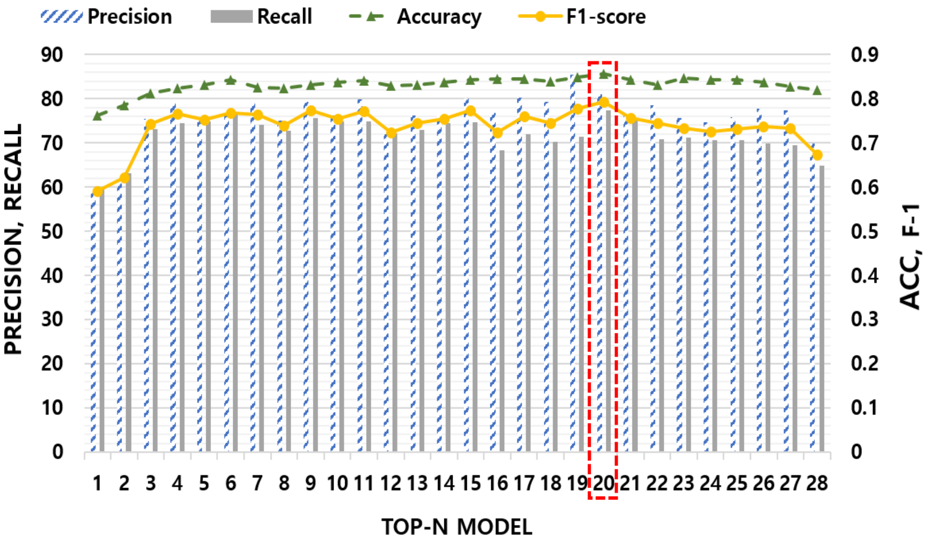


FIGURE 10. Performance evaluation results of Top-N models according to Accuracy and F-measure.

independent variables to extract variable importance. On the basis of the importance order, a total of twenty-eight Top-N groups are made. With the uses of the objective indexes *Accuracy* and *F-measure*, performance evaluation is applied to Top-N models in order to select a set of variables that show the best performance. Fig. 10 shows the visualized performance evaluation results of Top-N models according to *Accuracy* and *F-measure*. In Fig. 10, in terms of *Accuracy*, Top-20 group has 85.61%, which is the highest. In terms of *Precision*, Top-19 group has 85.50%, which is the highest. Regarding Recall, Top-20 had the highest value, or 77.27%. With an increase in the number of the derived variables incorporated in independent variables, the value of each performance evaluation index was measured to be high.

In case of Top-28 group that has twenty derived variables except for eTas (whose variable importance is 0.001 or less) included in independent variables, its *Accuracy* and *F-measure* are 3.69% and 12.01%, which are lower than those of Top-20 group. As the result of the objective performance evaluation with 28 groups, fifteen derived variables of Top-20 group, which are Mmge, eTge, nWas, nWge, Mmas, Mmiv, eTbv, nWss, Asge, Asss, Mmbv, Edas, eTed, eTss, and Edge, are finally extracted. Table 4 shows the optimized hyper-parameter values of Top-20 group.

Table 4 shows the hyper-parameter values extracted with the use of Grid Search and Parallel Processing. These hyper-parameters are used to minimize a value of mlogloss which is an evaluation index of a training set. After the optimization, eta that represents a learning rate was 0.10, gamma for specifying loss reduction was 0, max_depth that represents the maximum depth of an ensemble model tree was 4, and min_child_weight to adjust the minimum value of the sum of weights for all the observed values necessary to a child node was 1.

TABLE 4. Hyper-parameter values extracted with the use of grid search and parallel processing.

Parameter Name	Optimized Value (range)
eta (learning_rate)	0.10 ([0,1])
gamma (min_split_loss)	0 ([0,∞])
max_depth	4 ([0,∞])
min_child_weight	1 ([0,∞])
objective	'multi:softprob'

B. MODEL FIT EVALUATION

For performance evaluation, the dementia risk prediction model proposed in this study is compared with machine learning models in terms of goodness-of-fit. As the models to be compared with a XGBoost model, there are Decision Tree [38], Random Forest [39], SVM (Support Vector Machine) [19], and k-NN (k-Nearest Neighbor) [23]. Decision Tree is a model to solve a classification problem with the use of a binary tree. Random Forest uses Ensemble Bagging technique and generates a decision tree to predict a class, just as Gradient Boosting. SVM finds an optimal model to maximize the margin between two support vectors. k-NN finds the k number of data closest to the model and then classifies a class with high frequency. As the independent variables for these four models, the eight independent variables are used. A XGBoost model is divided into three types. The first type is a conventional XGBoost without hyper-parameter optimization and with no use of derived variables [40]. The second type is XGBoost Plus with hyper-parameter optimization. The third type is the proposed XGBoost model with hyper-parameter optimization and with the use of derived variables. At this time, the XGBoost model based on derived parameters utilizes Top-20 group. nrounds, which represents Epoch in XGBoost at the time of learning, is set to 1000. As an evaluation index for data validation, mlogloss loss function is used. mlogloss is a loss function used mainly for the classification problem called negative log-likelihood. It helps to improve learning flexibility, since it is possible to assume a variety of probability distribution in the modeling process. Each model's performance is compared with the uses of *Accuracy* and *F-measure*. Table 5 shows the goodness-of-fit evaluation results according to performance evaluation indexes.

TABLE 5. Goodness-of-fit evaluation results according to performance evaluation indexes.

Model	Accuracy	Precision	Recall	F1-measure
Decision Tree[38]	80.18	54.45	55.54	54.98
Random Forest[39]	83.68	84.00	66.60	74.29
SVM[19]	80.53	55.51	53.88	54.68
k-NN[23]	65.96	40.87	42.61	41.72
XGBoost[40]	83.16	72.36	66.69	69.40
XGBoost Plus	84.04	76.49	75.25	75.87
Our Model	85.61	81.40	77.27	79.28

As shown in Table 5, compared to other machine learning models, the proposed model has the highest *Accuracy*, or 85.61%. In addition, *Accuracy* and *F-measure* of the proposed model are 1.57% and 4.91% higher than those of XGBoost Plus, so that it has relatively better performance. k-NN model has the lowest Accuracy, or 65.96%. In terms of the classification of CDR as dementia risk index, most models have a higher *Recall* value than *Precision* value. It means that these models expect that a ratio of actual CDR values among predicted CDR values is higher than a ratio of predicted CDR values among actual CDR values. In conclusion, derived variables were extracted from existing independent variables according to Cover, Frequency, and Gain as variable importance indexes, and Gradient boosting based hyper-parameter optimization was applied to Top-N groups that include different numbers of derived variables. Parameters were optimized precisely in line with the data characteristics of each one of twenty-eight Top-N groups. Therefore, the proposed model better performed CDR prediction and produced better performance than other models.

V. CONCLUSION

For dementia risk prediction, this study proposed XGBoost model through the derived variable extraction and hyper-parameter optimization with the use of gradient boosting. It is a CART based ensemble learning model using Boosting that enables weak classifier models to learn sequentially and thereby generates a strong classifier model. A strong classifier is generated in the way of reducing residual through gradient descent technique from an initial classifier. For dementia risk prediction, the variable importance of independent variables is drawn. Top-N groups are generated in order to add significant independent variables that can positively influence CDR prediction. A Top-N group is extracted on the basis of the variable importance of existing independent variables and the derived variables extracted from the independent variables. After variable importance is drawn with the use of gradient boosting, top N variables are grouped sequentially. Of Top1 to Top28 groups, Top-20 group with the best performance was selected by objective performance comparison. Accordingly, for the prediction of Clinical Dementia Rating as a dependent variable, the independent variables of a XGBoost model were finally determined. Top-20 group consists of the following variables: MMSE, Mmge, eTge, nWas, nWge, nWBV, Mmas, Mmiv, Age, ASF, eTbv, nWss, Asge, Asss, Mmbv, MF, Edas, eTed, eTss, and Edge. For the effective prediction of dementia risk, the hyper-parameters of XGBoost are optimized. In line with data characteristics, a user can optimize important parameters directly. Grid Search technique and parallel processing are applied to find optimized parameters more efficiently. For the performance evaluation of the dementia risk prediction based on the derived variable extraction and optimized XGBoost model, different types of XGBoost models and different classification models are compared. A XGBoost model is divided into three types: a type of the model without parameter optimization and with no use of derived parameters; a type of the model with parameter optimization and with no use of derived parameters; a type of the model with parameter optimization and with the use of Top-20 group through the extraction of derived variables. The parameter optimization results of the Top-20 group with the best performance are as follows: eta (learning rate) = 0.10, gamma = 0, max_depth = 4, and min_child_weight = 1. As different classification models, there are Decision Tree, Random Forest, Support Vector Machine, and k-Nearest Neighbor. As evaluation indexes, Accuracy, Precision, Recall, and F1-score are used. They can be obtained from a confusion matrix. According to the evaluation, the XGBoost model proposed in this study had 85.61% accuracy and 79.28% F1-score. Compared to other classification models, the proposed model showed the best performance. It proves that the XGBoost based on derived variable extraction can generate effective performance for dementia risk prediction.

The purpose of this study is to apply a XGBoost model, which shows strong performance for classification prediction, to OASIS dementia data and thereby to proposed a more effective model of dementia risk prediction. For performance improvement, derived variables were extracted and hyper-parameters were optimized. Given the drawn results in terms of performance evaluation indexes, the proposed method effectively predicted dementia risk. This study has the following limitations: Firstly, all the hyper-parameters of XGBoost failed to be optimized. Secondly, the size of OASIS data set was not large enough. For this reason, there is a possibility to increase performance. If these limitations are overcome in a future study, it will be possible to improve the performance of the proposed dementia risk prediction model.

REFERENCES

[1] L. Jia, M. Quan, Y. Fu, T. Zhao, Y. Li, C. Wei, Y. Tang, Q. Qin, F. Wang, Y. Qiao, S. Shi, Y.-J. Wang, Y. Du, J. Zhang, J. Zhang, B. Luo, Q. Qu, C. Zhou, S. Gauthier, and J. Jia, ''Dementia in China: Epidemiology, clinical management, and research advances,'' *Lancet Neurol.*, vol. 19, no. 1, pp. 81–92, Jan. 2020.

[2] J.-W. Baek and K. Chung, ''Context deep neural network model for predicting depression risk using multiple regression,'' *IEEE Access*, vol. 8, pp. 18171–18181, Jan. 2020.

[3] *Statistics Korea*. Accessed: Jul. 21, 2020. [Online]. Available: http://kostat.go.kr/

[4] United Nations. *World Population Prospects*. Accessed: Jul. 21, 2020. [Online]. Available: https://population.un.org/wpp/

[5] *National Health Insurance Corporation*. Accessed: Jul. 21, 2020. [Online]. Available: https://www.nhis.or.kr/

[6] *Ministry of Health and Welfare, Central Dementia Center*. Accessed: Jul. 21, 2020. [Online]. Available: https://www.nid.or.kr/

[7] G. Lombardi, G. Crescioli, E. Cavedo, E. Lucenteforte, G. Casazza, A.-G. Bellatorre, C. Lista, G. Costantino, G. Frisoni, G. Virgili, and G. Filippini, ''Structural magnetic resonance imaging for the early diagnosis of dementia due to Alzheimer's disease in people with mild cognitive impairment,'' *Cochrane Database Systematic Rev.*, pp. 6–14, Mar. 2020.

[8] C.-M. Kim, E.-J. Hong, K. Chung, and R.-C. Park, ''Driver facial expression analysis using LFA-CRNN-based feature extraction for health-risk decisions,'' *Appl. Sci.*, vol. 10, no. 8, pp. 2956–2976, Apr. 2020.

[9] K. Chung, H. Yoo, and D.-E. Choe, ''Ambient context-based modeling for health risk assessment using deep neural network,'' *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 4, pp. 1387–1395, Apr. 2020.

[10] I. E. van de Vorst, N. M. Golüke, I. Vaartjes, M. L. Bots, and H. L. Koek, ''A prediction model for one-and three-year mortality in dementia: results from a nationwide hospital-based cohort of 50,993 patients in The Netherlands,'' *Age Ageing*, vol. 49, no. 3, pp. 361–367, May 2020.

[11] Z. B. Miled, K. Haas, C. M. Black, R. K. Khandker, V. Chandrasekaran, R. Lipton, and M. A. Boustani, ''Predicting dementia with routine care EMR data,'' *Artif. Intell. Med.*, vol. 102, Jan. 2020, Art. no. 101771, doi: 10.1016/j.artmed.2019.101771.

[12] T. Chen and C. Guestrin, ''Xgboost: A scalable tree boosting system,'' in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[13] D. Nielsen, ''Tree boosting with XGboost—Why does XGboost win 'every' machine learning competition,'' NTNU, Trondheim, Norway, Tech. Rep., Dec. 2016, pp. 33–44.

[14] J.-C. Kim and K. Chung, ''Neural-network based adaptive context prediction model for ambient intelligence,'' *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 4, pp. 1451–1458, Apr. 2020.

[15] B. Yu, W. Qiu, C. Chen, A. Ma, J. Jiang, H. Zhou, and Q. Ma, ''SubMito-XGBoost: Predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting,'' *Bioinformatics*, vol. 36, pp. 1074–1081, Oct. 2019.

[16] D. W. Hah, Y. M. Kim, and J. J. Ahn, ''A study on KOSPI 200 direction forecasting using XGBoost model,'' *J. Korean Data Inf. Sci. Soc.*, vol. 30, no. 3, pp. 655–669, May 2019.

[17] K. Chung and H. Jung, ''Knowledge-based dynamic cluster model for healthcare management using a convolutional neural network,'' *Inf. Technol. Manage.*, vol. 21, no. 1, pp. 41–50, Mar. 2020.

[18] A. J. Mitchell, ''The mini-mental state examination (MMSE): Update on its diagnostic accuracy and clinical utility for cognitive disorders,'' in *Cognitive Screening Instruments*. Cham, Switzerland: Springer, 2017, pp. 37–48.

[19] H. M. T. Ullah, Z. Onik, R. Islam, and D. Nandi, ''Alzheimer's disease and dementia detection from 3D brain MRI data using deep convolutional neural networks,'' in *Proc. 3rd Int. Conf. Converg. Technol. (I2CT)*, Pune, India, Apr. 2018, pp. 1–3.

[20] J. Islam and Y. Zhang, ''A novel deep learning based multi-class classification method for Alzheimer's disease detection using brain MRI data,'' in *Proc. Int. Conf. Brain Inform.*, vol. 10654, Nov. 2017, pp. 213–222.

[21] A. Veeramuthu, S. Meenakshi, and K. Ashok Kumar, ''A neural network based deep learning approach for efficient segmentation of brain tumor medical image data,'' *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4227–4234, May 2019.

[22] J.-C. Kim and K. Chung, ''Knowledge-based hybrid decision model using neural network for nutrition management,'' *Inf. Technol. Manage.*, vol. 21, no. 1, pp. 29–39, Mar. 2020.

[23] A. Manandhar, S. Gautam, D. K. Shrestha, S. Sauden, and D. R. Pant, ''Identifying dementia in MRI scans using artificial neural network and K-nearest neighbor,'' *Zerone Scholar*, vol. 1, no. 1, pp. 22–25, Dec. 2016.

[24] J. Tohka, E. Moradi, and H. Huttunen, ''Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia,'' *Neuroinformatics*, vol. 14, pp. 279–296, Jan. 2016.

[25] K. Chung and R. C. Park, ''P2P-based open health cloud for medicine management,'' *Peer–Peer Netw. Appl.*, vol. 13, pp. 610–622, Mar. 2020.

[26] K. Chung and H. Yoo, ''Edge computing health model using P2P-based deep neural networks,'' *Peer–Peer Netw. Appl.*, vol. 13, no. 2, pp. 694–703, Mar. 2020.

[27] *Open Access Series of Imaging Studies (OASIS)*. Accessed: Jul. 21, 2020. [Online]. Available: http://www.oasis-brains.org/

[28] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, ''Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults,'' *J. Cognit. Neurosci.*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007.

[29] J. Ashburner, ''Symmetric diffeomorphic modeling of longitudinal structural MRI,'' *Frontiers Neurosci.*, vol. 6, p. 197, Feb. 2013.

[30] C. E. Coffey, G. Ratcliff, J. A. Saxton, R. N. Bryan, L. P. Fried, and J. F. Lucke, ''Cognitive correlates of human brain aging: A quantitative magnetic resonance imaging investigation,'' *J. Neuropsychiatry Clin. Neurosci.*, vol. 13, pp. 471–485, Nov. 2001.

[31] S. Park, M. Ji, and J. Chun, ''2D human pose estimation based on object detection using RGB-D information,'' *KSII Trans. Internet Inf. Syst., South Korea*, vol. 12, pp. 800–816, Feb. 2018.

[32] Y. K. Jain and S. K. Bhandare, ''Min max normalization based data perturbation method for privacy protection,'' *Int. J. Comput. Commun. Technol.*, vol. 2, no. 8, pp. 45–50, Oct. 2011.

[33] V. Safak, ''Min-Mid-Max scaling, limits of agreement, and agreement score,'' 2020, *arXiv:2006.12904*. [Online]. Available: http://arxiv.org/abs/2006.12904

[34] S. Bhattacharya, S. R. K. S, P. K. R. Maddikunta, R. Kaluri, S. Singh, T. R. Gadekallu, M. Alazab, and U. Tariq, ''A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU,'' *Electronics*, vol. 9, no. 2, p. 219, Jan. 2020.

[35] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. Mohammadian, ''Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis,'' *Accident Anal. Prevention*, vol. 136, Mar. 2020, Art. no. 105405.

[36] W. Kim and J. Chun, ''An improved approach for 3D hand pose estimation based on a single depth image and haar random forest,'' *KSII Trans. Internet Inf. Syst., South Korea*, vol. 9, no. 8, Aug. 2015.

[37] H. Kim and J. Chun, ''A SCORM-based e-learning process control model and its modeling system,'' *KSII Trans. Internet Inf. Syst., South Korea*, vol. 5, no. 11, Nov. 2011.

[38] C.-S. Pan, C.-W. Wang, M.-H. Tsai, C.-B. Kuo, and C.-H. Kuo, ''Classification of dementia based on over-sampling approach and decision tree,'' in *Proc. Int. Conf. Big Data Edu. (ICBDE)*, 2018, pp. 1–4.

[39] M. Dauwan, J. J. Zande, E. Dellen, I. E. C. Sommer, P. Scheltens, A. W. Lemstra, and C. J. Stam, ''Random forest to differentiate dementia with lewy bodies from Alzheimer's disease,'' *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitor.*, vol. 4, no. 1, pp. 99–106, Jan. 2016.

[40] D. Stamate, ''A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical-Information Framework for Alzheimer disease biomarker discovery cohort,'' *Alzheimer's & Dementia: Transl. Res. Clin. Intervent.*, vol. 5, no. 1, pp. 933–938, Jan. 2019.

SEONG-EUN RYU is currently pursuing the bachelor's degree with the Division of Computer Science and Engineering, Kyonggi University, Suwon, South Korea. He has been a Researcher with the Data Mining Laboratory, Kyonggi University. His research interests include data mining, artificial intelligent, healthcare, biomedical and health informatics, knowledge systems, VR/AR, and deep learning.

DONG-HOON SHIN received the B.S. degree from the Department of Computer Engineering, Dongseo University, South Korea, in 2019. He is currently pursuing the master's degree with the Department of Computer Science, Kyonggi University, Suwon, South Korea. He has been a Researcher with the Data Mining Laboratory, Kyonggi University. His research interests include data mining, artificial intelligent, healthcare, biomedical and health informatics, knowledge systems, VR/AR, and deep learning.

KYUNGYONG CHUNG received the B.S., M.S., and Ph.D. degrees from the Department of Computer Information Engineering, Inha University, South Korea, in 2000, 2002, and 2005, respectively. He has worked with the Software Technology Leading Department, Korea IT Industry Promotion Agency (KIPA). From 2006 to 2016, he was a Professor with the School of Computer Information Engineering, Sangji University, South Korea. Since 2017, he has been a Professor with the Division of Computer Science and Engineering, Kyonggi University, Suwon, South Korea. His research interests include data mining, artificial intelligent, healthcare, biomedical and health informatics, knowledge systems, HCI, and recommendation systems. He was named as a 2017 Highly Cited Researcher by Clarivate Analytics.