Journal of Biomedical and Translational Research
pISSN 2508-1357 eISSN 2508-139X

Original Article
J Biomed Transl Res 2022;23(1):17-28
https://doi.org/10.12729/jbtr.2022.23.1.17

Machine learning-based risk factor analysis for periodontal disease from a Korean National Survey

Ho Sun Shon[1], Eun Sun Choi[2], Yan-Sub Cho[3], Eun Jong Cha[4], Tae-Geon Kang[5*], Kyung Ah Kim[4*]

[1]*Medical Research Institute, School of Medicine, Chungbuk National University, Cheongju 28644, Korea*
[2]*Department of Big Data Cooperative Course, Chungbuk National University, Cheongju 28644, Korea*
[3]*Department of Management Information Systems, Chungbuk National University, Cheongju 28644, Korea*
[4]*Department of Biomedical Engineering, School of Medicine, Chungbuk National University, Cheongju 28644, Korea*
[5]*Institute for Trauma Research, College of Medicine, Korea University, Seoul 02841, Korea*

Received: Mar 10, 2022
Revised: Mar 21, 2022
Accepted: Mar 21, 2022

***Corresponding author**
Tae-Geon Kang
Institute for Trauma Research, College of Medicine, Korea University, Seoul 02841, Korea
Tel: +82-2-2626-2473
E-mail: kangtg@kumc.co.kr

Kyung Ah Kim
Department of Biomedical Engineering, School of Medicine, Chungbuk National University, Cheongju 28644, Korea
Tel: +82-43-261-2852
E-mail: kimka@chungbuk.ac.kr

ORCID
Ho Sun Shon
http://orcid.org/0000-0002-6717-7869
Eun Sun Choi
https://orcid.org/0000-0001-8130-1128
Yan-Sub Cho
https://orcid.org/0000-0002-4395-1979
Eun Jong Cha
http://orcid.org/0000-0002-8554-4132
Tae-Geon Kang
http://orcid.org/0000-0002-8575-0120

Abstract

Periodontal disease is a chronic but treatable condition which often does not cause pain during the initial stages of the illness. Lack of awareness of symptoms can delay initiation of treatment and worsen health. The aim of this study was to develop and compare different risk prediction models for periodontal disease using machine learning algorithms. We obtained information on risk factors for periodontal disease from the Korea National Health and Nutrition Examination Survey (KNHANES) dataset. Principal component analysis and an auto-encoder were used to extract data on risk factors for periodontal disease. A synthetic minority oversampling technique algorithm was used to solve the problem of data imbalance. We used a combination of logistic regression analysis, support vector machine (SVM) learning, random forest, and AdaBoost to classify and compare risk prediction models for periodontal disease. In cases where we used principal component analysis (PCA) to extract risk factors, the recall was higher than the feature selection method in the logistic regression and support-vector machine learning models. AdaBoost's recall was 0.98, showing the highest performance of both feature selection and PCA. The F1 score showed relatively high performance in AdaBoost, logistic regression, and SVM learning models. By using the risk factors extracted from the research results and the predictive model based on machine learning, it will be able to help in the prevention and diagnosis of periodontal disease, and it will be used to study the relationship with various diseases related to periodontal disease.

Keywords: periodontal disease; risk factors; feature extraction; machine learning; prediction model

INTRODUCTION

Periodontal disease is a chronic condition characterized by alveolar bone loss which affects more than 30% of Korean adults aged 30 years or older. According to the 2020 Health Insur-

Kyung Ah Kim
http://orcid.org/0000-0002-8814-6973

Conflict of Interest
No potential conflict of interest relevant to this article was reported.

Acknowledgements
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2019R1F1A1051569 and 2020R1I1A1A01065199).

Ethics Approval
Not applicable.

ance Review and Assessment Service frequent disease statistics, periodontal disease was the main reason for visiting the out-patient department and ranked first in terms of the total cost of medical care benefits. If periodontal disease is not treated in time, it can lead to tooth loss by exacerbating inflammation of the surrounding supportive tissue [1]. Periodontal disease is caused by bacterial infection, with the development and progression of plaque formation and inflammation being dependent on multiple individual, genetic, social, and environmental factors [2]. Periodontal disease is one of the factors directly related to quality of life, and it has been suggested that the risk of quality-of-life deterioration is 1.32 times higher in the periodontal disease group than in the healthy periodontal group [3]. In uncontrolled periodontal disease inflammation can become chronic, in turn increasing risk for other illnesses such as heart disease, ocular disorders, and pulmonary fibrosis [4–7]. Several studies have examined the socio-economic and population-based factors thought to influence risk for periodontal disease in Korean adults [8–10].

The Korea National Health and Nutrition Examination Survey (KNHANES) is a nationwide health and nutrition survey conducted in accordance with Article 16 of the National Health Promotion Act. For the KNHANES, data were collected through demographic factors, medical examination results, and nutritional intake. Many studies using KNHANES data to examine periodontal disease in the Korean population have also been conducted. In a study using data from the 4th KNHANES, a complex sample logistic regression analysis was performed to confirm the effects of different predictors on the stage-related prevalence of periodontal disease. Efforts to resolve social and economic inequalities have also been made [11]. Research has also shown that the metabolic syndrome is associated with an elevated risk for periodontal disease, which increases in parallel with the number of metabolic syndrome features present [12]. Therefore, the severity of dyslipidemia, including low high-density lipoprotein (HDL), high low-density lipoprotein (LDL) cholesterol, and elevated triglycerides, is correlated with the severity of periodontal disease [13, 14]. In a study comparing the prevalence of periodontal disease based on oral health behaviors using data from the 6th KNHANES 3rd year (2015), statistically significant differences were found in periodontal disease treatment, tooth brushing time, dental floss, interdental brush use and subjective oral health status [15]. In one study, age, sex, body mass index (BMI), and lipid profiles were examined in 45 patients with periodontitis. Total cholesterol and triglyceride levels were higher in the periodontitis group, and HDL and LDL cholesterol levels were high in the case group. However, these differences were not statistically significant [16]. Other studies using multiple logistic regression analysis have identified hypertension, but not type II diabetes mellitus, heart disease, dyslipidemia, or anemia, as a risk factor for periodontal disease [17, 18]. In studies using decision trees, logistic regression analysis, and artificial neural networks to develop prediction models for periodontal disease, decision trees have shown the highest accuracy [19, 20]. To solve the class imbalance of categorical variables, machine learning algorithms have been used, and various methods have been used to evaluate the performance of predictive models [21–23].

Most of these previous studies dealt only with the relationship between some risk factors and periodontal disease, and studies that applied various machine learning algorithms to build

https://doi.org/10.12729/jbtr.2022.23.1.17

predictive models are insufficient. In response to this knowledge gap, the aim of this study was to develop and compare different risk prediction models for periodontal disease using machine learning algorithms based on data obtained from the 6th KNHANES survey [24]. The ultimate goal of this study was to improve the quality of care and reduce healthcare costs for patients with periodontal disease through improved detection and treatment of important clinical risk factors.

MATERIALS AND METHODS

Material

We extracted raw data from the 6th KNHANES, conducted from 2013 to 2015. The total number of households surveyed was 22,948, and we considered a total of 655 variables of interest. The analysis was performed using data from 11,102 participants who completed pre-processing.

Feature selection

In this study, we selected demographic, health behavior-related, and oral-health related variables associated with periodontal disease in previous studies [25].

Demographic characteristics included area of residence, sex, age, private health insurance subscription, household income, number of household members, basic livelihood benefits, housing ownership, marital status, health insurance type, education level, and economic activity status. Health behavior-related characteristics included a history of hypertension, dyslipidemia, stroke, myocardial infarction, angina pectoris, type II diabetes mellitus, obesity, subjective health status, history of health check-ups, subjective body type recognition, lifetime history of alcohol use, lifetime smoking status, average hours of sleep per day, systolic blood pressure, diastolic blood pressure, BMI, serum fasting blood glucose, glycated hemoglobin, lipid profiles, aspartate aminotransferase, alanine aminotransferase, and leukocyte levels, as well as urinalysis for proteins, glucose, ketones, and bilirubin [26, 27].

For oral health characteristics, we considered the use of permanent dental caries, perceived oral health status, experience of toothache over the past year, orthodontic treatment experience, chewing problems, complaints of chewing discomfort, speaking problems, and daily dental hygiene habits. In keeping with other studies, we selected the number of brushing times and the number of oral products used as variables of interest.

Analysis method

Fig. 1 summarizes the overall process used for the classification and analysis of the risk factors for periodontal disease extracted from the database. To optimize statistical analysis of the data, outliers and missing values were removed. The refined dataset that was subjected to pre-processing process included 11,102 samples and 54 data sets, which together examined 15 numerical variables of interest, including average sleep time, systolic blood pressure, and BMI, as well as 39 categorical variables of interest, including gender income, and education

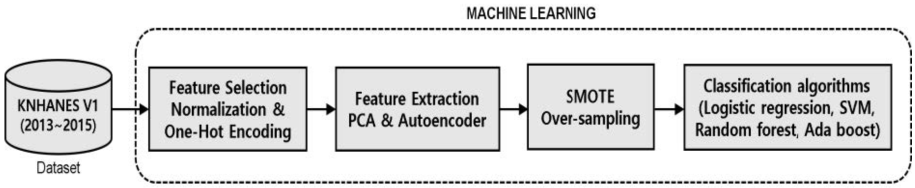
MACHINE LEARNING
KNHANES V1 (2013~2015)
Dataset
Feature Selection Normalization & One-Hot Encoding
Feature Extraction PCA & Autoencoder
SMOTE Over-sampling
Classification algorithms (Logistic regression, SVM, Random forest, Ada boost)

Fig. 1. Overall process of periodontal disease analysis.

level. MinMaxScaler was used to normalize numerical data. In the AdaBoost model for categorical data, one-hot encoding was performed to apply a deep learning model.

The periodontal disease risk factor data were extracted using principal component analysis (PCA) and an auto-encoder. Variable extraction is a technique for creating new variables by combining existing ones. It is different from the random selection of isolated variables of interest. PCA is a technique that converts a new variable from a high-dimensional variable into a low-dimensional variable without linear correlation by linearly combining the data while preserving the variance as much as possible.

An auto-encoder is a kind of artificial network used for unsupervised machine learning. It consists of two parts: an encoder that transforms the input into an internal representation, and a decoder that transforms the internal representation into an output. Except for the number of neurons in the input and output layers being the same, an autoencoder (AE) has the same structure as a general multilayer perceptron.

Autoencoder reduces the dimension by compressing data by making the number of neurons in the input and output layers the same and making the number of neurons in the hidden layer smaller than the input layer. Stacked autoencoder forms a network by stacking as many stacks as the number determined by the designer in advance. The network formed in this way can extract important features from the input data. In this study, a stacked auto-encoder, i.e., one with several symmetrical hidden layers, was used. A layered auto-encoder has the advantage of being able to learn more complex features by adding layers. In addition, we used a synthetic minority oversampling technique (SMOTE) algorithm to solve the data imbalance. SMOTE is a commonly used oversampling technique for generating synthetic data. Oversampling was applied to adjust the unbalanced data so that the ratio of the data was 1:1. As a result of oversampling, 7,549 cases (50.38%) with periodontal disease and 7,434 cases (49.62%) without periodontal disease were corrected. To implement the periodontal disease classification model, data for learning were divided into 70%, and data for verification were divided into 30%. The learning data consisted of 50.15% cases without periodontal disease and 49.85% cases with periodontal disease, and the validation data consisted of 50.93% cases without periodontal disease and 49.07% cases with periodontal disease.

We used a combination of logistic regression analysis, support vector machine (SVM) learning, random forest, and AdaBoost to classify risk prediction models for periodontal disease. Logistic regression analysis is a statistical technique in which the effects of multiple predictors on which a binary dependent outcome is estimated using a binomial or ordinal polynomial logistic function. The optimal model was backward elimination according to standard criteria.

SVM performs classification in the direction with the highest margin. The larger the margin, the higher the classification performance when new data not used for training are used for input. Extending the dimension of non-linear data has the advantages of linear separation, less impact on erroneous data, and less tendency to overfit data when compared to other models. Random forest is an ensemble machine learning model used for classification and regression analysis. It is a machine learning technique proposed by Breiman (2001) to randomly select an optimal reference variable. Random forest easily handles missing data and is effective for processing large amounts of data. In addition, model accuracy can be improved by avoiding overfitting of the data, and it is possible to select relatively important variables in the classification model. AdaBoost is a type of boosting algorithm that sequentially learns and predicts several weak learners, and extracts features that improve model performance during the training process by assigning weights to incorrectly predicted data to improve errors. In the learning process, each sample is weighted, and the processing speed is increased by adapting previously unprocessed samples to the next learning phase through modelling. To evaluate the test data, the data were divided in a 70:30 ratio, with 7,771 cases in the training set and 3,331 cases in the test set. The results derived from machine learning were evaluated based on the accuracy, precision, recall, and F1 score of the confusion matrix. Statistical analyses were performed using the R (version 4.0.5), Jupyter Notebook, and Python (version 4.1) software packages.

RESULTS

Characteristics of risk factors for periodontal disease

The demographic, health behavior-related, and oral health characteristics of the participants are shown in Tables 1 to 3 below. The demographic characteristics of participants with and without periodontal disease were compared (Table 1). Experiments with the chi-square test showed significant between group differences with respect to gender, age, education level, economic activity, and periodontal disease. Only a minority of participants had a known diagnosis of hypertension, dyslipidemia, stroke, myocardial infarction, angina pectoris, or type II diabetes mellitus. In total, 32.89% of participants were classified as obese. In total, 51.67% of the respondents considered their health to be “normal”, while 65.56% of the respondents had previously received a health checkup. In total, 41.11% of respondents reported normal subjective body type recognition, 89.24% of respondents had a lifetime history of alcohol consumption, 59.75% of respondents had never smoked, and 82.77% of respondents had HDL cholesterol levels within the normal range. Urinary protein, sugar, bilirubin, and ketones were absent in 93.15%, 96.71%, 96.24%, and 94.54% of patients, respectively.

The health behaviors of patients with and without periodontal disease were compared, as summarized in Table 2. There were significant between-group differences evident for all variables of interest, including blood pressure, hyperlipidemia, heart disease, type II diabetes mellitus, health check-up status, history of alcohol consumption, smoking history, and HDL cholesterol level.

Regarding the oral characteristics of the study participants, 41.79% of the respondents

Table 1. Significance test between demographic characteristics and periodontal disease using Chi-square

Demographic characteristics	Division	Periodontal disease				Chi-square (*p*-value)
		No	%	Yes	%	
Total		7,578	68.26	3,524	31.74	
Gender	Male	3,048	40.22	1,960	55.62	229.68 (< 0.001)
	Female	4,530	59.78	1,564	44.38	
Age group	20–39	3,016	39.8	392	11.12	1,001.8 (< 0.001)
	40–64	3,457	45.62	2,112	59.93	
	65 ≤	1,105	14.58	1,020	28.94	
Basic living allowance	Yes	426	5.62	256	7.26	10.978 (< 0.001)
	No	7,152	94.38	3,268	92.74	
Education	Elementary school	893	11.78	835	23.69	587.75 (< 0.001)
	Middle school	756	9.98	549	15.58	
	High school	2,164	28.56	1,178	33.43	
	College/university	3,336	44.02	833	23.64	
	Graduate school	429	5.66	129	3.66	
Economic activity status	Yes	4,660	61.49	2,262	64.19	7.3258 (0.006)
	No	2,918	38.51	1,262	35.81	

Table 2. Significance test between health behavior and periodontal disease

Characteristics related to health	Division	Periodontal disease				Chi-square (*p*-value)
		No	%	Yes	%	
Total		7,578	68.26	3,524	31.74	
High blood pressure	No	6,365	83.99	2,454	69.64	302.62 (< 0.001)
	Yes	1,213	16.01	1,070	30.36	
Dyslipidemia	No	6,570	86.70	2,892	82.07	40.634 (< 0.001)
	Yes	1,008	13.30	632	17.93	
Myocardial infarction	No	7,541	99.51	3,477	98.67	21.786 (< 0.001)
	Yes	37	0.49	47	1.33	
Angina	No	7,479	98.69	3,425	97.19	30.165 (< 0.001)
	Yes	99	1.31	99	2.81	
Diabetes	No	7,149	94.34	3,086	87.57	152.1 (< 0.001)
	Yes	429	5.66	438	12.43	
Health check-up	Yes	4,847	63.96	2,432	69.01	26.958 (< 0.001)
	No	2,731	36.04	1,092	30.99	
Drinking	No	751	9.91	444	12.60	17.83 (< 0.001)
	Yes	6,827	90.09	3,080	87.40	
Smoking	≤ 99	210	2.77	51	1.45	328.5 (< 0.001)
	100 ≤	2,444	32.25	1,764	50.06	
	No	4,924	64.98	1,709	48.50	
HDL	No	6,479	85.50	2,710	76.90	124.03 (< 0.001)
	Yes	1,099	14.50	814	23.10	
Urine glucose	Negative	7,398	97.62	3,339	94.75	71.873 (< 0.001)
	Trace	49	0.65	32	0.91	
	Positive (+)	52	0.69	58	1.65	
	Positive (++)	39	0.51	59	1.67	
	Positive (+++)	40	0.53	36	1.02	

HDL, high-density lipoprotein.

described their perceived oral status as "normal", 61.32% of the respondents had not experienced toothache over the past year, and 94.51% of the respondents did not receive orthodontic treatment or have chewing problems. Int total, 36.63% of the respondents said that they were "not at all uncomfortable" with their health. In this subgroup, 98.93% had brushed their teeth the previous day, 68.26% said they had no history of periodontal disease, and 72.01% said they had no permanent dental caries.

The oral characteristics of the study participants with and without periodontal disease were compared, as shown in Table 3. Variables such as oral health status, toothache experience in the past year, chewing problems, speech problems, brushing yesterday, and permanent tooth decay were significantly associated with the presence or absence of periodontal disease.

Performance comparison of classification model

Following the extraction of significant variables related to periodontal disease, the classification performance of the model was predicted using a machine learning algorithm. PCA revealed four main components which explained 60.23% of the total data (Table 4). The first principal component had a strong relationship with the region, and the second principal component was related to education level and chewing problems. The third principal component was related to chewing problems, BMI, and ALT (SGPT), and the fourth principal component

Table 3. Significance test between oral characteristics and periodontal disease

Oral health characteristics	Division	Periodontal disease				Chi-square (*p*-value)
		No	%	Yes	%	
Total		7,578	68.26	3,524	31.74	
Oral health status	Very good	101	1.33	43	1.22	381.67 (< 0.001)
	Good	1,142	15.07	347	9.85	
	Ordinary	3,459	45.65	1,180	33.48	
	Bad	2,408	31.78	1,443	40.95	
	Very bad	468	6.18	511	14.50	
Toothache experience over the past year	No	4,870	64.26	1,938	54.99	86.772 (< 0.001)
	Yes	2,708	35.74	1,586	45.01	
Chewing problem	Very uncomfortable	251	3.31	264	7.49	569.22 (< 0.001)
	Uncomfortable	1,008	13.30	909	25.79	
	Ordinary	1,183	15.61	735	20.86	
	Not uncomfortable	1,935	25.53	750	21.28	
	Not uncomfortable at all	3,201	42.24	866	24.57	
Speaking problem	Very uncomfortable	85	1.12	89	2.53	355.08 (< 0.001)
	Uncomfortable	389	5.13	388	11.01	
	Ordinary	707	9.33	561	15.92	
	Not uncomfortable	1,346	17.76	717	20.35	
	Not uncomfortable at all	5,051	66.65	1,769	50.20	
Brushing teeth yesterday	No	69	0.91	50	1.42	5.3916 (< 0.001)
	Yes	7,509	99.09	3,474	98.58	
Permanent tooth decay	No	5,578	73.61	2,416	68.56	30.174 (< 0.001)
	Yes	2,000	26.39	1,108	31.44	

Table 4. Importance of components by principal component analysis

	PC1	PC2	PC3	PC4
Standard deviation	4.53350	2.07233	1.68388	1.32427
Proportion of variance	0.42050	0.08787	0.05801	0.03588
Cumulative proportion	0.42050	0.50837	0.56639	0.60227

was related to chewing problems and the number of household members. The performance of the classification models was compared according to the presence or absence of periodontal disease (Table 5). Based on the classification performance results, the accuracy, precision, and recall of the logistic regression model were 0.73, 0.77, and 0.87, respectively, and the F1 score was 0.82. The model with the number of dimensions reduced using PCA was derived to have an accuracy, precision, and recall of 0.7, 0.73, and 0.89, respectively, and the F1 score was 0.8. The recall thus increased when all significant variables were included in the model. In the model that resolved the class imbalance using PCA and SMOTE, accuracy, precision, and recall were 0.67, 0.68, and 0.68, respectively, and the F1 score was 0.68. In the model reduced by applying the auto-encoder, accuracy, precision, and recall were 0.62, 0.68, and 0.84, respectively, and the F1 score was 0.75. When the class imbalance was resolved using an auto-encoder and SMOTE, the accuracy, precision, and recall of the model were 0.50, 0.51, and 0.54, respectively, and the F1 score was 0.52.

The accuracy, precision and recall of the SVM model were 0.71, 0.75, and 0.86, respectively, and the F1 score was 0.8. In particular, the accuracies of the model with reduced dimensions obtained by applying PCA were 0.71, 0.73, and 0.93, respectively, and the F1 score was

Table 5. Performance evaluation by classification model

Machine learning	Evaluation	Feature selection	PCA	PCA + SMOTE	AE	AE + SMOTE
Logistic regression	Accuracy	0.73	0.70	0.67	0.62	0.50
	Precision	0.77	0.73	0.68	0.68	0.51
	Recall	0.87	0.89	0.68	0.84	0.54
	F1 score	0.82	0.80	0.68	0.75	0.52
SVM	Accuracy	0.71	0.71	0.61	0.61	0.50
	Precision	0.75	0.73	0.60	0.68	0.51
	Recall	0.86	0.93	0.70	0.82	0.68
	F1 score	0.80	0.81	0.65	0.74	0.58
Random forest	Accuracy	0.72	0.70	0.64	0.59	0.49
	Precision	0.75	0.75	0.61	0.67	0.50
	Recall	0.90	0.85	0.81	0.76	0.73
	F1 score	0.82	0.79	0.70	0.72	0.60
AdaBoost	Accuracy	0.70	0.70	0.68	0.61	0.49
	Precision	0.70	0.72	0.65	0.32	0.48
	Recall	0.98	0.94	0.72	0.19	0.43
	F1 score	0.82	0.81	0.69	0.24	0.46

PCA, principal component analysis; SMOTE, synthetic minority oversampling technique; AE, autoencoder; SVM, support vector machine.

0.81. The recall and F1 scores of the model using PCA were more accurate compared to those using all selected variables. The accuracy, precision, and recall of the model that resolved the class imbalance using PCA and SMOTE were 0.61, 0.60, and 0.70, respectively, and the F1 score was 0.65. The accuracy, precision, and recall of the model reduced using the auto-encoder were 0.61, 0.68, and 0.82, respectively, and the F1 score was 0.74. The accuracy, precision, and recall of the model that solved the class imbalance using an auto-encoder and SMOTE were 0.50, 0.51, and 0.68, respectively, and the F1 score was 0.58.

In the random forest model, the accuracy, precision, and recall were 0.72, 0.75, and 0.90, respectively, and the F1 score was 0.82. The accuracy, precision, and recall of the reduced-dimensional model by applying PCA were 0.70, 0.75, and 0.85, respectively, and the F1 score was 0.79. The accuracy, precision, and recall of the model that resolved the class imbalance using PCA and SMOTE were 0.64, 0.61, and 0.81, respectively, and the F1 score was 0.7. The accuracies of the dimensionally reduced model obtained by applying the auto-encoder were 0.59, 0.67, and 0.76 respectively, and the F1 score was 0.72. The accuracy, precision, and recall of the model that resolved the class imbalance using the auto-encoder and SMOTE were 0.49, 0.50, and 0.73, respectively, and the F1 score was 0.60.

The accuracy, precision, and recall of the AdaBoost model using the extracted variables were 0.70, 0.70, and 0.98, respectively, and the F1 score was 0.82. The accuracy, precision, and recall of the model by applying PCA were 0.70, 0.72, and 0.94, respectively, and the F1 score was 0.81. The accuracy, precision, and recall of the model that resolved the class imbalance using PCA and SMOTE were 0.68, 0.65, and 0.72, respectively, and the F1 score was 0.69. The accuracy, precision, and recall of the auto-encoder-applied model were 0.61, 0.32, and 0.19, respectively, and the F1 score was 0.24. The accuracies of the model to which the auto-encoder and SMOTE were applied were 0.49, 0.48, and 0.43, respectively, and the F1 score was 0.46.

Among the models used, the model with the highest accuracy was the one to which variable selection was applied. In the case of the model using PCA, recall increased in the model to which logistic regression analysis and SVM were applied. In the case of SVM, both recall and F1 scores showed the highest performance. Additionally, the model in which PCA was applied to AdaBoost was evaluated as having the highest precision.

DISCUSSION

In this study, we extracted data on risk factors for periodontal disease from the 6th KNHANES, and developed risk prediction models for periodontal disease using logistic regression, SVM, random forest, and AdaBoost algorithms. We performed dimensionality reduction using PCA and a stacked auto-encoder, and data imbalance was dealt with using the SMOTE oversampling technique to improve the performance of the model. However, feature extraction using a deep-learning-based auto-encoder showed a lower performance compared to other feature extraction techniques. This was related to oversampling owing to the class imbalance problem.

In this study, the risk prediction models for periodontal disease were evaluated based on a confusion matrix for a comparative analysis of machine learning algorithms. Logistic regression analysis provided the model with the greatest accuracy and precision, while AdaBoost provided the model with the highest recall value. The F1 score value increased in the model to which PCA was applied compared to the model to which variable selection was applied using SVM. Based on these results, it will be possible to use feature extraction techniques and machine learning algorithms to analyze various medical data and help medical professionals prevent and predict periodontal disease.

Although various studies on periodontal disease have been conducted, they are hampered by problems with data quality, such as poor data pre-processing or an imbalance in the target data. Because these problems degrade the performance of risk classification and prediction models, there is a need for various methods to purify and argue data. In addition, a more systematic method is required for the extraction of risk factors to develop risk prediction models. To solve this problem, we are conducting research on classification prediction models, such as feature extraction and multi-layer perceptron, using deep learning algorithms.

There are two difficulties in analyzing data on periodontal disease obtained from the KNHANES. First, there is the problem of class imbalance in the dataset. Although this imbalance was resolved by applying oversampling using the SMOTE technique, the performance of the model indeed deteriorated. In the process of oversampling data, we think underfitting might be involved, because it fails to learn useful data that affects risk prediction in a meaningful way. In future research, instead of SMOTE, a type of generation model called an auto-encoder will be applied to generate and apply a synthetic dataset close to the actual data. Second, it is difficult to reduce dimensionality using an auto-encoder. Although we scaled down the dimensions, we observed a problem with the dramatic drop in the performance of the model. In future research, we aim to apply a deep learning-based predictive model that designs input and output layers in a more diversified manner in order to compare, analyze, and improve model performance.

REFERENCES

1. Jeong MA, Kim JH. Association between cardiovascular disease and periodontal disease prevalence. J Korea Converg Soc 2011;2:47-52.
2. Nunn ME. Understanding the etiology of periodontitis: an overview of periodontal risk factors. Periodontololy 2000 2003;32:11-23.
3. Yu J, Hwang S. A convergence study on the effect of periodontal disease on health-related quality of life in adults over 40s. J Korea Converg Soc 2021;12:49-56.
4. Lee SJ, Hong YM, Back JH, Nam YS. The relationship between metabolic syndrome and periodontal disease except patients with hypertension and type 2 diabetes mellitus in Korea. J Korean Acad Oral Health 2009;33:451-460.
5. Shin HS, Ahn YS, Lim DS. Association between the number of existing permanent teeth and chronic obstructive pulmonary disease. J Dent Hyg Sci 2016;16:217-224.
6. Shin HE, Kim JH, Jung YS, Kim EK, Choi YH, Song KB. Relation between rheumatoid ar-

thritis and periodontal diseases: using the fifth Korea National Health and Nutrition Examination Survey. J Korean Acad Oral Health 2014;38:232-237.

7. Sim YH, Kim HL, Park HJ, Choi EY, Byun AR, Chun HJ, Shim KW, Lee HS, Lee SH. The association between periodontitis and chronic disease in Korean adult population. Korean J Fam Pract 2015;5:S726-S731.
8. Jung JO, Oh GJ. A study of the relationship between socioeconomic status, oral health behaviors and periodontitis in the elderly Korean population. J Korean Acad Oral Health 2011;35:57-66.
9. Choi JS, Lee YJ, Jeon SB, Kim HM, Jeong EH, Jo EJ. The association between self-reported oral health problems and oral health-related quality of life. J Korean Acad Oral Health 2010;34:411-421.
10. Won YS, Choi CH, Oh HN. Risk factors of periodontal disease in Korean adults. J Korean Acad Oral Health 2014;38:176-183.
11. Woo DH, You HY, Kim MJ, Kim HN, Kim JB, Jeong SH. Risk indicators of periodontal disease in Korean adults. J Korean Acad Oral Health 2013;37:95-102.
12. Kang HJ. A study on periodontal disease and tooth loss in metabolic syndrome patient. J Dent Hyg Sci 2015;15:445-456.
13. Lee S, Im A, Burm E, Ha M. Association between periodontitis and blood lipid levels in a Korean population. J Periodontol 2018;89:28-35.
14. Choi JS. Association between periodontitis and hypertriglyceridemia in Korean adults aged 30 and older: based on data from 2015 Korea National Health and Nutrition Examination Survey. J Korean Soc Dent Hyg 2020;20:53-62.
15. Lee SY, Lee YH. A convergence study of adults' oral health behaviors and periodontal disease. J Korea Converg Soc 2019;10:63-70.
16. Golpasand Hagh L, Zakavi F, Hajizadeh F, Saleki M. The association between hyperlipidemia and periodontal infection. Iran Red Crescent Med J 2014;16:e6577.
17. Kim Y. The association between periodontitis and systemic disease among Korean adults. J Korean Acad Oral Health 2016;40:244-249.
18. Kim HK, Choi KH, Lim SW, Rhee SW. Development of prediction model for prevalence of metabolic syndrome using data mining: Korea National Health and Nutrition Examination Study. J Digit Converg 2016;14:325-332.
19. Yoo SH, Park IS, Kim YM. A decision-tree analysis of influential factors and reasons for unmet dental care in Korean adults. Health Soc Welf Rev 2017;37:294-335.
20. Lee IJ, Lee J. Predictive of osteoporosis by tree-based machine learning model in post-menopause woman. J Radiol Sci Technol 2020;43:495-502.
21. Min B, Yoo J, Kim S, Shin D, Shin D. Network intrusion detection with one class anomaly detection model based on auto encoder. J Internet Comput Serv 2021;22:13-22.
22. Seo WH, Ma PS, Woo JH, Sun KH, Kim B, Kim BS. Data & knowledge-based anomaly detection of rotating machine using variational auto-encoder. Trans Korean Soc Noise Vib Eng 2021;31:289-297.
23. Kim H, Lee T. Stacked autoencoder based malware feature refinement technology research. J

Korea Inst Inf Sec Cryptol 2020;30:593-603.

24. Lee JH. The relationship between metabolic syndrome components and the number of remaining teeth in Korean adults. J Korean Acad Oral Health 2020;44:130-137.
25. Kang HJ. The convergence relationship between health behavior and cardiovascular disease and periodontitis. J Korea Converg Soc 2019;10:233-239.
26. Kim HD, Paik DM, Kho DH, Paik DI. Influence of cardiovascular related disease on periodontitis. J Korean Acad Dent Health 2006;30:46-55.
27. Cho YY. Association between periodontal disease, number of remaining teeth and high-sensitivity C-reactive. J Korean Soc Dent Hyg 2020;20:313-324.