

## 딥러닝을 활용한 상실치아 수 예측의 가능성: 파일럿 스터디

김선집<sup>1</sup>, 임도형<sup>2</sup>, 허정욱<sup>3</sup>, 조현재<sup>1</sup><sup>1</sup>서울대학교 치의학대학원 예방치학교실, <sup>2</sup>연세대학교 인지과학 협동과정, <sup>3</sup>하단 굿윌치과 병원Possibility of predicting missing teeth using deep learning:  
a pilot studySeon-Jip Kim<sup>1</sup>, Dohyoung Rim<sup>2</sup>, Jeong Uk Heo<sup>3</sup>, Hyun-Jae Cho<sup>1</sup><sup>1</sup>Department of Preventive Dentistry and Public Oral Health, School of Dentistry, Seoul National University, <sup>2</sup>Department of Cognitive Science, Yonsei University, Seoul, <sup>3</sup>Goodwill Dental Hospital at Hadan, Busan, Korea**Received:** November 23, 2019**Revised:** December 3, 2019**Accepted:** December 4, 2019**Corresponding Author:** Hyun-Jae ChoDepartment of Preventive Dentistry and  
Public Oral Health, School of Dentistry,  
Seoul National University, 101 Daehakro,  
Jongno-gu, Seoul 03080, Korea  
Tel: +82-2-740-8677

Fax: +82-2-765-1722

E-mail: stbluewi@snu.ac.kr

https://orcid.org/0000-0002-3079-8629

\*This research was supported by a program  
of data-science buildup from the Big Data  
Institute, Seoul National University and  
the ICT & Future Planning Program of  
the National Research Foundation (grant  
number: 2017R1C1B5017915).**Objectives:** The primary objective of this study was to determine if the number of missing teeth could be predicted by oral disease pathogens, and the secondary objective was to assess whether deep learning is a better way of predicting the number of missing teeth than multivariable linear regression (MLR).**Methods:** Data were collected through review of patient's initial medical records. A total of 960 participants were cross-sectionally surveyed. MLR analysis was performed to assess the relationship between the number of missing teeth and the results of real-time PCR assay (done for quantification of 11 oral disease pathogens). A convolutional neural network (CNN) was used as the deep learning model and compared with MLR models. Each model was performed five times to generate an average accuracy rate and mean square error (MSE). The accuracy of predicting the number of missing teeth was evaluated and compared between the CNN and MLR methods.**Results:** Model 1 had the demographic information necessary for the prediction of periodontal diseases in addition to the red and the orange complex bacteria that are highly predominant in oral diseases. The accuracy of the convolutional neural network in this model was 65.0%. However, applying Model 4, which added yellow complex bacteria to the total bacterial load, increased the expected extractions of dental caries to 70.2%.

On the other hand, the accuracy of the MLR was about 50.0% in all models. The mean square error of the CNN was considerably smaller than that of the MLR, resulting in better predictability.

**Conclusions:** Oral disease pathogens can be used as a predictor of missing teeth and deep learning can be a more accurate analysis method to predict the number of missing teeth as compared to MLR.**Key Words:** Deep learning, Linear regression, Missing teeth, Real-time PCR, Periodontitis

## Introduction

Periodontal disease is a major oral disease, which threatens the oral health of humans<sup>1)</sup>. This disease can result in missing teeth, including the destruction of the connective tissue and supporting bone<sup>2,3)</sup>. Dental biofilm is an important etiologic

factor of periodontitis<sup>4)</sup>. Dental biofilm contains various types of microorganisms, which Socransky et al. classified by potential pathogenicity<sup>5,6)</sup>. Generally, the proportion of anaerobic Gram-negative bacteria is important regarding the development of periodontitis<sup>7,8)</sup>.

Generally, missing teeth can be a useful marker of past and

current periodontal disease<sup>9</sup>. There is no prior study regarding the relationship between the number of missing teeth and periodontal pathogen as a covariate. Although some researchers used the number of missing teeth as a covariate or outcome variable, these studies did not consider periodontal pathogens<sup>10-12</sup>.

The convolutional neural network (CNN) is a form of deep learning, and can be used as an image classification tool through learning of characteristic features<sup>13</sup>. Recently, the CNN model was also used in a cross-sectional study, which should use logistic regression or multivariable linear regression (MLR)<sup>14,15</sup>.

A recent article in the 2017 world workshop on the classification of periodontal and peri-Implant diseases and conditions, which redefines the stages of periodontal disease, suggests that tooth loss is a major component of periodontal disease<sup>16</sup>. Using this information, the tooth loss was considered as a powerful risk factor, a research has been published to predict future tooth loss by stages of periodontal disease<sup>17</sup>. Long-term follow-up over 10 years of this study resulted in a significant increase in the risk of tooth loss at higher stages of periodontal disease. In addition, many studies have investigated tooth loss as a risk factor for cardiovascular disease, metabolic syndrome, and cognitive impairment<sup>18-20</sup>. In many respects, tooth loss affects not only oral health, but also systemic disease, and is a risk factor to watch out for.

The first aim of this study is to determine if the number of tooth loss can be predicted by periodontal disease pathogen. The second aim is to assess whether deep learning is a better way of predicting the number of teeth lost than MLR.

## Materials and Methods

### 1. Subjects

This study was approved by the institutional review board of the School of Dentistry, Seoul National University, Seoul, Korea (IRB number S-D20170023), and was carried out at the Hadan Goodwill Dental Hospital, Busan.

We collected the data by searching the patient's initial medical record. A total of 1,017 subjects were selected, who visited as new patients from August 1 to December 30, 2017. All patients were asked of demographic information and systemic diseases such as age, sex, smoking, number of cigarettes per a day, hypertension, diabetes mellitus, heart disease, and lung disease. A treatment plan was established for all patients when patients visited, and the number of planned tooth extractions was recorded without considering wisdom teeth. The number of miss-

ing teeth was counted using oral examination of medical chart without wisdom teeth. Microbiologic analyses were performed to measure the risk of periodontitis using real-time PCR. Finally, a total of 960 participants (456 men and 504 women) were included, excluding patients who refused microbiologic analyses.

### 2. Measurements

Participants' saliva specimens were collected after 30 seconds rinsing with mouthwash 10 ml (Garglin Dental Solution Regular, Dong-A Pharm Inc., Seoul, Korea) and processed for real-time PCR. Bacterial chromosomal DNA in saliva was extracted using a DNA extraction Kit (Exgene Clinic SV mini, GeneAll Inc., Seoul, Korea). The samples were analyzed using easyperio (BIOYD, Seongnam, Korea).

Socransky et al. classified the plaque bacteria into five groups according to the degree of influencing periodontal disease and the various periods of plaque formation<sup>21</sup>, among them, 11 oral disease pathogens were selected from red and orange complexes, which are highly pathogenic, causing periodontal disease. 11 oral disease pathogens are as follows: *Aggregatibacter actinomycetemcomitans* (Aa), *Porphyromonas gingivalis* (Pg), *Tannerella forsythus* (Tf), *Treponema denticola* (Td), *Fusobacterium nucleatum* (Fn), *Prevotella intermedia* (Pi), *Prevotella nigrescens* (Pn), *Streptococcus mitis* (Sm), *Streptococcus mutans* (Smu), *Streptococcus sobrinus* (Ss), *Lactobacillus casei* (Lc), and total bacterial load. Each bacteria DNA was amplified by a specific primer using functional genes (rgpB, waaA, gtf). Total bacteria 16s rDNA control was used to detect DNA from total bacteria load species. The samples in the DNA polymerase (AmpONE Hot-start Taq DNA polymerase 250u, GeneAll Inc., Seoul, Korea) assay were analyzed in a 20 µl reaction mixture containing genomic DNA 2 µl, specific primer, probe set, and PCR reaction buffer. The thermal program chosen was 45 cycles of 95°C for 15 seconds, 55°C for 15 seconds, and 72°C for 20 seconds, with an initial denaturation at 95°C for 15 minutes. All data were analyzed using sequenced-detection system software (ABI 7500 Fast Real-Time PCR System, Applied Biosystems, Life Technologies Inc., Carlsbad, CA, USA). Standard curves were used to convert cycle threshold (Ct) scores into the number of bacterial cells using samples with known amounts of bacterial-specific DNA. DNA was 10-fold serially diluted from 10<sup>0</sup> to 10<sup>5</sup> copies and subjected to real-time PCR to create a standard curve by plotting threshold cycles against the copy number of the plasmid DNA as previously described.

### 3. Statistical analysis

Data was divided into training set and test set categories.

The training set is the data for training a model and the test set is data to verify the performance of the model. A randomization sequence was generated using the RAND function in Excel (Microsoft Corporation, Redmond, WA, USA), and used to divide the full dataset (N=960) into a training set (N=658) and a test set (N=302). Data analysis was performed by SPSS software version 23.0 (IBM Co., Armonk, NY, USA). Significance was determined at  $\alpha=0.05$  for all tests. Each variable between the test set and training set was analyzed by a Chi-square test for non-continuous variables and an independent samples t-test. After adjusting for demographic factors, systemic diseases, and the number of bacteria by species, MLR was used for estimating the number of missing teeth in the training set. Four multi-variable regression

models were designed to analyze influences of each group of covariates as follows: model 1 has covariates that are essential for periodontitis prediction, such as age, sex, smoking, diabetes, hypertension, and red and orange complex; model 2 adds yellow complexes to caries-related bacteria and total bacterial load; model 3 adds smoking parameters that ignore multiple collinearity that should not be used in linear regression analysis, and adds less reliable heart and lung disease questionnaires; model 4 adds up to the tooth that was planned for extraction. Using constants and intercept calculated by MLR, the accuracy rate was calculated within a margin of error ( $-1 < x < 1$ ) in the training set.

A CNN was used in the deep learning model. This study

**Table 1.** Demographic information of the study population

	Total (N=960)	Training set (N=658)	Test set (N=302)
Age*	36.42±15.94	36.46±15.94	36.34±15.96
Sex, N, (%)			
Male	456 (47.5)	318 (48.3)	138 (45.7)
Female	504 (52.5)	340 (51.7)	164 (54.3)
Smoking, N, (%)			
None	773 (80.5)	527 (80.1)	246 (81.5)
Past	25 (2.6)	20 (3)	5 (1.7)
Current	162 (16.9)	111 (16.9)	51 (16.9)
Hypertension, N, (%)			
No	879 (91.6)	596 (90.6)	283 (93.7)
Yes	81 (8.4)	62 (9.4)	19 (6.3)
Diabetes mellitus, N, (%)			
No	935 (97.4)	642 (97.6)	293 (97.0)
Yes	25 (2.6)	16 (2.4)	9 (3.0)
Heart diseases, N, (%)			
No	959 (99.9)	657 (99.8)	302 (100.0)
Yes	1 (0.1)	1 (0.2)	0 (0.0)
Lung disease, N, (%)			
No	900 (93.8)	621 (94.4)	279 (92.4)
Yes	60 (6.3)	37 (5.6)	23 (7.6)
Number of cigarettes per day*	2.32±5.87	2.37±5.99	2.21±5.61
Number of missing teeth*	1.22±2.88	1.20±2.91	1.25±2.82
Number of planned tooth extractions*	0.23±1.25	0.18±1.00	0.34±1.68
Periodontal pathogens*			
<i>Aggregatibacter actinomycetemcomitans</i>	2.15*10 <sup>4</sup> ±1.74*10 <sup>5</sup>	1.98*10 <sup>4</sup> ±1.35*10 <sup>5</sup>	2.54*10 <sup>4</sup> ±2.38*10 <sup>5</sup>
<i>Porphyromonas gingivalis</i>	1.81*10 <sup>6</sup> ±4.14*10 <sup>6</sup>	1.78*10 <sup>6</sup> ±4.12*10 <sup>6</sup>	1.88*10 <sup>6</sup> ±4.19*10 <sup>6</sup>
<i>Tannerella forsythus</i>	2.09*10 <sup>6</sup> ±3.65*10 <sup>6</sup>	2.14*10 <sup>6</sup> ±3.88*10 <sup>6</sup>	1.98*10 <sup>6</sup> ±3.08*10 <sup>6</sup>
<i>Treponema denticola</i>	6.80*10 <sup>4</sup> ±3.10*10 <sup>5</sup>	5.90*10 <sup>4</sup> ±2.98*10 <sup>5</sup>	8.76*10 <sup>4</sup> ±3.36*10 <sup>5</sup>
<i>Fusobacterium nucleatum</i>	2.90*10 <sup>6</sup> ±6.76*10 <sup>6</sup>	2.91*10 <sup>6</sup> ±5.46*10 <sup>6</sup>	2.88*10 <sup>6</sup> ±8.97*10 <sup>6</sup>
<i>Prevotella intermedia</i>	8.26*10 <sup>5</sup> ±2.77*10 <sup>6</sup>	8.96*10 <sup>5</sup> ±2.96*10 <sup>6</sup>	6.74*10 <sup>5</sup> ±2.30*10 <sup>6</sup>
<i>Prevotella nigrescens</i>	8.64*10 <sup>4</sup> ±3.46*10 <sup>5</sup>	8.69*10 <sup>4</sup> ±3.74*10 <sup>5</sup>	8.55*10 <sup>4</sup> ±2.75*10 <sup>5</sup>
<i>Streptococcus mitis</i>	7.95*10 <sup>6</sup> ±9.16*10 <sup>6</sup>	7.79*10 <sup>6</sup> ±8.87*10 <sup>6</sup>	8.29*10 <sup>6</sup> ±9.76*10 <sup>6</sup>
<i>Streptococcus mutans</i>	2.80*10 <sup>6</sup> ±5.80*10 <sup>6</sup>	2.65*10 <sup>6</sup> ±5.48*10 <sup>6</sup>	3.13*10 <sup>6</sup> ±6.44*10 <sup>6</sup>
<i>Streptococcus sobrinus</i>	1.01*10 <sup>4</sup> ±1.59*10 <sup>5</sup>	9.86*10 <sup>4</sup> ±1.76*10 <sup>5</sup>	1.07*10 <sup>4</sup> ±1.14*10 <sup>5</sup>
<i>Lactobacillus casei</i>	1.21*10 <sup>5</sup> ±1.00*10 <sup>6</sup>	9.87*10 <sup>4</sup> ±8.61*10 <sup>5</sup>	1.70*10 <sup>5</sup> ±1.27*10 <sup>6</sup>
Total bacterial load	1.74*10 <sup>8</sup> ±1.52*10 <sup>8</sup>	1.74*10 <sup>8</sup> ±1.59*10 <sup>8</sup>	1.73*10 <sup>8</sup> ±1.38*10 <sup>8</sup>

\*SD, Standard deviation.

Table 2. Multiple linear regression results by model

	Model 1 ( $R^2=0.206$ , $P<0.001$ )			Model 2 ( $R^2=0.214$ , $P<0.001$ )			Model 3 ( $R^2=0.215$ , $P<0.001$ )			Model 4 ( $R^2=0.227$ , $P<0.001$ )		
	Log $\beta$	SE	P	Log $\beta$	SE	P	Log $\beta$	SE	P	Log $\beta$	SE	P
Constant	-1.433	0.454	0.002			0.004	-1.361	0.480	0.005	-1.327	0.477	0.006
Age	0.065	0.008	<0.001	0.065	0.008	<0.001	0.065	0.008	<0.001	0.063	0.008	<0.001
Sex	0.068	0.221	0.758	0.060	0.221	0.786	0.062	0.223	0.781	0.068	0.221	0.758
Number of planned tooth extraction										0.341	0.106	0.001
Smoking	0.144	0.147	0.328	0.157	0.148	0.290	0.110	0.270	0.683	0.105	0.268	0.695
Number of cigarettes per day							0.007	0.034	0.838	0.009	0.034	0.795
Hypertension	1.378	0.378	<0.001	1.367	0.382	<0.001	1.365	0.384	<0.001	1.287	0.382	0.001
Diabetes mellitus	-0.970	0.696	0.164	-0.978	0.698	0.162	-0.967	0.700	0.168	-0.976	0.695	0.161
Heart disease							2.082	2.628	0.428	2.180	2.609	0.404
Lung disease							-0.148	0.458	0.746	-0.132	0.454	0.771
<i>Aggregatibacter actinomycetemcomitans</i>	-1.1E-07	7.8E-07	0.886	-8.6E-08	7.8E-07	0.912	-9.3E-08	7.8E-07	0.905	-1.2E-07	7.8E-07	0.878
<i>Porphyromonas gingivalis</i>	8.1E-08	3.1E-08	0.010	7.9E-08	3.2E-08	0.013	7.9E-08	3.2E-08	0.013	5.6E-08	3.2E-08	0.080
<i>Tannerella forsythus</i>	-3.2E-08	3.5E-08	0.360	-2.6E-08	3.5E-08	0.465	-2.6E-08	3.5E-08	0.468	-1.4E-08	3.5E-08	0.682
<i>Treponema denticola</i>	-1.6E-08	3.7E-07	0.966	2.0E-07	3.9E-07	0.604	2.1E-07	4.0E-07	0.592	2.2E-07	3.9E-07	0.585
<i>Fusobacterium nucleatum</i>	-1.6E-08	2.0E-08	0.421	-9.4E-09	2.0E-08	0.645	-1.1E-08	2.0E-08	0.606	-9.9E-09	2.0E-08	0.625
<i>Prevotella intermedia</i>	-4.3E-08	3.7E-08	0.246	-4.4E-08	3.7E-08	0.232	-4.3E-08	3.7E-08	0.251	-4.6E-08	3.7E-08	0.215
<i>Prevotella nigrescens</i>	1.3E-07	2.8E-07	0.641	9.0E-08	2.9E-07	0.755	9.1E-08	2.9E-07	0.755	2.7E-08	2.9E-07	0.926
<i>Streptococcus mitis</i>				1.9E-08	1.5E-08	0.204	1.9E-08	1.5E-08	0.207	2.0E-08	1.5E-08	0.171
<i>Streptococcus mutans</i>				-5.3E-09	2.3E-08	0.821	-6.4E-09	2.4E-08	0.786	-6.8E-09	2.3E-08	0.770
<i>Streptococcus sobrinus</i>				9.4E-07	5.9E-07	0.110	9.4E-07	5.9E-07	0.111	1.0E-06	5.9E-07	0.086
<i>Lactobacillus casei</i>				-1.4E-07	1.2E-07	0.240	-1.4E-07	1.2E-07	0.238	-1.3E-07	1.2E-07	0.268
Total bacterial load				-1.2E-09	8.7E-10	0.165	-1.2E-09	8.7E-10	0.176	-1.3E-09	8.7E-10	0.148

Model 1 was adjusted for age, sex, smoking, hypertension, diabetes mellitus, *A. actinomycetemcomitans*, *P. gingivalis*, *T. forsythus*, *F. nucleatum*, *P. intermedia*, and *P. nigrescens*. Model 2 was adjusted for age, sex, smoking, hypertension, diabetes mellitus, *A. actinomycetemcomitans*, *P. gingivalis*, *T. forsythus*, *F. nucleatum*, *P. intermedia*, and *P. nigrescens*. *S. mitis*, *S. mutans*, *S. sobrinus*, *L. casei*, and total bacterial load. Model 3 was adjusted for age, sex, smoking, number of cigarettes per a day, hypertension, diabetes mellitus, heart diseases, lung diseases, *A. actinomycetemcomitans*, *P. gingivalis*, *T. forsythus*, *F. nucleatum*, *P. intermedia*, and *P. nigrescens*. *S. mitis*, *S. mutans*, *S. sobrinus*, *L. casei*, and total bacterial load. Model 4 was adjusted for age, sex, number of planned tooth extractions, smoking, number of cigarettes per a day, hypertension, diabetes mellitus, heart diseases, lung diseases, *A. actinomycetemcomitans*, *P. gingivalis*, *T. forsythus*, *F. nucleatum*, *P. intermedia*, and *P. nigrescens*. *S. mitis*, *S. mutans*, *S. sobrinus*, *L. casei*, and total bacterial load.

SE, Standard error.

**Table 3.** Comparison of accuracy between convolutional neural network and multiple linear regression for each model

	Model 1		Model 2		Model 3		Model 4	
	Accuracy rate	MSE	Accuracy rate	MSE	Accuracy rate	MSE	Accuracy rate	MSE
CNN	65.0	4.90	69.5	3.07	68.3	3.33	70.2	3.54
MLR	50.9	95.21	50.0	69.90	50.0	59.68	51.2	60.15

Model 1 was adjusted for age, sex, smoking, hypertension, diabetes mellitus, *A. actinomycetemcomitans*, *P. gingivalis*, *T. forsythus*, *F. nucleatum*, *P. intermedia*, and *P. nigrescens*. Model 2 was adjusted for age, sex, smoking, hypertension, diabetes mellitus, *A. actinomycetemcomitans*, *P. gingivalis*, *T. forsythus*, *F. nucleatum*, *P. intermedia*, and *P. nigrescens*, *S. mitis*, *S. mutans*, *S. sobrinus*, *L. casei*, and total bacterial load. Model 3 was adjusted for age, sex, smoking, number of cigarettes per a day, hypertension, diabetes mellitus, heart diseases, lung diseases, *A. actinomycetemcomitans*, *P. gingivalis*, *T. forsythus*, *F. nucleatum*, *P. intermedia*, and *P. nigrescens*, *S. mitis*, *S. mutans*, *S. sobrinus*, *L. casei*, and total bacterial load. Model 4 was adjusted for age, sex, number of planned tooth extractions, smoking, number of cigarettes per a day, hypertension, diabetes mellitus, heart diseases, lung diseases, *A. actinomycetemcomitans*, *P. gingivalis*, *T. forsythus*, *P. intermedia*, *P. intermedia*, and *P. nigrescens*, *S. mitis*, *S. mutans*, *S. sobrinus*, *L. casei*, and total bacterial load.

CNN, Convolutional neural network; MLR, Multi-variable linear regression; MSE, Mean squared error.

used three convolutional layers and ten hidden nodes. Each convolutional layer has 25, 50, 25 filters of kernel size 1×3. A total of 200 epochs was used to reduce overfitting. Training network weights were learned using the Adam algorithm (learning rate=0.0001), a stochastic gradient descent method. Four CNN models, which had different input nodes, were designed to compare with MLR models. Each model was performed five times to generate average accuracy rate and mean square error (MSE). All CNN analysis was performed by Python 3.6.1 (Python Software Foundation, Wilmington, DE, USA) with a TensorFlow (Google, Mountain View, CA, USA) framework.

## Results

The demographic information of the study population is shown in Table 1. A total of 960 participants, 456 males (47.5%) and 504 females (52.5%), were included in this study. The mean number of missing teeth was  $1.22 \pm 2.88$  and mean number of planned tooth extraction was  $0.23 \pm 1.25$ . The number of non-smoker was 773 (80.5%), and those who had diabetes mellitus numbered 25 (2.6%). There was no significant difference between the training set (N=658) and test set (N=302) for any variable.

In the multi-variable linear regression, the number of missing teeth was significantly associated with age and hypertension. *Porphyromonas gingivalis* also showed significant association with the number of missing teeth, except for Model 4 (Table 2).

The accuracy of predicting the number of missing teeth was evaluated and compared between the CNN and MLR approaches. The accuracy of the CNN method was 65.0% in Model 1, which increased to 70.2% in Model 4 (Table 3). The accuracy of the MLR was about 50.0% in all models. The MSE of the CNN method was significantly smaller than the MLR.

## Discussion

This cross-sectional study assessed the relationship between the number of missing teeth and periodontal pathogens in saliva using deep learning compared to MLR. The study was unique in that CNN was used in the analysis of a cross-sectional study, not in the analysis of the image. Generally, CNN was investigated as a diagnostic accuracy tool using a medical image such as a radiograph<sup>15,22,23</sup>. The results showed that the CNN model had higher accuracy for the prediction of missing teeth than the MLR model.

Model 1 variables were clinically reliable and acceptable in the traditional statistical model. Several studies about periodontal pathogens have dealt with red complex and orange complex similar to the periodontal pathogen of Model 1<sup>6,9,24-26</sup>. Comparing Model 1, Model 2 used additionally *S. mitis*, *S. mutans*, *S. sobrinus*, *L. casei*, and total bacterial load, which were seldom used in the periodontal pathogen. In the CNN model, the accuracy rate was increased by 4.5%. This may be due to adjusting the number of bacteria related to dental caries in that the cause of tooth extraction is not only periodontal disease but also dental caries.

Model 3 used heart disease, lung disease, and the number of cigarettes per a day additionally. Heart disease and lung disease were measured using a simple questionnaire about history, not accurate medical examination. Moreover, adding the number of cigarettes per day could adjust collinearity, because the smoking variable already was adjusted. Model 4 used the number of planned tooth extraction which was calculated from the treatment plan of dental chart. However, this variable could be unreliable, because there was no consideration about prosthetic extraction or residual root. Therefore, Models 2, 3, and 4 could be considered as poor statistical models. However, all



CNN models showed higher accuracy and lower MSE than MLR. Interestingly, an increasing trend of accuracy was shown adding covariates. While the MLR model showed similar accuracy and MSE adding covariates.

Generally, the deep CNN model used several convolutional layers and fully connected layers, because object detection and image classification are complex problems. However, we used only a convolutional layer and hidden layer, because our data was not image, but cross-sectional data, which had only 22 variables. The data was not as large as other image studies. Hence, the only three convolutional layers used to facilitate deep learning performance.

This study had several limitations. First, this study was a retrospective cross-sectional design. Hence, the available variables were limited. Probing depth in dental chart information could be a clinically important variable. However, probing depth without calibration training could not be reliable. Moreover, there were many individuals without a probing depth record. Participants were all new patients, and the number of planned tooth extractions was skewed to zero or one. If the number of planned tooth extractions was distributed normally, then the number of planned tooth extractions could be an outcome variable.

The second limitation of this study was the number of participants. The number of participants in this study was 960. For the deep learning analysis, all data require labelling. Especially in this research area, dental professionals should label data by themselves. Therefore, it was difficult to increase the number of participants in this study. Considering these limitations, further prospective well-designed studies, which have more than 5,000 participants, are required.

Recently, the human microbiome has attracted attention as an emerging theme, because using only human genomic information can not solve the mechanism of disease progression<sup>5,27</sup>. Knowledge of the oral microbiome with deep learning will enable more accurate prediction of oral diseases.

## Conclusions

This study shows that oral disease-related bacteria can influence the prediction of tooth loss, and the deep learning analysis method can be used more effectively than the conventional MLR method. Although the association between oral disease bacteria and lost teeth has been well documented, few studies have been conducted on predictive models using deep learning. In future studies, prospective study design and the collection of clinical and epidemiological evidence will be needed, using a

greater number of samples. This study can demonstrate the use of deep learning in research to develop predictive models of oral health.

## ORCID

Seon-Jip Kim, <https://orcid.org/0000-0001-5909-5743>

Dohyoung Rim, <https://orcid.org/0000-0003-2022-6333>

Jeong Uk Heo, <https://orcid.org/0000-0001-7112-3196>

## References

1. Kassebaum NJ, Bernabé E, Dahiya M, Bhandari B, Murray CJ, Marcenes W. Global burden of severe periodontitis in 1990-2010: a systematic review and meta-regression. *J Dent Res* 2014;93:1045-1053.
2. Haffajee AD, Socransky SS. Microbial etiological agents of destructive periodontal diseases. *Periodontol* 2000 1994;5:78-111.
3. Schätzle M, Löe H, Lang NP, Bürgin W, Anerud A, Boysen H. The clinical course of chronic periodontitis. *J Clin Periodontol* 2004;31:1122-1127.
4. Theilade E, Theilade J. Role of plaque in the etiology of periodontal disease and caries. *Oral Sci Rev* 1976;9:23-63.
5. Pflughoeft KJ, Versalovic J. Human microbiome in health and disease. *Annu Rev Pathol* 2012;7:99-122.
6. Socransky SS, Haffajee AD. Periodontal microbial ecology. *Periodontol* 2000 2005;38:135-187.
7. Slots J. The predominant cultivable microflora of advanced periodontitis. *Scand J Dent Res* 1977;85:114-121.
8. Haffajee AD, Cugini MA, Tanner A, Pollack RP, Smith C, Kent RL Jr, et al. Subgingival microbiota in healthy, well-maintained elder and periodontitis subjects. *J Clin Periodontol* 1998;25:346-353.
9. Hyvärinen K, Salminen A, Salomaa V, Pussinen PJ. Systemic exposure to a common periodontal pathogen and missing teeth are associated with metabolic syndrome. *Acta Diabetol* 2015;52:179-182.
10. Piuvezam G, de Lima KC. Factors associated with missing teeth in the Brazilian elderly institutionalised population. *Gerodontology* 2013;30:141-149.
11. Choi HM, Han K, Park YG, Park JB. Associations between the number of natural teeth and renal dysfunction. *Medicine (Baltimore)* 2016;95:e4681.
12. A-Dan W, Jun-Qi L. Factors associated with the oral health-related quality of life in elderly persons in dental clinic: validation of a Mandarin Chinese version of GOHAI. *Gerodontology* 2011;28:184-191.
13. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems* 28. Neural Information Processing Systems Foundation Inc.;2012:1097-1105.
14. Trtica-Majnaric L, Zekic-Susac M, Sarlija N, Vitale B. Prediction of influenza vaccination outcome by neural networks and logistic regression. *J Biomed Inform* 2010;43:774-781.
15. Wang Z, Li L, Glicksberg BS, Israel A, Dudley JT, Ma'ayan A. Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age. *J Biomed Inform* 2017;76:59-68.
16. Tonetti MS, Greenwell H, Kornman KS. Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *J Periodontol* 2018;89 Suppl 1:S159-S172.
17. Ravidà A, Qazi M, Troiano G, Saleh MHA, Greenwell H, Kornman K,

- et al. Using periodontal staging and grading system as a prognostic factor for future tooth loss: A long-term retrospective study. *J Periodontol* 2019 Sep 9 [Epub]. DOI: 10.1002/JPER.19-0390.
18. Cheng F, Zhang M, Wang Q, Xu H, Dong X, Gao Z, et al. Tooth loss and risk of cardiovascular disease and stroke: A dose-response meta analysis of prospective cohort studies. *PLoS One* 2018;13:e0194563.
  19. Souza ML, Massignan C, Glazer Peres K, Aurelio Peres M. Association between metabolic syndrome and tooth loss: A systematic review and meta-analysis. *J Am Dent Assoc* 2019;150:1027-1039.e7.
  20. Saito S, Ohi T, Murakami T, Komiyama T, Miyoshi Y, Endo K, et al. Association between tooth loss and cognitive impairment in community-dwelling older Japanese adults: a 4-year prospective cohort study from the Ohasama study. *BMC Oral Health* 2018;18:142.
  21. Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL Jr. Microbial complexes in subgingival plaque. *J Clin Periodontol* 1998;25:134-144.
  22. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175:1828-1837.
  23. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402-2410.
  24. Hyvärinen K, Laitinen S, Paju S, Hakala A, Suominen-Taipale L, Skurnik M, et al. Detection and quantification of five major periodontal pathogens by single copy gene-based real-time PCR. *Innate Immun* 2009;15:195-204.
  25. Gomes SC, Nonnenmacher C, Susin C, Oppermann RV, Mutters R, Marcantonio RA. The effect of a supragingival plaque-control regimen on the subgingival microbiota in smokers and never-smokers: evaluation by real-time polymerase chain reaction. *J Periodontol* 2008;79:2297-2304.
  26. Gomes SC, Piccinin FB, Oppermann RV, Susin C, Nonnenmacher CI, Mutters R, et al. Periodontal status in smokers and never-smokers: clinical findings and real-time polymerase chain reaction quantification of putative periodontal pathogens. *J Periodontol* 2006;77:1483-1490.
  27. Wade WG. The oral microbiome in health and disease. *Pharmacol Res* 2013;69:137-143.