



OPEN

## Deep learning and clustering approaches for dental implant size classification based on periapical radiographs

Ji-Hyun Park<sup>1</sup>, Hong Seok Moon<sup>1</sup>, Hoi-In Jung<sup>2</sup>, JaeJoon Hwang<sup>3</sup>, Yoon-Ho Choi<sup>4</sup> & Jong-Eun Kim<sup>1</sup>✉

This study investigated two artificial intelligence (AI) methods for automatically classifying dental implant diameter and length based on periapical radiographs. The first method, deep learning (DL), involved utilizing the pre-trained VGG16 model and adjusting the fine-tuning degree to analyze image data obtained from periapical radiographs. The second method, clustering analysis, was accomplished by analyzing the implant-specific feature vector derived from three key points coordinates of the dental implant using the k-means++ algorithm and adjusting the weight of the feature vector. DL and clustering model classified dental implant size into nine groups. The performance metrics of AI models were accuracy, sensitivity, specificity, F1-score, positive predictive value, negative predictive value, and area under the receiver operating characteristic curve (AUC-ROC). The final DL model yielded performances above 0.994, 0.950, 0.994, 0.974, 0.952, 0.994, and 0.975, respectively, and the final clustering model yielded performances above 0.983, 0.900, 0.988, 0.923, 0.909, 0.988, and 0.947, respectively. When comparing the AI model before tuning and the final AI model, statistically significant performance improvements were observed in six out of nine groups for DL models and four out of nine groups for clustering models based on AUC-ROC. Two AI models showed reliable classification performances. For clinical applications, AI models require validation on various multicenter data.

The dental implant is a valuable treatment option for edentulous patients<sup>1</sup>. The long-term success rate and survival rate of dental implants are guaranteed, but mechanical and biological complications occur in patients with dental implants as time passes<sup>2-4</sup>. To manage various complications, detailed information on dental implants is essential.

In the case of mechanical complications, such as a fracture of dental prostheses, it is necessary to identify the dental implant systems. Afterward, the diameter of the implant should also be identified because the diameter of the dental implant determines the dimension of the connection between a superstructure and the dental implant. In the case of biological complications, especially peri-implantitis, radiographic measurement of bone level is a crucial factor<sup>5-7</sup>. In cases where previous radiographic examinations are unavailable, the diagnosis of peri-implantitis can be established when bone levels  $\geq 3$  mm apical to the most coronal aspect of the intra-osseous part of the implant are observed, accompanied by bleeding on probing<sup>5</sup>. Clinicians often measure bone loss using a relative ratio to the total implant length from the periapical radiograph. However, measuring exact bone loss or objectively comparing the rate of bone loss among patients with different implants is challenging using a relative bone loss ratio. A dental implant's length can serve as a reference metric in dental radiograph interpretation for radiographic measurement of bone level.

Numerous types of implants have emerged, so without medical records, getting specific and detailed information on dental implants is difficult<sup>8</sup>. In addition, after dental implant surgery, it is hard to observe or measure it directly, so periapical or panoramic radiographs are used to examine dental implants. However, identifying

<sup>1</sup>Department of Prosthodontics, Yonsei University College of Dentistry, Yonsei-ro 50-1, Seodaemun-gu, Seoul 03722, Korea. <sup>2</sup>Department of Preventive Dentistry and Public Oral Health, Yonsei University College of Dentistry, Seoul 03722, Korea. <sup>3</sup>Department of Oral and Maxillofacial Radiology, School of Dentistry, Dental Research Institute, Pusan National University, Busan 50612, Korea. <sup>4</sup>School of Computer Science and Engineering, Pusan National University, Busan 46241, Korea. ✉email: gomyou@yuhs.ac

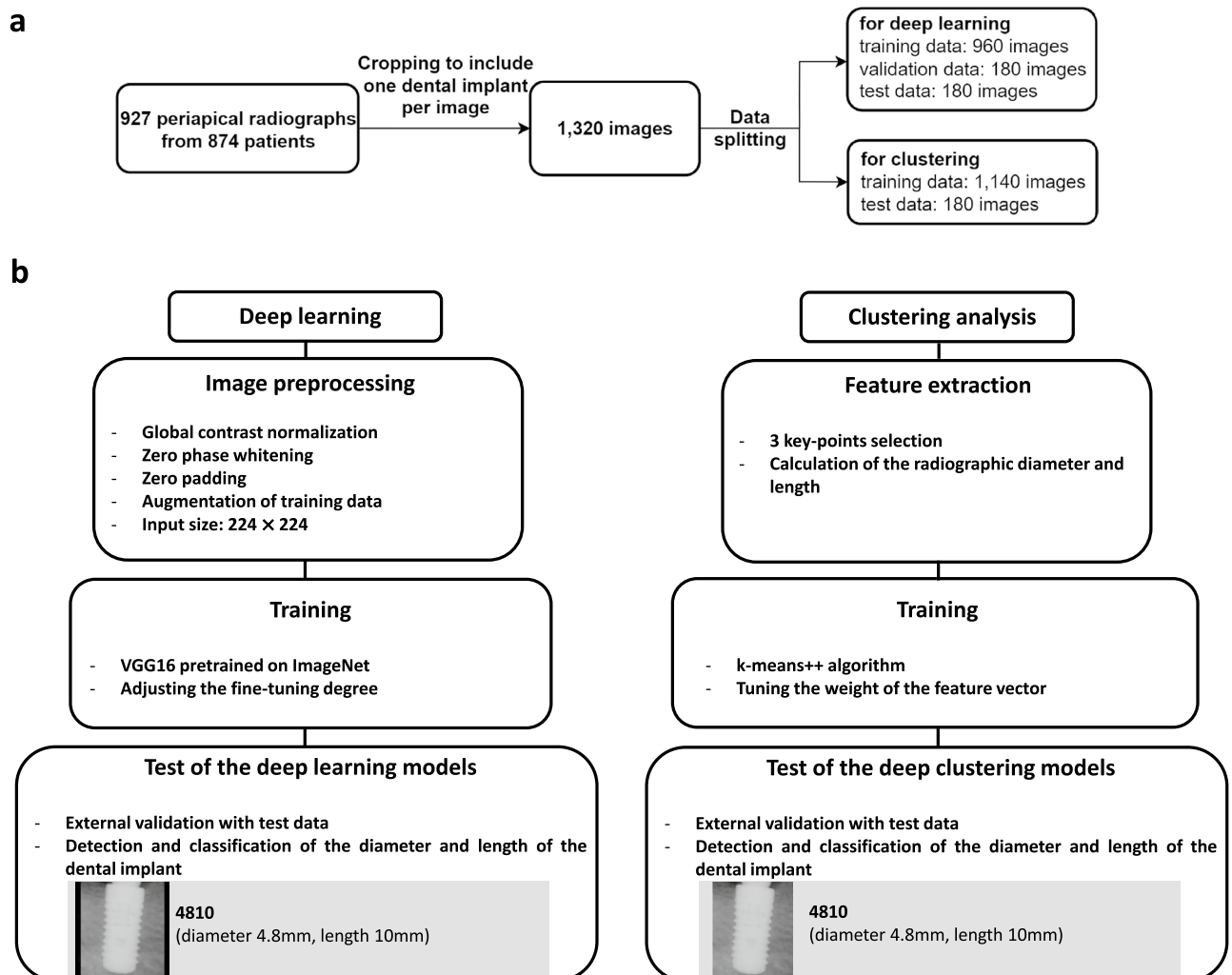
dental implants solely based on clinicians' experience can be time-consuming and costly without sufficient medical records.

The application of artificial intelligence (AI) in medical imaging analysis has expanded rapidly<sup>9–14</sup>. Also, in dentistry, AI models were used for the diagnosis of dental diseases such as dental caries or periodontal disease<sup>15–17</sup>. As a result, they improved the accuracy and reliability of diagnoses and aided dental professionals' performances<sup>17</sup>. Active AI research was conducted to identify dental implant systems from dental radiography<sup>18–24</sup>. There were studies to obtain high performance in identifying dental implant types using small datasets, various deep convolutional neural networks (CNNs), and transfer learning. Recently a study showed that the deep learning (DL) model yielded accurate and valid results in identifying and classifying different dental implant systems with large-scale multicenter data<sup>24</sup>.

However, few studies have investigated implant size, which plays an essential role in the production of implant prostheses and serves as a reference metric when interpreting dental radiographs. While recent research has focused on identifying dental implant systems, more investigations about automatically classifying implant size are needed. To address this research gap, this study aimed to develop and evaluate two automatic classification systems classifying the size of dental implants from periapical radiographs with DL and clustering (Fig. 1). It was hypothesized that the performance of the two AI models could be improved through the tuning process.

## Results

The label with a four-digit number represents the dental implant size, with the first two digits corresponding to the diameter and the last two digits corresponding to the length (3308, diameter 3.3 mm and length 8 mm; 3310, diameter 3.3 mm and length 10 mm; 3312, diameter 3.3 mm and length 12 mm; 4108, diameter 4.1 mm and length 8 mm; 4110, diameter 4.1 mm and length 10 mm; 4112, diameter 4.1 mm and length 12 mm; 4808, diameter 4.8 mm and length 8 mm; 4810, diameter 4.8 mm and length 10 mm; 4812, diameter 4.8 mm and length 12 mm).



**Figure 1.** A schematic description of deep learning and clustering approaches: (a) data acquisition and data splitting for deep learning and clustering; (b) comparing of deep learning and clustering process.

### Classification of the implant size with DL

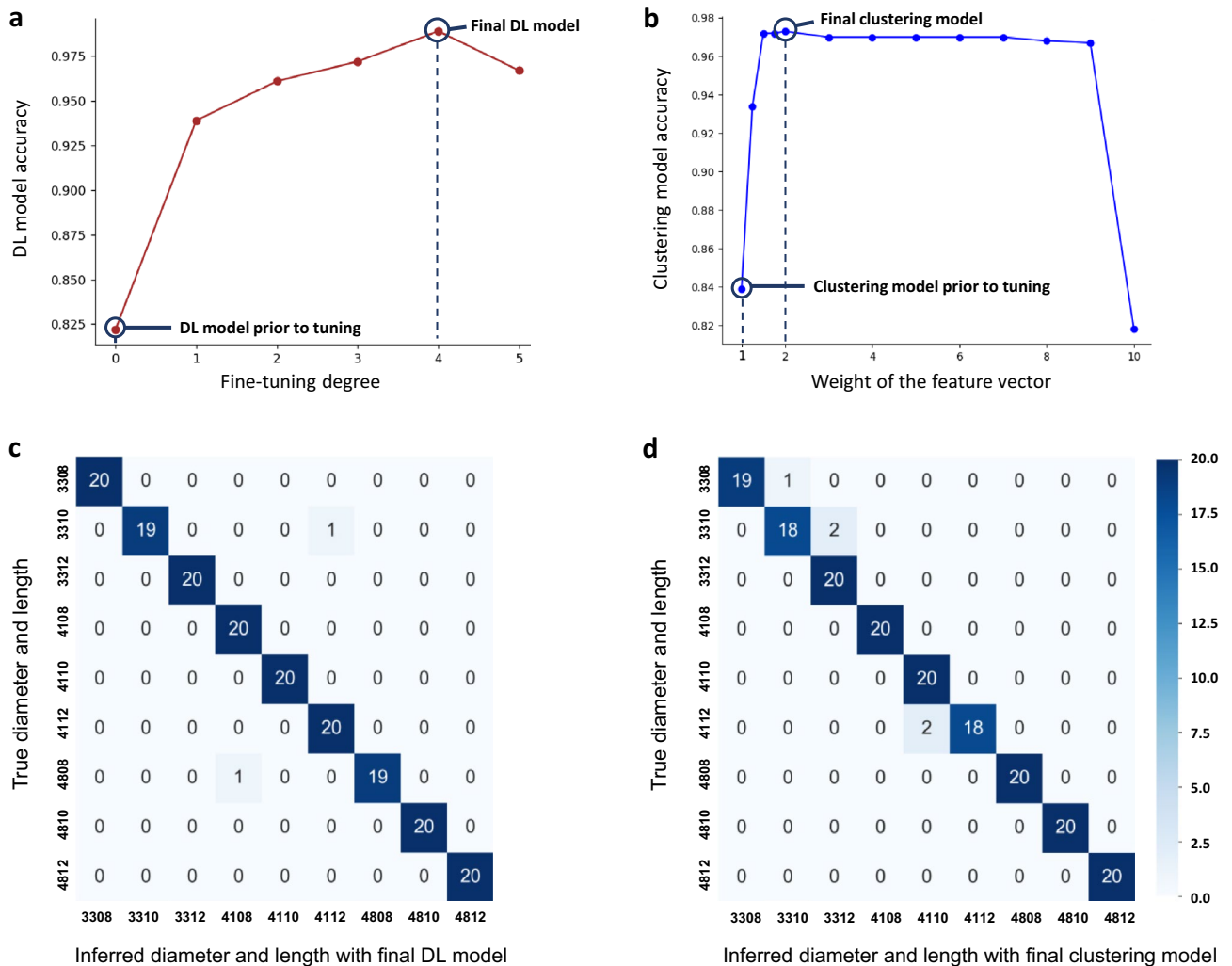
The final DL model with the best accuracy was chosen when the fine-tuning degree was four (Fig. 2a). Through adjusting the fine-tuning degree, the final DL model exhibited, in six out of nine groups, higher AUC-ROC values and statistically significant differences ( $p < 0.05$ ) in terms of AUC-ROC compared to the model with the fine-tuning degree of zero.

Figure 2c shows a confusion matrix constructed by the final DL model with test data. Across the result of nine groups, the accuracy, sensitivity, specificity, F1-score, positive predictive value, negative predictive value, and AUC-ROC are above 0.994, 0.950, 0.994, 0.974, 0.952, 0.994, and 0.975 (95% CI, 0.926–1.00) respectively (Table 1). In the confusion matrix of the final DL model, there are two false predictions. One represents an inaccurate inference of the length and diameter, and the other shows an accurate inference of the length and an inaccurate inference of diameter. A confusion matrix and the performance evaluation results of the DL model with the fine-tuning degree of zero are shown (Supplementary Fig. S3 and Supplementary Table S4). Receiver operating characteristic (ROC) curve and area under the ROC curve (AUC-ROC) of the DL models with the best accuracy and the fine-tuning degree of zero in nine groups, and the results of the chi-square test are shown (Supplementary Fig. S5).

The input images of bone level implants and heat maps generated by gradient-weighted class activation mapping (Grad-CAM) were superimposed on the corresponding images, and they are presented in Fig. 3. In each heat map, red regions indicate higher activation values or importance, while blue regions represent lower activation values or importance in the prediction process<sup>25</sup>.

### Classification of the implant size with clustering

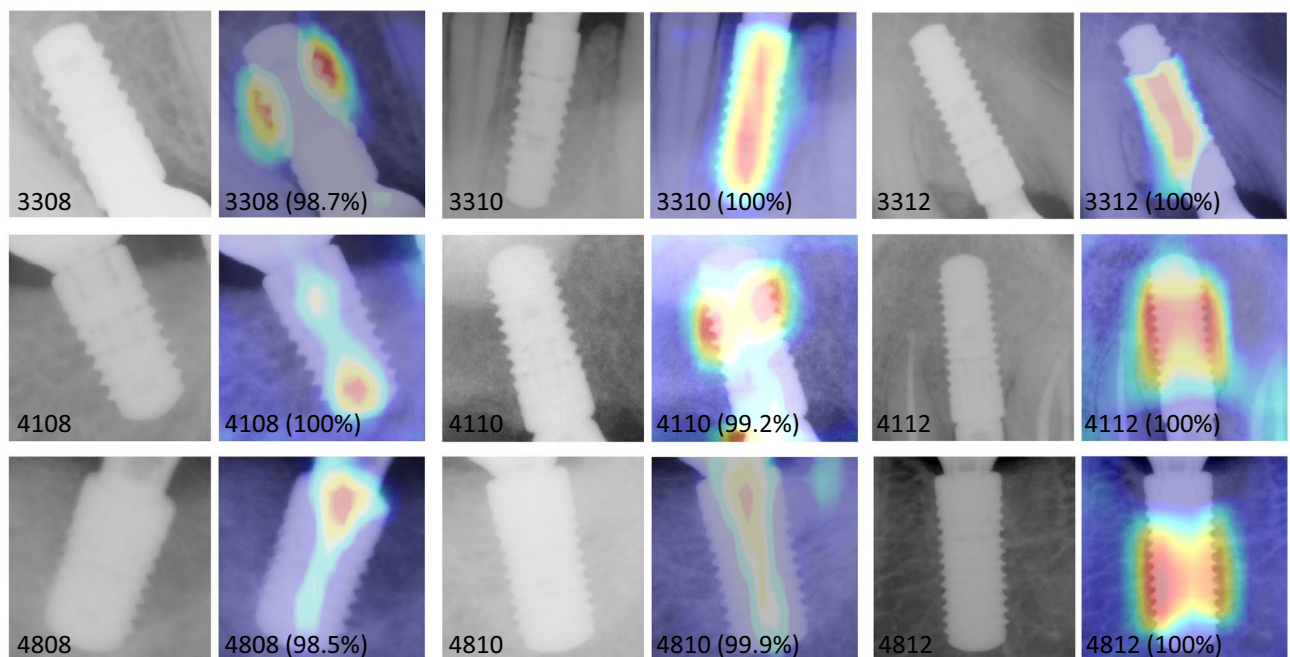
Using a test data set, external validation was conducted on the final clustering model with the best accuracy when the weight of the feature vector was two (Fig. 2b). Through adjusting the weight of the feature vector, the



**Figure 2.** Results for implant size classification using deep learning and clustering approaches: (a) relationship between the fine-tuning degree and deep learning model accuracy; (b) relationship between the weight of the feature vector and clustering model accuracy; (c) confusion matrix of the final DL model result; (d) confusion matrix of the final clustering model result.

Model	Label	TP	TN	FP	FN	ACC	SE	SP	F1-score	PPV	NPV	AUC-ROC (95% CI)
DL	3308	20	160	0	0	1.000	1.000	1.000	1.000	1.000	1.000	*1.000 (1.000–1.000)
	3310	19	160	0	1	0.994	0.950	1.000	0.974	1.000	0.994	*0.975 (0.926–1.000)
	3312	20	160	0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000 (1.000–1.000)
	4108	20	159	1	0	0.994	1.000	0.994	0.976	0.952	1.000	*0.997 (0.991–1.000)
	4110	20	160	0	0	1.000	1.000	1.000	1.000	1.000	1.000	*1.000 (1.000–1.000)
	4112	20	159	1	0	0.994	1.000	0.994	0.976	0.952	1.000	*0.997 (0.991–1.000)
	4808	19	160	0	1	0.994	0.950	1.000	0.974	1.000	0.994	0.975 (0.926–1.000)
	4810	20	160	0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.00 (1.000–1.000)
	4812	20	160	0	0	1.000	1.000	1.000	1.000	1.000	1.000	*1.00 (1.000–1.000)
CL	3308	19	160	0	1	0.994	0.950	1.000	0.974	1.000	0.994	0.975 (0.926–1.00)
	3310	18	159	1	2	0.983	0.900	0.994	0.923	0.947	0.988	*0.947 (0.879–1.00)
	3312	20	158	2	0	0.989	1.000	0.988	0.952	0.909	1.000	0.994 (0.985–1.00)
	4108	20	160	0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000 (1.000–1.000)
	4110	20	158	2	0	0.989	1.000	0.988	0.952	0.909	1.000	*0.994 (0.985–1.000)
	4112	18	160	0	2	0.989	0.900	1.000	0.947	1.000	0.988	*0.950 (0.883–1.000)
	4808	20	160	0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000 (1.000–1.000)
	4810	20	160	0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000(1.000–1.000)
	4812	20	160	0	0	1.000	1.000	1.000	1.000	1.000	1.000	*1.000 (1.000–1.000)

**Table 1.** Performance evaluation results of the final DL and clustering models. *DL* deep learning, *CL* clustering, *TP* true positive, *TN* true negative, *FP* false positive, *FN* false negative, *ACC* accuracy, *SE* sensitivity, *SP* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *AUC-ROC* area under the receiver operating characteristic curve, *CI* confidence interval. The asterisk (\*) indicates a statistically significant difference between the model performance before and after tuning, with a significance level of  $p < 0.05$ .



**Figure 3.** Bone level implant images and their Grad-CAM of the final deep learning model, described with true label, predicted label, and softmax value.

final clustering model exhibited, in four out of nine groups, higher AUC-ROC values and statistically significant differences ( $p < 0.05$ ) in terms of AUC-ROC compared to the model with a weight of the feature vector set to one.

Figure 2d shows a confusion matrix constructed by the final clustering model with test data. Across the result of nine groups, the accuracy, sensitivity, specificity, F1-score, positive predictive value, negative predictive value, and AUC-ROC were above 0.983, 0.900, 0.988, 0.923, 0.909, 0.988, and 0.947 (95% CI, 0.879–1.00) respectively (Table 1). In the clustering model, each data point is assigned in each cluster based on Euclidean similarity, so the false predictions always appear on the cluster near the true predictions on the scatter plot of

the feature vectors. In the confusion matrix of the final clustering model, the false prediction results represent accurate inferences on diameter and inaccurate inferences on length. The radiographic diameter and length are indicated on the *x*- and *y*-axes, respectively, on the scatter plot, and the cluster of each data is represented with color coding (Fig. 4). A confusion matrix and the performance evaluation results of the clustering model with the weight of the feature vector set to one are shown (Supplementary Fig. S6 and Supplementary Table S7). ROC curves and AUC-ROC values of the clustering models with the best accuracy and the weight of the feature vector set to one in nine groups and the results of the chi-square test are shown (Supplementary Fig. S8).

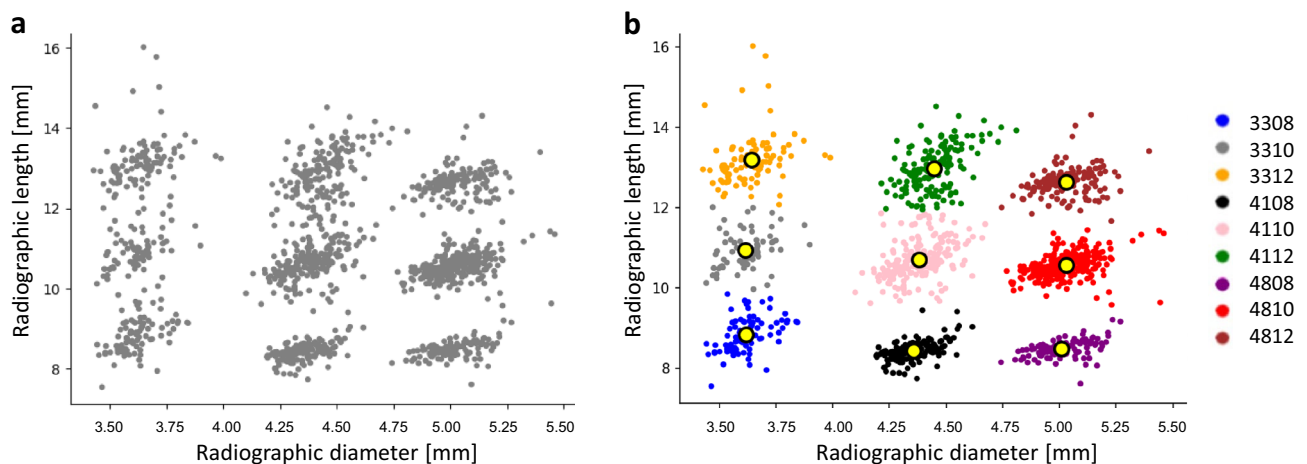
## Discussion

AI models in implant dentistry have the potential to recognize implant types, predict implant success, and optimize implant designs<sup>26</sup>. Many studies have attempted to develop methods for identifying the manufacturers and models of implants based on AI models<sup>20–24</sup>. Nevertheless, few studies have focused on the detailed diameter and length of dental implants in radiographic images. To the best of our knowledge, using deep learning or clustering to classify the size of dental implants has not yet been studied. The diameter of a dental implant plays an important role when selecting components for prosthetic restorations, while its length can serve as a reference for radiographic interpretations. Therefore, accurate information on the exact dimensions of the implant is essential for its long-term maintenance, including the management of mechanical and biological complications.

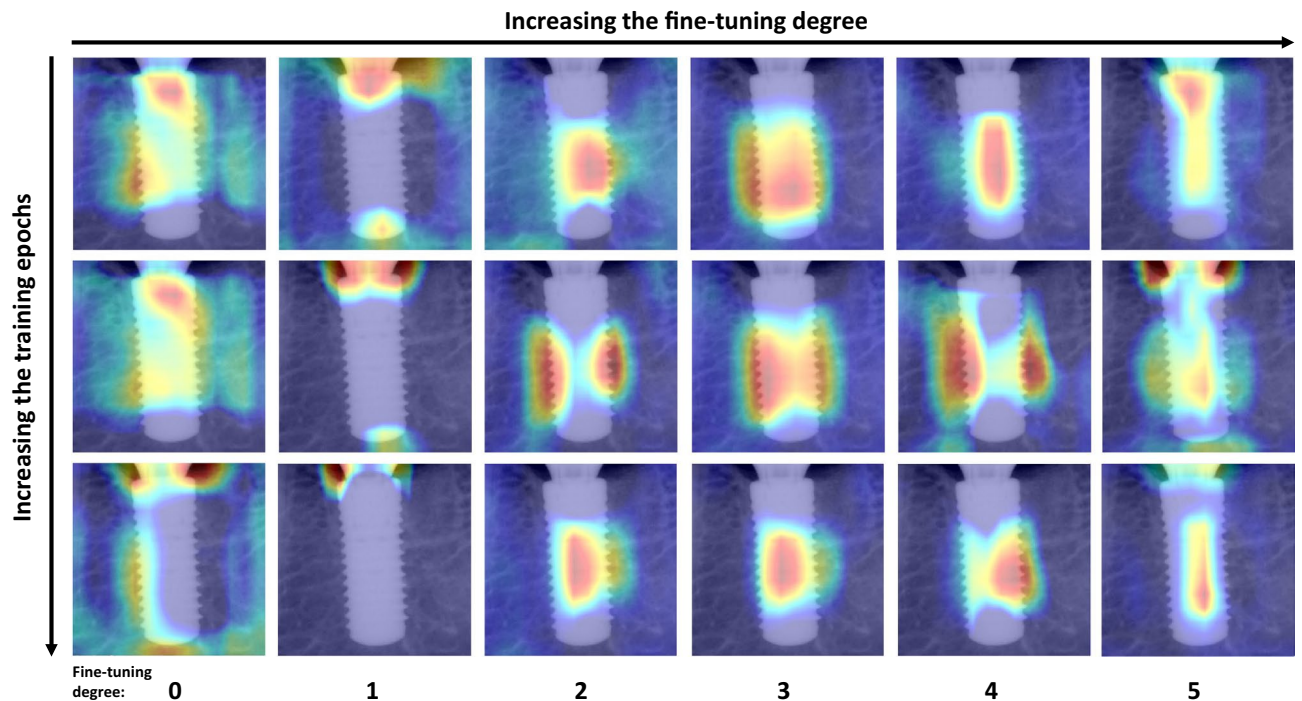
A study used a DL model to find six key points on periapical radiographs of dental implants and calculated the ratio of bone loss for the entire implant<sup>27</sup>. An automated key points detection system was proposed for calculating the percentage of bone loss to assess the severity of peri-implantitis. However, for the same ratio of the radiographic bone loss over the total implant length, the actual amount of bone loss varies proportionally with the implant length. Measuring bone loss amount rather than bone loss ratio is necessary for accurately interpreting radiographs for peri-implantitis diagnosis. Therefore, the present study aimed to develop automatic systems to identify the diameter and length of dental implants and to make it possible to measure the objective amount of bone loss.

In the clustering methods of this study, we employed not only three key points but also clustering analysis to infer the diameter and length of dental implants. In various prior studies of automatic lateral cephalogram analysis, deep learning was used for size measurements of various structures through landmark detection. In contrast, in this study, after extracting three key points, we utilized clustering analysis to infer the diameter and length of dental implants. This approach stems from differences in imaging conditions between lateral cephalograms and intraoral radiographs. In the case of lateral cephalograms, where the patient's sagittal plane is positioned parallel to the radiographic plate during imaging, the derived values can contribute to cephalogram analysis. However, from periapical radiographs of implants captured within the oral cavity, direct calculation of the implant's diameter and length is complicated by factors such as magnification resulting from radiographic imaging and reduction caused by the angle between the implant and the digital sensor. Particularly, the angular issue between the object and the digital sensor introduces limitations based on the anatomical structures within the oral cavity, varying according to the inclination of implants not visible within the bone and different dental regions such as incisors, premolars, and molars. This study explored the feasibility of transforming radiographic diameter and length, obtained through landmark usage, into actual implant diameter and length through classification, employing machine learning techniques, specifically clustering.

In medical image analysis, AI models should be explainable to ensure the clinical relevance and reliability of the model's results for medical practitioners. Unfortunately, DL models often operate like black boxes, so various techniques, including Grad-CAM, were developed<sup>25</sup>. The range of layers to be fine-tuned is an important factor in achieving optimal performance on a new task in transfer learning. Figure 5 shows the Grad-CAM images with various freezing ranges and training epochs. When the fine-tuning degree of a DL model is zero or one, there is a tendency for the heatmap to appear in the background or outside the implant area. As the fine-tuning



**Figure 4.** Scatter plot of the clustering analysis: (a) scatter plot of the feature vectors for clustering; (b) scatter plot of the final clustering model with the color code and representation of centroids of clusters (yellow circles).



**Figure 5.** Grad-CAM images for different fine-tuning degrees and training epochs in deep learning approach.

degree increases, there is a tendency for the heat map to concentrate and appear on the implant area. However, training a CNN architecture with only a small amount of data can result in overfitting and limited generalization. Therefore, when applying a DL model to limited data, it is crucial to perform appropriate fine-tuning<sup>28</sup>.

In applying a clustering analysis, a feature vector composed of the radiographic diameter and length of the dental implant was used in this study. All images were visualized in the scatter plot of the feature vectors, so it was possible to interpret how closely the feature vector of the dental implant image was positioned to the centroid or if it was positioned between two different clusters. This scatter plot serves as a map-like guide for dividing implant groups, assisting in the explainability of the clustering model.

In the case of using deep learning methods for bone level implants, there were two false predictions. In these cases, either the length and diameter were incorrectly predicted to be larger, or the length was predicted correctly, but the diameter was incorrectly predicted. In the case of using clustering methods for bone level implants, there were five false predictions. The diameter was predicted correctly in these cases, but the length was predicted incorrectly. The clustering model exhibited reliable results in inferring the diameter. If this trend is similarly observed across a wider range of implant systems and more datasets on external validation, there is a possibility that using both the DL model and clustering model in tandem for analyzing a single implant image could serve complementary roles. In cases where the predicted diameter and length of an implant differ between the two models, it is conceivable to give precedence to the predictions of the clustering model, which is speculated to be more reliable for diameter prediction. Subsequently, confirming the vector corresponding to the radiographic diameter and length of that image at the position of the feature map, and then comparing the predicted lengths from both the DL model and clustering model, while simultaneously considering distortions caused by the image's radiographic angulation, would facilitate a more rational decision-making process.

In the study, the performances of two AI models were improved through the tuning process. The fine-tuning degree was chosen as four through external validation with test data in the DL approach. When comparing the results of the tuning process before and after, based on the AUC-ROC values for the test dataset, a statistically significant performance improvement was observed in six out of the nine groups after tuning, when the tuning degree was set to four. In the clustering approach, comparing the results of the weight of the feature vector set to one and two, based on the AUC-ROC values for the test dataset, four of the nine groups demonstrated a statistically significant improvement in performance after tuning. As far as we know, no research has been conducted on utilizing deep learning or clustering approaches to classify the sizes of dental implants. Therefore, we could not compare the performance of our study's results with those of other studies.

This study had limitations. First, we utilized a limited set of images from periapical radiographs with Straumann bone level implants. Due to the limitations in data availability, our study utilized only periapical radiographs of a limited dataset of Straumann bone level implants to classify dental implant sizes using deep learning and clustering methods. Subsequent research should encompass more diverse dental implant systems and account for complex clinical scenarios to evaluate the performance of these AI models. For generalizability, the model should be trained on a large and diverse dataset that adequately represents the variations and complexities of the target problem<sup>29</sup>. Recently, a study identified 25 different systems of dental implants using 37,442 periapical and 113,291 panoramic images<sup>24</sup>. Also, for enhancing the reliability and reproducibility of AI models detecting and classifying the diameter and length of various dental implants, further investigation is crucial to establish a

comprehensive, large-scale dataset including various dental implant systems and diameters and lengths of dental implants. Second, the dataset included only a periapical modality. In clinical practice, the limited information may include radiographs from different modalities, such as panoramic radiographs. Therefore, when constructing a dataset and conducting further studies, it is necessary to expand the scope by including radiographs from other modalities. Third, to evaluate the improvement in model performance, we compared the results before and after the tuning process for each AI approach<sup>30,31</sup>. For clinical validity assessment, it is necessary to compare the performance of dental professionals across multiple groups based on their experience, or it would be necessary to investigate how much these automated systems can enhance the decision performance of dental professionals. In previous studies focusing on classifying dental implant systems, they presented the results comparing the performance of AI models and dental professionals. In many of these studies, experts often conducted performance evaluations on human experts assuming prior knowledge in distinguishing dental implant systems. Dental professionals are skilled at distinguishing dental implant systems. However, progressing beyond implant system differentiation, distinguishing the diameter and length of these implants requires even more prior knowledge. Therefore, presenting human experts with a comparable quantity of training images annotated with dental implant sizes, akin to AI models, and accurately and efficiently evaluating the extent to which their performance improves, is imperative. Fourth, training models with imbalanced data can cause bias and performance deterioration in minority groups<sup>32</sup>. The model tends to prioritize the majority class due to its higher prevalence, reducing performance for the underrepresented classes. To overcome these limitations, future studies should aim to collect additional data from minority groups. Additionally, utilizing synthesized data from suitable generative models could effectively address these limitations<sup>33,34</sup>. Fifth, it will be essential to conduct tests to determine which scenarios deal with various dental implant systems and sizes with various AI approaches to find more effective methods. The results of this paper discussed the model performance of the fine-tuned VGG16. While our initial testing of ResNet50 and InceptionV3 in the study yielded performance lower than that of VGG16, they still exhibited reliable performance. ResNet50 and InceptionV3 demonstrated consistent performance not only in our study but also across various other research studies. Future research is needed to focus on a wider range of architectures, highlighting deep learning using various CNN architectures, such as DenseNet, MobileNet, EfficientNet, Xception, ResNeXt, SENet, and RegNet.

The AI models developed in this study have limitations in detecting various dental implant systems and different diameters and lengths. However, this research went beyond previous studies that focused on classifying implant systems and aimed to provide more detailed information about dental implants from periapical radiographs. By further advancing this approach, it has the potential to efficiently manage patients with dental implants in a clinical setting and offer an objective metric reference for dental radiograph interpretation.

## Conclusion

Automatic classification of the size of bone level implants can be achieved through DL and clustering. The performances of two AI models were improved through the tuning process. DL involves obtaining features through a training process with transfer learning and fine-tuning. On the other hand, the clustering model was developed by selecting an appropriate feature using three key points and tuning the weight of the feature vector. As a result, they can improve the efficiency and accuracy of implant assessment, assist dental professionals in making informed decisions, and enhance patient outcomes in dental implant treatments.

## Materials and methods

### Ethics

The Institutional Review Board (IRB) of Yonsei University Dental Hospital approved this study (Approval number: 2-2022-0067). The IRB of Yonsei University Dental Hospital waived the requirement to obtain individual informed consent, so no participants were provided written or verbal informed consent since this study had a noninterventive retrospective design, and all data were evaluated anonymously. All methods in this study were performed in accordance with the relevant guidelines and regulations.

### Data acquisition and data splitting

This study focused on the Straumann bone level implant (Institut Straumann, Basel, Switzerland). Periapical radiographs of Straumann bone level implants were included based on the inclusion criteria, which allowed for verification of diameter and length from the electronic medical records of dental implant first surgery. To replicate diverse clinical scenarios, the included periapical radiographs contained loaded implants, implants with healing abutments, and implants with cover screws. Cases, where dental implant images exhibited significant blurring or distortion due to movement during imaging or unusual imaging angles, were excluded from the study. We collected 927 periapical radiographs obtained from 874 patients aged 19–85 who underwent periapical radiography using the paralleling technique with 60 kVp, 7 mA, and 0.08–0.1 s between 2005 and 2022.

All periapical radiographs were cropped to display one dental implant per image. After cropping 927 periapical radiographs, 1320 images containing bone level implants were obtained (Fig. 1a). Each image was annotated with a four-digit number representing the dental implant's diameter and length, with the first two digits corresponding to the diameter and the last two corresponding to the length. A board-certified prosthodontist performed the process. Subsequently, cropped images underwent a thorough verification process. This involved cross-referencing each image with the corresponding electronic medical record entry. Initially, the process encompassed labeling to ascertain the number of threads, the characteristic appearance of the implant apex, as well as the proportional representation of diameter and length on the image itself. This comprehensive approach ensured the refinement of erroneously annotated data. The resulting process yielded a meticulously curated ground truth dataset. For the DL process, to prevent class imbalance, the validation dataset and testing dataset

were structured with an equal distribution of 20 data points per class, while the remaining data was designated as the training dataset. The entire 1320 images were divided into 960 images for training, 180 images for validation, and 180 images for testing (Supplementary Table S1). For the clustering process, the entire 1320 images were divided into 1140 images for training and 180 images for testing (Supplementary Table S2). The testing dataset used for the performance evaluation of both AI models consisted of the same set of 180 images. The study was conducted according to the checklist regarding AI in dental research<sup>31</sup>.

### DL approach for bone level implant size classification

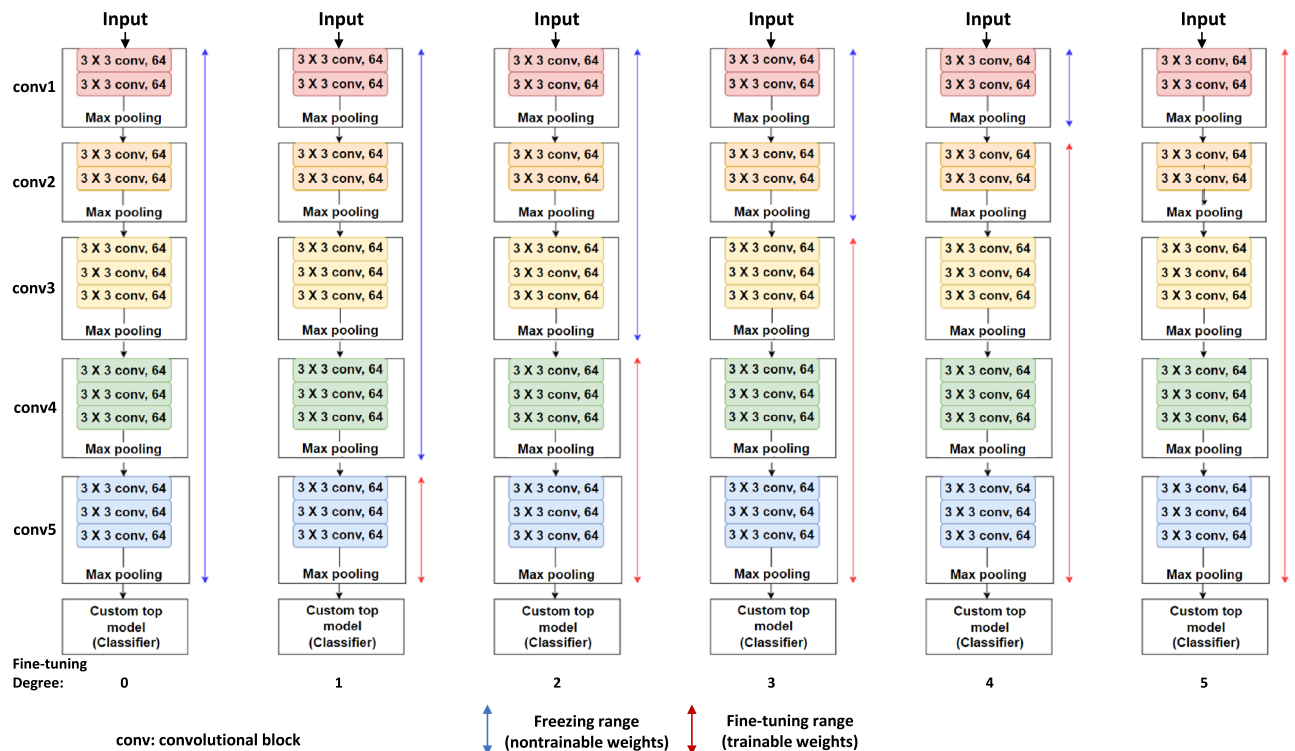
The cropped images were preprocessed with global contrast normalization and zero-phase whitening. The diversity of the training data was increased by applying various transformations to the images. The training data was randomly augmented using horizontal flipping, vertical flipping, rotation (range of  $\pm 10^\circ$ ), width shifting (range of  $\pm 0.1$ ), height shifting (range of  $\pm 0.04$ ), zooming (range of  $\pm 0.02$ ), and shear (range of  $\pm 0.01$ ). As a result, the training data has been augmented by 16 times. The input image size was 224 by 224.

VGG16, ResNet50, and InceptionV3 were tested as potential CNN architectures. The results showed accuracy values of 0.989, 0.961, and 0.967, respectively, with the test dataset, and VGG16 was chosen<sup>35</sup>. VGG16 model, pre-trained with ImageNet, was used for transfer learning<sup>36</sup>. VGG16 architecture comprises five convolutional blocks. In this study, the fine-tuning degree, in the range of zero to five, was defined as the number of blocks, which was trained with our training data, and the other convolutional blocks were frozen as the weight of ImageNet (Fig. 6). External validation was performed on six fine-tuned VGG16 models using the test data. The model achieving the best accuracy was chosen for the final DL model<sup>37</sup>.

Early stopping was used with the patience set to five to prevent overfitting. The learning rate was 0.00002 using the Adam optimizer. The batch size was set to 16. Dropout regularization with a rate of 0.5 was applied to prevent overfitting. The DL models were trained for 15 epochs with a possibility of early stopping and developed in Python 3.8.16 and TensorFlow 2.9.2 under a single NVIDIA RTX 3090.

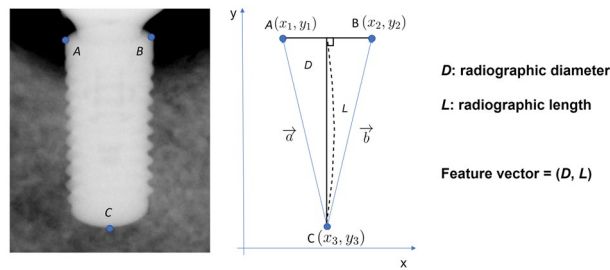
### Clustering approach for bone level implant size classification

Three key points (points A, B, and C) were labeled on cropped bone level implant images and used to calculate the radiographic diameter and length (Fig. 7). In this study, the diameter of the dental implant is the implant body, except for thread depth.  $\overline{AB}$  corresponds to the diameter of the dental implant on a periapical radiograph, and point C corresponds to the dental implant apex. Key point annotation was performed using the annotation tool LabelMe (MIT Computer Science and Artificial Intelligence Laboratory, retrieved from <https://github.com/wkentaro/labelme>). The type of annotation was a polygon. The  $x$ - and  $y$ -coordinates of the three key points were extracted from a JSON format file.  $D$  means radiographic diameter,  $L$  means radiographic length, and  $S$  means the area of triangle ABC yields (Eqs. (1)–(5)).



**Figure 6.** Six cases of the pre-trained VGG16 model by adjusting the fine-tuning degree.





**Figure 7.** Key point selection and feature extraction for clustering.

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

$$S = \frac{1}{2}DL \quad (2)$$

$$S = \frac{1}{2} |\vec{a} \times \vec{b}| \quad (3)$$

$$S = \frac{1}{2} |x_1y_2 + x_2y_3 + x_3y_1 - x_2y_1 - x_3y_2 - x_1y_3| \quad (4)$$

$$L = \frac{1}{D} |x_1y_2 + x_2y_3 + x_3y_1 - x_2y_1 - x_3y_2 - x_1y_3| \quad (5)$$

Through the above equations,  $D$  and  $L$  were calculated in pixel units, so they were changed to the values in millimeters by multiplying themselves by the imager pixel spacing value, a DICOM metadata element, corresponding to the physical distance measured at the front plane of the image receptor housing between the centers of adjacent pixels.  $D$  and  $L$  were extracted from each image as a feature vector to be used for clustering.

The radiographic diameter and length indicated on the  $x$ - and  $y$ -axes, respectively, and the ground truth cluster of each data was represented with color coding. The three key points in each of the 1320 images of bone level implants were used to calculate the radiographic diameter and length for making a feature vector to train and test the clustering model.

This study used a type of  $k$ -means clustering called  $k$ -means++, which can improve the clustering process produced by the  $k$ -means algorithm by selecting centroids that are well-distributed throughout the data set<sup>38,39</sup>. The centroid represents the central point of a cluster. The centroid serves as a representative value or can be used to measure distances between clusters. As the number of clusters,  $k$  was set to nine because the study aimed to classify nine groups of the different diameters and lengths of the dental implants. A two-dimensional coordinate space was used with the radiographic diameter and length set as the  $x$ - and  $y$ -axes, respectively (Fig. 7). The study aimed to find the centroid for assigning clusters well to the ground-truth group from nine clusters by tuning the weight of the feature vector.

$$f : R^2 \rightarrow R^2, \quad \text{where } f(D, L) = (wD, L) \text{ and } w > 0 \quad (6)$$

Function  $f$  denotes a linear transformation that modifies the diameter component of a vector by scaling with a factor  $w$  while keeping the vertical component unchanged (Eq. (6)). This transformation stretched or compressed the vector along the horizontal axis while keeping its length constant along the vertical axis.

The changes in the vector space according to the value of  $w$  and the resulting  $k$ -means clustering outcomes were evaluated. This research set  $w$  within the range of one to 10, and we adjusted it to improve the performance of the clustering model. The same computing resources used to develop the DL model were also employed to develop the clustering model.

### Statistical analysis and model performance evaluation

Statistical analysis was performed with the Python sklearn library and Stata software (StataCorp, College Station, TX) version 18. With the results of the two final models with DL and clustering approaches, the accuracy, sensitivity, specificity, F1-score, positive predictive value (PPV), negative predictive value (NPV), and AUC-ROC were calculated based on the confusion matrix (Eqs. (7)–(12), TP: true positive, TN: true negative, FP: false positive, FN: false negative, PPV: positive predictive value, NPV: negative predictive value. The recall is also known as sensitivity, and the precision is also known as PPV).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (9)$$

$$F1 - \text{score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (10)$$

$$PPV = \frac{TP}{TP + FP} \quad (11)$$

$$NPV = \frac{TN}{TN + FN} \quad (12)$$

Comparisons of performances were conducted in each two AI approaches between the model with the best accuracy and the model before tuning process. The chi-square test was used as a statistical test, and the statistical significance level was set to  $p = 0.05$ .

### Data availability

The data sets generated or analyzed during the current study are not publicly available in order to preserve patient confidentiality but are available from the corresponding authors on reasonable request.

Received: 14 June 2023; Accepted: 9 September 2023

Published online: 06 October 2023

### References

- Howe, M. S., Keys, W. & Richards, D. Long-term (10-year) dental implant survival: A systematic review and sensitivity meta-analysis. *J. Dent.* **84**, 9–21 (2019).
- Albrektsson, T. & Donos, N. Implant survival and complications. The Third EAO consensus conference 2012. *Clin. Oral Implants Res.* **23**(Suppl 6), 63–65 (2012).
- Jung, R. E., Zembic, A., Pjetursson, B. E., Zwahlen, M. & Thoma, D. S. Systematic review of the survival rate and the incidence of biological, technical, and aesthetic complications of single crowns on implants reported in longitudinal studies with a mean follow-up of 5 years. *Clin. Oral Implants Res.* **23**(Suppl 6), 2–21 (2012).
- Papaspyridakos, P., Chen, C. J., Singh, M., Weber, H. P. & Gallucci, G. O. Success criteria in implant dentistry: A systematic review. *J. Dent. Res.* **91**, 242–248 (2012).
- Berglundh, T. *et al.* Peri-implant diseases and conditions: Consensus report of workgroup 4 of the 2017 world workshop on the classification of periodontal and peri-implant diseases and conditions. *J. Periodontol.* **89**, S286–S291 (2018).
- Schwarz, F., Derks, J., Monje, A. & Wang, H.-L. Peri-implantitis. *J. Periodontol.* **89**, S267–S290 (2018).
- Monje, A. *et al.* Morphology and severity of peri-implantitis bone defects. *Clin. Implant Dent. Relat. Res.* **21**, 635–643 (2019).
- Jokstad, A. & Ganeles, J. Systematic review of clinical and patient-reported outcomes following oral rehabilitation on dental implants with a tapered compared to a non-tapered implant design. *Clin. Oral Implants Res.* **29**(Suppl 16), 41–54 (2018).
- Esteve, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
- Pereira, S., Pinto, A., Alves, V. & Silva, C. A. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* **35**, 1240–1251. <https://doi.org/10.1109/TMI.2016.2538465> (2016).
- Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
- Lee, J. G. *et al.* Deep learning in medical imaging: General overview. *Korean J. Radiol.* **18**, 570–584 (2017).
- Yang, S. Y. *et al.* Histopathology-based diagnosis of oral squamous cell carcinoma using deep learning. *J. Dent. Res.* **101**, 1321–1327 (2022).
- Mohammad-Rahimi, H. *et al.* Deep learning in periodontology and oral implantology: A scoping review. *J. Periodontol. Res.* **57**, 942–951 (2022).
- Krois, J. *et al.* Deep learning for the radiographic detection of periodontal bone loss. *Sci. Rep.* **9**, 8495. <https://doi.org/10.1038/s41598-019-44839-3> (2019).
- Lee, S. *et al.* Deep learning for early dental caries detection in bitewing radiographs. *Sci. Rep.* **11**, 16807. <https://doi.org/10.1038/s41598-021-96368-7> (2021).
- Putra, R. H., Doi, C., Yoda, N., Astuti, E. R. & Sasaki, K. Current applications and development of artificial intelligence for digital dental radiography. *Dentomaxillofac. Radiol.* **51**, 20210197. <https://doi.org/10.1259/dmfr.20210197> (2022).
- Michelinakis, G., Sharrock, A. & Barclay, C. W. Identification of dental implants through the use of implant recognition software (IRS). *Int. Dent. J.* **56**, 203–208 (2006).
- Lee, J. H., Kim, Y. T., Lee, J. B. & Jeong, S. N. A performance comparison between automated deep learning and dental professionals in classification of dental implant systems from dental imaging: A multi-center study. *Diagnostics (Basel)* <https://doi.org/10.3390/diagnostics10110910> (2020).
- Kim, J. E. *et al.* Transfer learning via deep neural networks for implant fixture system classification using periapical radiographs. *J. Clin. Med.* **9**, 1117. <https://doi.org/10.3390/jcm9041117> (2020).
- Lee, J. H. & Jeong, S. N. Efficacy of deep convolutional neural network algorithm for the identification and classification of dental implant systems, using panoramic and periapical radiographs: A pilot study. *Medicine (Baltimore)* **99**, e20787. <https://doi.org/10.1097/md.00000000000020787> (2020).
- Takahashi, T. *et al.* Identification of dental implants using deep learning-pilot study. *Int. J. Implant Dent.* **6**, 53. <https://doi.org/10.1186/s40729-020-00250-6> (2020).
- Kim, H. S. *et al.* Transfer learning in a deep convolutional neural network for implant fixture classification: A pilot study. *Imaging Sci. Dent.* **52**, 219–224 (2022).

24. Park, W., Schwendicke, F., Krois, J., Huh, J. K. & Lee, J. H. Identification of dental implant systems using a large-scale multicenter data set. *J. Dent. Res.* <https://doi.org/10.1177/00220345231160750> (2023).
25. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*. 618–626. <https://doi.org/10.1109/ICCV.2017.74> (2017)
26. Revilla-León, M. *et al.* Artificial intelligence applications in implant dentistry: A systematic review. *J. Prosthet. Dent.* **129**, 293–300 (2023).
27. Cha, J. Y., Yoon, H. I., Yeo, I. S., Huh, K. H. & Han, J. S. Peri-implant bone loss measurement using a region-based convolutional neural network on dental periapical radiographs. *J. Clin. Med.* <https://doi.org/10.3390/jcm10051009> (2021).
28. Tajbakhsh, N. *et al.* Convolutional neural networks for medical image analysis: Full training or fine tuning?. *IEEE Trans. Med. Imaging* **35**, 1299–1312. <https://doi.org/10.1109/tmi.2016.2535302> (2016).
29. Schwendicke, F., Samek, W. & Krois, J. Artificial intelligence in dentistry: Chances and challenges. *J. Dent. Res.* **99**, 769–774 (2020).
30. Mongan, J., Moy, L., Charles, E. & Kahn, J. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiology Artificial Intelligence* **2**, e200029. <https://doi.org/10.1148/ryai.2020200029> (2020).
31. Schwendicke, F. *et al.* Artificial intelligence in dental research: Checklist for authors, reviewers, readers. *J. Dent.* **107**, 103610 (2021).
32. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
33. Kohlakala, A., Coetzer, J., Bertels, J. & Vandermeulen, D. Deep learning-based dental implant recognition using synthetic X-ray images. *Med. Biol. Eng. Comput.* **60**, 2951–2968 (2022).
34. Kim, M. *et al.* Synthesizing realistic high-resolution retina image by style-based generative adversarial network and its utilization. *Sci. Rep.* **12**, 17307. <https://doi.org/10.1038/s41598-022-20698-3> (2022).
35. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
36. Kim, H. E. *et al.* Transfer learning for medical image classification: A literature review. *BMC Med. Imaging* **22**, 69. <https://doi.org/10.1186/s12880-022-00793-7> (2022).
37. Lee, K.-S. *et al.* Evaluation of scalability and degree of fine-tuning of deep convolutional neural networks for COVID-19 screening on chest X-ray images using explainable deep-learning algorithm. *J. Pers. Med.* **10**, 213 (2020).
38. Kapoor, A. & Singhal, A. A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms. In *3rd International Conference on Computational Intelligence & Communication Technology*. 1–6. <https://doi.org/10.1109/CICT.2017.7977272> (2017).
39. Yoder, J. & Priebe, C. E. Semi-supervised k-means++. *J. Stat. Comput. Simul.* **87**, 2597–2608 (2017).

## Acknowledgements

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI21C1716).

## Author contributions

J.-H.P. contributed to conception and design, data acquisition, analysis, and interpretation, drafted and critically revised manuscript; H.S.M. and H.-I.J. contributed to data analysis, data interpretation, and critically revised manuscript; J.H. and Y.-H.C., contributed to data interpretation, critically revised manuscript; J.-E.K. contributed to conception and design, data acquisition, analysis, and interpretation, critically revised manuscript; All authors gave their final approval and agree to be accountable for all aspects of the work.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-42385-7>.

**Correspondence** and requests for materials should be addressed to J.-E.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023