IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 February 2025, accepted 12 March 2025, date of publication 17 March 2025, date of current version 26 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3552203

RESEARCH ARTICLE

Assessment of the Impact of Missing Teeth on Deep Learning-Based Tooth Numbering Method From Panoramic X-Rays

ZHONGBO TANG[1], (Member, IEEE), TOMOAKI MAMENO[2], TOMONORI HAYAMI[3], CHONHO LEE[4], SUSUMU DATE[3],(Member, IEEE), KAZUNORI IKEBE[2], AND KAZUNORI NOZAKI[5]

[1]Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan
[2]Department of Prosthodontics, Gerodontology, and Oral Rehabilitation, Graduate School of Dentistry, Osaka University, Osaka 565-0871, Japan
[3]D3 Center, Osaka University, Osaka, Ibaraki 567-0047, Japan
[4]Department of Information Science and Engineering, Okayama University of Science, Okayama 700-0005, Japan
[5]Division for Oral Dental Informatics, Osaka University Dental Hospital, Osaka 565-0871, Japan

Corresponding author: Kazunori Nozaki (nozaki.kazunori.dent@osaka-u.ac.jp)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Graduate School of Dentistry, Osaka University Ethics Committee under Application R4-E25.

ABSTRACT Automatically analyzing panoramic dental X-rays through AI-based approaches has attracted attention and concern among dentists with the recent advancement of AI techniques. Recent studies have shown that these approaches have achieved high accuracy in recognizing and numbering teeth. However, the datasets utilized in such studies contain only a small proportion of panoramic X-rays that include missing teeth, despite such X-rays often being observed in actual diagnosis. Accordingly, the suitability of these approaches for practical panoramic X-rays remains unconfirmed. This study verifies the efficacy of three powerful and efficient object detection models for tooth numbering on panoramic X-rays with varying numbers of missing teeth. The experimental results indicate that the performance of AI models is affected by the existing tooth count appearing in panoramic X-ray dataset images, and images with a lower tooth count tend to exhibit a lower tooth-numbering accuracy that reaches a precision of 62.2%. Additionally, our study highlights the features in low tooth-count images that influence this accuracy. Our results thus also verify the unsuitability of AI models in numbering on panoramic X-ray images with a low tooth count. Importantly, this study suggests that disclosing the tooth count distribution of datasets is indispensable for proving the effectiveness of AI-based approaches in the tooth numbering task.

INDEX TERMS Artificial intelligence, deep learning, dental informatics, dental panoramic X-rays, medical informatics application, orthodontics, tooth numbering.

I. INTRODUCTION

The panoramic dental X-ray is an effective diagnostic imaging technique extensively utilized in dental clinics. It provides comprehensive information for assessing various dental conditions and pathologies, including the alveolar bone resorption due to periodontal disease, the presence and condition of impacted and supernumerary teeth, various tumors of the jaws, the status of the temporomandibular joints, and the sinuses [1], [2], [3], [4], [5]. The findings from panoramic dental X-ray images are diverse. To incorporate these findings from dental diagnoses into dental records, tooth numbering, which involves assigning a unique ID to each tooth in a panoramic X-ray image, is indispensable. However, manually processing tooth numbering on panoramic X-rays necessitates extensive expertise and time commitment from dental professionals. Hence, there is a high demand for an automatic and intelligent system that assists dentists with tooth numbering in order to free up their precious time for more valuable clinical diagnosis.

Recent advances in deep-learning techniques have facilitated the application of diverse object detection models

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

TABLE 1. Overview of recent research on tooth numbering.

References	Test-set	Implants	Inst.	Avg.	Approach	Results
Tuofzz et al. [6]	222	excluded	5053	22.8	Faster R-CNN, VGG-16, Heuristic method	Sensitivity 98.00%, Specificity 99.94%
Bilgir et al. [7]	249	included	7190	28.9	Faster R-CNN Inception v2	Sensitivity 95.59%, Precision 96.52% F-measure 96.06%
Estai et al. [8]	unclear	excluded	1824	29.0	UNet-based segmentation, Faster R-CNN VGG-16	Recall 98.03%, Specificity 99.93% F1 score 98.03%
Karaoglu et al. [9]	447	included	12567	28.1	Mask R-CNN, Heuristic method	mAP 92.49%, Precision 96.08% Recall 95.65%, F1-score 95.87%
Yilmaz et al. [10]	200	excluded	5265	26.3	YOLOv4, Faster R-CNN	Precision 99.90%, Recall 99.18% F1 score 99.54%
Ali et al. [11]	470	included	12497	26.6	YOLOv7, Knowledge-based algorithm	mAP 98.20%
Putra et al. [12]	100	excluded	2970	29.7	YOLOv4	Accuracy 88.5%, Precission 87.7% Recall 100.0%, F1 score 93.4%

for automated tooth detection and numbering in dental panoramic X-rays. Table 1 summarizes the test dataset size, whether implants were included, instance count (total tooth count), average tooth count per image, approach, and evaluation metric results of recent research. In a study, Tuofzz et al. [6] first utilized faster-RCNN [13] for tooth detection and subsequently a convolutional neural network, VGG-16 [14], for tooth number classification. Their approach for tooth detection achieved a sensitivity (recall) of 99.41% and a precision of 99.45% using their own datasets. Yilmaz et al. [10] investigated the performance with faster-RCNN and a detection model, YOLOv4 [15], and they showed a precision of 93.67% and 99.90%, respectively. Their work indicated that YOLOv4 outperformed faster-RCNN in a tooth numbering task. Additionally, Ali et al. [11] utilized a CNN-based object detector, YOLOv7 [16], model and prior knowledge-based algorithm in tooth detection with six-fold cross validation. They observed an impressive mean average precision (mAP) of 99.20%. Putra et al. [12] assessed the performance of YOLOv4 for automated tooth numbering. The accuracy, precision, recall, and F1-score for tooth detection and numbering were 88.5%, 87.7%, 100.0%, and 93.4%, respectively.

All of these works demonstrated the effectiveness of their proposed approaches using their own private datasets gathered from regional dental hospitals. Since differences between private datasets exist, the versatility of their approaches is unclear. Due to the lack of detailed information on these datasets, we compared the differences between them using the average tooth count, which is shown in Table 1. An adult can have a maximum of thirty-two teeth, and the average tooth count of datasets used in these works reveals that the reported results are limited.

Furthermore, missing teeth are usually accompanied by two features, increasing the complexity of tooth numbering on panoramic X-rays. The first feature is the increased likelihood of dental implants being present. Dental implants are prosthetic devices designed to replace missing teeth, serving as artificial tooth roots that support replacement teeth. The second feature is an edentulous space (ES), which is a space composed of adjacent missing teeth. The presence of ES is an inevitable consequence of tooth loss. Fig. 1 illustrates an example of ES. Normal teeth are denoted by a solid line, missing teeth are denoted by a dashed line, and the area enclosed by a red line is ES. Although Table 1 shows that that some datasets include implants, the performance results with implants are not explicit. Assessing the influence of these features is essential for comprehensively evaluating panoramic X-rays with varying numbers of missing teeth.

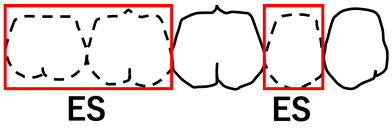


FIGURE 1. Example of edentulous space (ES).

Our study aims to evaluate the efficacy of applying deep learning-based approaches to tooth numbering on panoramic X-rays with varying numbers of missing teeth. Experiments are conducted and evaluated on datasets grouped by different tooth counts. Also, we analyze the two features accompanying missing teeth and assess their influence on tooth numbering performance.

II. MATERIALS AND METHODS

Our experiments were conducted on an Ubuntu server with a version 22.04.1 LTS operating system. The server was equipped with one NVIDIA TITAN RTX GPU containing 24 GB of VRAM.

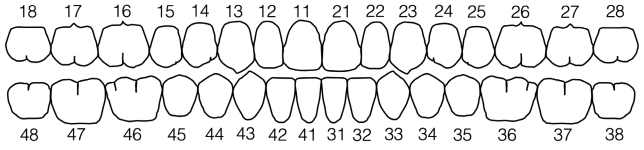


FIGURE 2. FDI numbering system.

A. FDI NUMBERING SYSTEM

Fig. 2 illustrates an overview of the FDI numbering system. Each tooth is assigned with a unique two-digit number. In the tooth numbering task in this study, all teeth in panoramic X-rays are labeled with an FDI number.

B. DEFINITION OF TOOTH-COUNT GROUPS

The tooth count of an image is defined as the number of teeth present in a panoramic X-ray image. This study aimed to verify the efficacy of deep learning-based models on panoramic X-rays with varying numbers of missing teeth. Therefore, we categorized the panoramic X-rays into eight groups based on the tooth count of each image, as illustrated in Table 2.

TABLE 2. Tooth-count groups list.

Group No.	Tooth count	Group No.	Tooth count
1	1-4	5	17-20
2	5-8	6	21-24
3	9-12	7	25-28
4	13-16	8	29-32

The eight groups were adopted to maintain a minimum image count (over four) for training set sampling. Sampling from each group ensured that the training set contained images from each group, reducing potential effects caused by an unbalanced number of images from each group. Also, categorizing the test set utilized in this study into eight groups was effective for maintaining an acceptable number of test cases. Ensuring a sufficient number of test cases helps to reduce instability and enhances reliability of evaluation results.

C. DATASET

Two databases were used in this experiment.

- Osaka University Dental Hospital Database (OUD): 307 panoramic X-ray images of voluntary patients receiving treatments in the Division of Oral Rehabilitation and Geriatric Dentistry including Implantology, Osaka University Dental Hospital. All images were annotated by a unique dentist. Each tooth in the images was enclosed by a bounding box with three labels. These labels contained an FDI [17] tooth number, a boolean value for the existence of tooth pulp, and the type of dental treatment, such as implants and composite fillings.
- Tufts Dental Database (TDD) [18]: 1000 panoramic X-rays with multi-modal annotation. It consists of 885 panoramic X-ray images with only permanent teeth, 114 panoramic X-ray images with both permanent teeth and deciduous teeth, and 1 panoramic X-ray image without teeth. Tooth number annotations with bounding boxes from TDD were used in this study.

Panoramic X-rays containing fixed partial dentures, supernumerary teeth and deciduous teeth were excluded in this study due to an insufficient number of test images for reliable evaluation. Also, X-rays without any teeth, which have no influence on experimental results, were excluded. Since Gulum et al. [19] mentioned the effect of data size on tooth numbering, and large numbers of samples are essential for model training in order to achieve a reliable evaluation. TDD has an acceptable data size for model training, containing 885 panoramic X-rays. Additionally, the annotated tooth numbers in TDD did not use the FDI system; thus, the original TDD tooth numbers were converted to the FDI system.

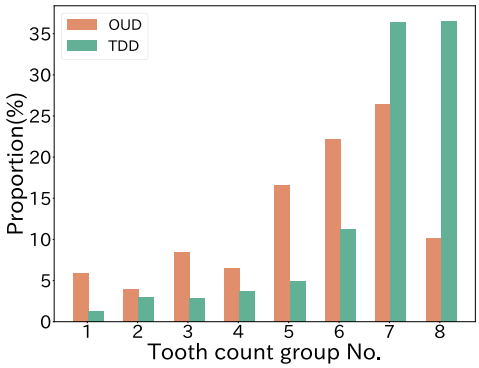


FIGURE 3. Difference in proportion of tooth-count group between OUD and TDD.

The average tooth count of OUD is 20.4, which is significantly lower than that of datasets used in recent research. Fig. 3 shows the proportion of each tooth-count group of OUD and TDD. OUD contains a higher proportion of images belonging to low tooth-count groups (Groups 1-4). There is not that huge of a difference in the distribution in tooth-count groups for OUD compared with TDD, so OUD is suitable for assessing the influence of tooth count on deep learning-based models. Therefore, 885 panoramic X-ray images from TDD were used as a training set and a validation set, and 307 panoramic X-ray images from OUD were used for testing in this experiment.

D. PROCESSING PIPELINE

The entire processing pipeline is demonstrated in Fig. 4. It involves model training, tooth detection, and detection results filtering. We tried three powerful and efficient object detection models, YOLOv8 [17], YOLOv11, and Real-Time Detection Transformer [20] (RT-DETR) for verification. All models were trained to number teeth using thirty-two labels.

1) TRAINING MODEL

The composition of the training set and the validation set is demonstrated in Fig. 4. Stratified random sampling was used in this study to ensure that an appropriate proportion of images belonged to the different tooth-count groups. This sampling method first splits data into various sub-groups by one feature and randomly samples from these sub-groups in order to maintain the representation of all sub-groups. In this experiment, all X-ray images from TDD were divided into eight groups in accordance with their tooth-count group. X-ray images in each group were randomly split into a training set and a validation set at a 3:1 ratio. The resulting training and validation sets were composed of the training and validation sets from each tooth-count group, respectively.

2) TOOTH DETECTION

During the tooth detection phase, the 307 images in OUD were used to evaluate models trained by TDD. Each bounding

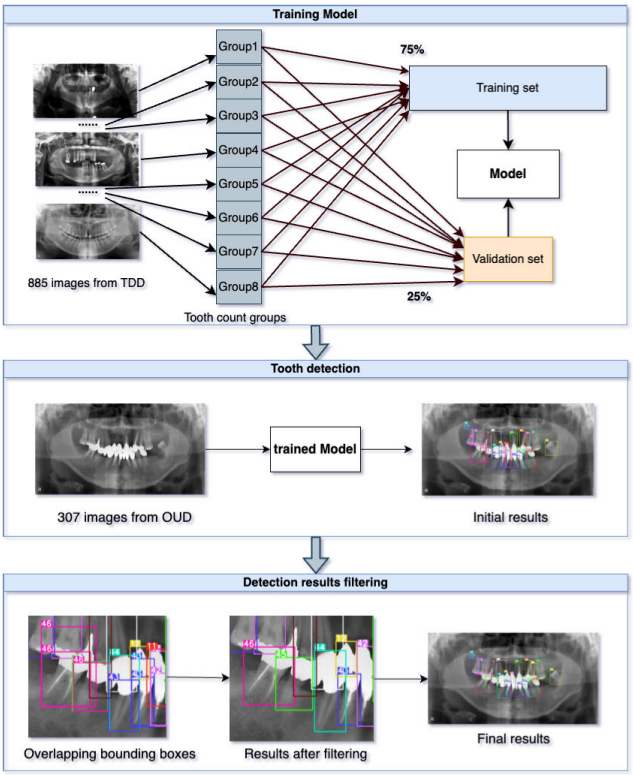


FIGURE 4. Overview of experiment pipeline.

box was attached with a predicted label (FDI tooth number) on the top left corner and a confidence score (hidden for clarity) that indicated the reliability of the predicted bounding box and the predicted label. The color of the box varied from its predicted tooth number in order to facilitate visual discrimination between different tooth numbers.

The models generated multiple predictions for a single object, resulting in overlapping bounding boxes enclosing an identical tooth. Since each box has one predicted label, overlapping boxes will label multiple predicted tooth numbers to the identical tooth, leading to inaccurate numbering. Therefore, filtering the overlapping boxes is essential to eliminate redundant predictions and ensure accurate numbering.

3) DETECTION RESULTS FILTERING

The bottom part of Fig. 4 illustrates the changes before and after filtering out the overlapping bounding boxes generated by the object detection models. For each tooth enclosed by multiple boxes, we selected the box with the highest confidence score as the final numbering result, ensuring no more overlapping boxes remained on the identical tooth.

E. EVALUATION METRICS

1) METRICS FOR OBJECT DETECTION

Intersection over union (IoU), precision, recall, and F1-score are common metrics when evaluating object detection models. All these metrics were calculated using the Ultralytics library. IoU measures the degree of predicted bounding boxes matching with the ground truth (real teeth). Equation (1) is defined to calculate the IoU. The areas of the ground truth bounding box and the predicted bounding box are denoted as S_{truth} and $S_{\text{predicted}}$, respectively. S_{inter} indicates the overlapping areas between S_{truth} and $S_{\text{predicted}}$. The default minimum IoU threshold is set to 0.5. Consequently, if the IoU between a ground-truth bounding box and a predicted one falls below this threshold, the prediction is deemed invalid.

$$\text{IoU} = \frac{S_{\text{inter}}}{S_{\text{truth}} + S_{\text{predicted}} - S_{\text{inter}}} \tag{1}$$

Precision (P) and recall (R) are defined in (2) and (3), respectively. TP denotes the number of predicted boxes that overlap with the ground truth boxes with an IoU of over 0.5 and with correct label prediction. FP denotes the number of predicted boxes that overlap with the ground truth boxes with an IoU below 0.5 or with incorrect label prediction. FN denotes the number of ground truths that are not detected.

P indicates the proportion of bounding boxes that were correctly predicted among all predicted bounding boxes with the same tooth number.

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

R reveals how well the model can detect for a specific tooth number.

$$\text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

F1-score (F1) is a harmonic mean of precision and recall, defined by (4). It ranges from 0 to 1, with 1 being the best possible score and 0 the worst.

$$\text{F1} = \frac{2 \times (P \times R)}{P + R} \tag{4}$$

2) METRICS FOR TOOTH NUMBERING

The detection rate (DR) and numbering accuracy (NA) were used to evaluate the tooth numbering performance instead of metric precision and recall calculated by using the Ultralytics library. DR is calculated using (5), defining the ratio of the detected tooth count N_{detected} to the total tooth count N.

$$\text{DR} = \frac{N_{\text{detected}}}{N} \tag{5}$$

The NA metric was used to evaluate how well the model can assign the correct tooth number. It is calculated using (6), defining the ratio of the correctly numbered tooth count N_{correct} to the total tooth count N. This metric is commonly used and can directly reveal the performance of any approach in the tooth numbering task.

$$\text{NA} = \frac{N_{\text{correct}}}{N} \tag{6}$$

3) METRICS FOR VARIABLE RELATION

This study used the correlation coefficient (CC) to evaluate the relationship between arbitrary two variables X and Y, defined by (7). This value ranges from -1 and 1. Generally,

values between 0.0-0.3, 0.3-0.5, 0.5-0.7, 0.7-0.9, and 0.9-1.0 are regarded as indicating negligible, low, moderate, high, and very high positive correlation, respectively.

$$\mathrm{CC} = \frac{\sum (X - \overline{X}) \times (Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \times \sum (Y - \overline{Y})^2}} \tag{7}$$

F. MISSING TEETH FEATURES CATEGORIZATION

As mentioned above, we introduced two features associated with missing teeth: dental implants and edentulous space (ES). Our investigation assessed the impact of these features from two perspectives:

- The type of a tooth (implant or natural tooth).
- The proximity of a tooth to ES.

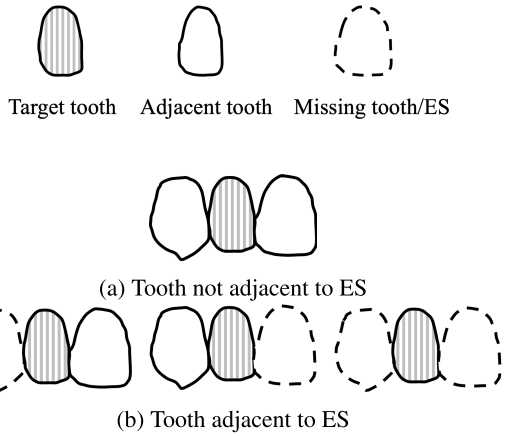


FIGURE 5. Patterns of tooth adjacent to ES and not adjacent to ES.

The proximity of a tooth is illustrated in Fig. 5. In this study, a tooth with neither side adjacent to ES was classified as not adjacent to ES.

Consequently, all teeth were categorized into four classes:

- Implant adjacent to ES (IES)
- Implant not adjacent to ES (InES)
- Natural tooth adjacent to ES (NES)
- Natural tooth not adjacent to ES (NnES)

These four categories were evaluated in this study.

III. RESULTS

We evaluated the tooth numbering performance using the 307 X-ray images from OUD through three object detection models (YOLOv8, YOLOv11, RT-DETR). Table 3 presents the number of X-rays, the number of instances (teeth, using Inst as abbreviation), and the average tooth count for each group (Avg). The metrics including precision (P), recall (R), F1-score (F1), detection rate (DR), and numbering accuracy (NA) were evaluated for each tooth-count group. We will analyze the results obtained from RT-DETR since it outperformed all other models we tried.

A. OVERVIEW RESULTS

The overall P of OUD was 90.0% with an R of 90.8% and F1 of 90.4%, respectively. Group 1, which contained the fewest teeth, exhibited the lowest P, R, and F1 at 62.2%, 55.3%, and 58.5%, respectively. In contrast, Group 8, containing the most teeth, achieved the highest performance results with a P of 98.3%, R of 97.9%, and F1 of 98.1%.

The overall DR achieved was 98.0% and each of groups demonstrated DRs surpassing 96.0%. Group 1 and 2 reached the highest DR value of 100.0%.

The NA across the entire dataset was 92.9%. Groups 1 and 2 exhibited a significantly lower NA of 49.0% and 73.3%, respectively, compared with the other tooth-count groups. The NA for each of Groups 3-8 exceeded 88.0%, and Groups 7 and 8 reached 95.0%.

The CC between tooth count and numbering accuracy for OUD was calculated at 0.59. This indicated a moderate positive correlation between tooth count and numbering accuracy, which suggests that tooth count moderately affects the numbering accuracy.

B. RESULTS OF FEATURES DERIVED FROM MISSING TEETH

On the basis of the overview results for each tooth-count group, all instances (teeth) were categorized into four types to further assess their DR and NA.

Table 4 shows the total count of instances, DR, and NA for the four categories of tooth features. Consequently, NnES obtained the highest DR of 98.3%, whereas IES and InES exhibited the lowest DR of 93.8%. Meanwhile, NnES reached the highest NA of 96.3%, whereas IES exhibited the lowest NA of 55.4%. NES achieved a low detection rate and numbering accuracy.

IV. DISCUSSION

We evaluated the performance difference among tooth-count groups in the tooth numbering task through three object detection models. We further calculated the detection rate and numbering accuracy of four tooth features to assess the influence of implants and ES. Furthermore, we highlighted the influence of distinct data biases derived from the practical panoramic X-ray datasets. Our findings may guide dentists in considering more specific factors when developing a generalized AI-based approach that is adaptable to various dental features.

The performance of each model, including precision (P), recall (R), and F1-score (F1), improved as the tooth count increased (corresponding to an increase in tooth-count group number) except for Group 3. The difference in these metrics between tooth-count groups suggested that datasets containing different proportions of images from different tooth-count groups could exhibit different results. The overall P, R, and F1 of our testing dataset (OUD) did not reach as high as those of related research (P: 99.90%, R: 99.18%), as listed in Table 1. However, the average tooth count of the testing dataset was 20.4, which was lower than all related research. This suggested that our dataset contained a larger proportion of low tooth-count groups (Group 1-4), contributing to our relatively low results.

TABLE 3. Results of YOLOv8, YOLOv11, RT-DETR on each tooth-count group.

Model				YOLOv8					YOLOv11					RT-DETR				
Group	X-rays	Inst.	Avg.	P	R	F1	DR	NA	P	R	F1	DR	NA	P	R	F1	DR	NA
1	18	49	2.7	65.3	71.2	68.1	87.8	51.0	51.2	78.0	61.8	87.8	46.9	62.2	55.3	58.5	100.0	49.0
2	12	86	7.2	65.9	86.0	74.6	100.0	68.6	58.8	82.1	68.5	98.8	80.2	66.9	73.0	69.8	100.0	73.3
3	26	270	10.4	87.7	82.0	84.8	95.2	85.2	75.8	82.2	78.9	95.6	82.2	80.8	87.3	83.9	97.8	88.5
4	20	287	14.4	80.9	81.9	81.4	97.6	84.3	80.3	80.4	80.3	95.8	83.6	79.2	85.3	82.1	97.9	88.1
5	51	952	18.7	80.4	84.2	82.2	95.1	87.6	81.6	81.8	81.7	95.3	85.8	84.7	82.6	83.6	96.2	88.9
6	68	1538	22.6	83.3	86.1	84.7	96.4	88.3	85.7	81.6	83.7	95.6	88.3	87.3	89.	88.1	97.0	92.2
7	81	2148	26.5	91.4	89.4	90.4	98.8	94.3	92.6	91.5	92.0	98.3	94.9	93.2	93.1	93.1	98.8	95.6
8	31	942	30.4	96.4	96.9	96.6	99.0	97.9	94.7	96.3	95.5	98.3	96.5	98.3	97.9	98.1	99.4	98.7
1-8	307	6272	20.4	89.0	87.4	88.2	97.4	90.8	87.3	87.8	87.5	96.9	90.5	**90.0**	**90.8**	**90.4**	**98.0**	**92.9**

TABLE 4. Results of different tooth features.

Tooth features	Instances	DR	NA
IES	112	93.8	55.4
InES	48	93.8	83.3
NES	1643	97.6	86.4
NnES	4469	98.3	**96.3**

When comparing the DRs among tooth-count groups, no significant disparities were observed. Regarding the NA, an upward trend was observed with the increase in tooth count as shown in Table 3. This upward trend indicated that datasets containing images from higher tooth-count groups yield better model performance, DR, and NA.

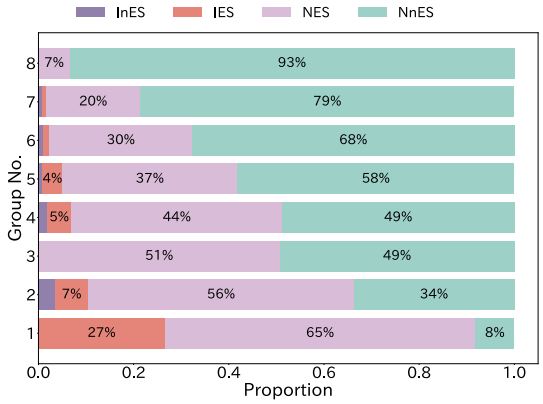
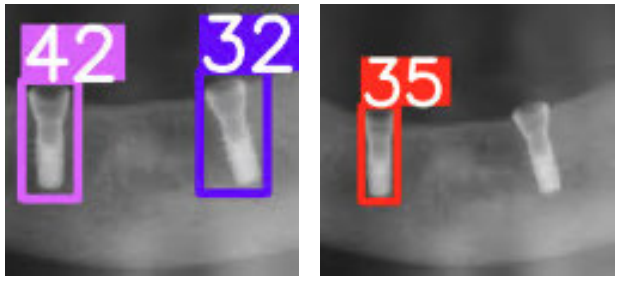


FIGURE 6. Proportion of four tooth features in each tooth-count groups.

The performance differences between each tooth-count group can be confirmed in Table 3, and we will explore the reasons for the variation in numbering accuracy. The proportion of the aforementioned four tooth features in each tooth-count group is shown in Fig. 6. Obviously, the low tooth-count groups (Group 1-4) occupied a large percentage of NES (Groups 1-4 contained 65.3%, 55.8%, 50.7%, and 44.3%, respectively) and tended to contain higher proportions of implants (IES and InES). According to the results of the four tooth features in Table 4, the proportion of each feature affected both the DR and the NA of each group. Due to the high NA of NnES, datasets with a large proportion of such teeth will obtain a higher numbering accuracy, making the influence of biases less apparent. Therefore, an upward trend in numbering accuracy could be explained because high tooth-count groups have a high proportion of NnES. Additionally, Group 3 did not contain IES, which exhibited the lowest numbering accuracy. The slightly higher NA of Group 3 (0.4% more than Group 4) can be attributed to the equal proportion of NnES in both groups and the higher proportion of implants in Group 4.

The tooth numbering results for implants and NES in terms of NA demonstrates that the deep learning-based models were generally less suitable for panoramic X-rays with numerous missing teeth. The impact of implants was primarily attributed to their distinctive shape compared with natural teeth. Fig. 7 illustrates examples of tooth numbering errors. Original tooth 42 was predicted as 35, and 32 was incorrectly detected.



(a) Ground truth (b) Numbering result

FIGURE 7. Examples of implant numbering error.

Additionally, datasets containing implants labeled with all FDI tooth numbers are difficult to obtain due to the complexity and high cost of conducting implant treatment. The lack of sufficient implant data increases the challenges of tooth numbering for implants. The influence of NES was likely due to the lack of adjacent tooth (teeth) information. Fig. 8 illustrates examples of bounding boxes that partially enclose the adjacent tooth (teeth). When annotating with boxes, the adjacent tooth (teeth) of a tooth would inevitably be enclosed due to the presence of inclined teeth or a curved oral structure. The adjacent tooth (teeth) part was learned by the models during the training process, which is expected to contribute to accurate tooth numbering.

Our study clarified that differences in tooth numbering performance among tooth-count groups exist. Also, we investigated the influence of implants and NES on tooth numbering performance. However, there were three limitations in our study. The first was that our datasets excluded panoramic X-rays containing bridges and deciduous teeth, which are pervasive features in panoramic X-rays. The reason for the exclusion is due to the limited number of data and the complexity of labeling the bridge treatment. Therefore, the effects of bridges and deciduous teeth on tooth numbering

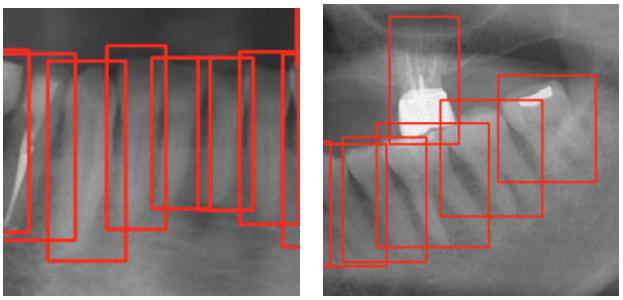
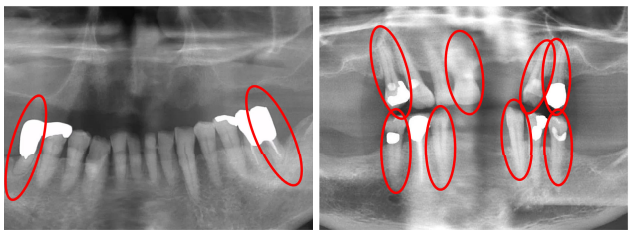


FIGURE 8. Examples of bounding box that enclosing adjacent tooth.

were not verified. The second limitation was that this study did not consider the differences in NES count of panoramic X-rays within the same tooth-count group since the image count of low tooth-count groups (Groups 1-4) was not sufficient for further subdivision. The NES count is dependent on the spatial distribution of teeth. An example is illustrated in Fig. 9. All teeth are adjacent in the lower half of Fig. 9a, whereas teeth are located in both the upper and the lower halves of Fig. 9b. A difference in NES count can eventually result in a divergence in numbering accuracy. The third limitation was that our experiment was conducted using a private dataset (OUD), which is not public and cannot be verified by other researchers. Although this situation is common in the dental field, further benchmarks for comprehensive verification in the field would be preferable.



(a) An image contains two NES (numbering accuracy: 88.9%). (b) An image contains eight NES (numbering accuracy: 54.5%).

FIGURE 9. An example of two images belonging to Group 3 with different NES count and numbering accuracy.

Considering the necessity of professional dental knowledge in labeling on panoramic X-ray images and the necessity of addressing privacy concerns, gathering datasets that include these features along with annotations is labor-intensive. Additionally, the NES number for the same tooth-count group varies depending on the distribution of tooth spatial positions. Obtaining datasets that contain all possible distribution patterns of teeth is a challenging task. Therefore, generative AI models [21] can be applied to synthesize labeled data containing various pervasive dental features, and enriching the variety of tooth position distribution patterns is future work.

V. CONCLUSION

This study assessed the performance of a deep learning-based approach for tooth numbering on panoramic X-rays within eight tooth-count groups. As our results indicated, although groups with higher tooth counts achieved better performance in the tooth numbering task, disparities in performance were still observed among different groups. Therefore, the distribution of the groups influences the performance of deep learning-based models for tooth numbering, which represents a significant bias in panoramic X-ray datasets. Therefore, disclosing the distribution of groups of panoramic X-ray datasets is essential for objectively evaluating the applicability of deep learning-based approaches to tooth numbering. Revealing the distribution along with the proportion of features containing implants and NES could be more beneficial for researchers in developing a generalized AI for panoramic X-ray image identification.

REFERENCES

[1] R. Takahashi, C. Muramatsu, T. Hara, T. Hayashi, T. Fukui, A. Katsumata, and H. Fujita, "Semiautomatic method for measuring alveolar bone resorption level with alveolar crest line detection by thin plate splines deformation and gradient analysis on dental panoramic radiographs," *IEICE Tech. Rep., IEICE Tech. Rep.*, vol. 115, no. 401, pp. 327–330, Jan. 2016.

[2] C. Kuwada, Y. Ariji, M. Fukuda, Y. Kise, H. Fujita, A. Katsumata, and E. Ariji, "Deep learning systems for detecting and classifying the presence of impacted supernumerary teeth in the maxillary incisor region on panoramic radiographs," *Oral Surg., Oral Med., Oral Pathol. Oral Radiol.*, vol. 130, no. 4, pp. 464–469, Oct. 2020.

[3] H. Yang, E. Jo, H. J. Kim, I.-H. Cha, Y.-S. Jung, W. Nam, J.-Y. Kim, J.-K. Kim, Y. H. Kim, T. G. Oh, S.-S. Han, H. Kim, and D. Kim, "Deep learning for automated detection of cyst and tumors of the jaw in panoramic radiographs," *J. Clin. Med.*, vol. 9, no. 6, p. 1839, Jun. 2020.

[4] H. Crow, E. Parks, J. Campbell, D. Stucki, and J. Daggy, "The utility of panoramic radiography in temporomandibular joint assessment," *Dentomaxillofacial Radiol.*, vol. 34, no. 2, pp. 91–95, Mar. 2005.

[5] L. Maestre-Ferrín, S. Galan-Gil, C. Carrillo-García, and M. Peñarrocha-Diago, "Radiographic findings in the maxillary sinus: Comparison of panoramic radiography with computed tomography," *Int. J. Oral Maxillofacial Implants*, vol. 26, no. 2, pp. 341–346, Apr. 2011.

[6] D. V. Tuzoff, L. N. Tuzova, M. M. Bornstein, A. S. Krasnov, M. A. Kharchenko, S. I. Nikolenko, M. M. Sveshnikov, and G. B. Bednenko, "Tooth detection and numbering in panoramic radiographs using convolutional neural networks," *Dentomaxillofacial Radiol.*, vol. 48, no. 4, May 2019, Art. no. 20180051.

[7] E. Bilgir, İ. Ş. Bayrakdar, Ö. Çelik, K. Orhan, F. Akkoca, H. Sağlam, A. Odabas, A. F. Aslan, C. Ozcetin, M. Killi, and I. Rozylo-Kalinowska, "An artificial intelligence approach to automatic tooth detection and numbering in panoramic radiographs," *BMC Med. Imag.*, vol. 21, pp. 1–9, Dec. 2021.

[8] B. O. Gurses, E. Sener, and P. Guneri, "Deep learning for automated detection and numbering of permanent teeth on panoramic images," *Dentomaxillofacial Radiol.*, vol. 51, no. 5, Jul. 2022, Art. no. 20210296.

[9] A. Karaoglu, C. Ozcan, A. Pekince, and Y. Yasa, "Numbering teeth in panoramic images: A novel method based on deep learning and heuristic algorithm," *Eng. Sci. Technol., Int. J.*, vol. 37, Jan. 2023, Art. no. 101316.

[10] S. Yilmaz, M. Tasyurek, M. Amuk, M. Celik, and E. M. Canger, "Developing deep learning methods for classification of teeth in dental panoramic radiography," *Oral Surg., Oral Med., Oral Pathol. Oral Radiol.*, vol. 138, no. 1, pp. 118–127, Jul. 2024.

[11] M. A. Ali, D. Fujita, and S. Kobashi, "Teeth and prostheses detection in dental panoramic X-rays using CNN-based object detector and a priori knowledge-based algorithm," *Sci. Rep.*, vol. 13, no. 1, p. 16542, Oct. 2023.

[12] R. H. Putra, E. R. Astuti, D. K. Putri, M. Widiasri, P. A. M. Laksanti, H. Majidah, and N. Yoda, "Automated permanent tooth detection and numbering on panoramic radiograph using a deep learning approach," *Oral Surg., Oral Med., Oral Pathol. Oral Radiol.*, vol. 137, no. 5, pp. 537–544, May 2024.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[16] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.
[17] S. Keiser-Nielsen, "Fédération dentaire internationale two-digit system of designating teeth," *Int. Dent. J.*, vol. 21, pp. 104–106, Jan. 1971.
[18] K. Panetta, R. Rajendran, A. Ramesh, S. P. Rao, and S. Agaian, "Tufts dental database: A multimodal panoramic X-ray dataset for benchmarking diagnostic systems," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 4, pp. 1650–1659, Apr. 2022.
[19] S. Gülüm, S. Kutal, K. C. Aydin, G. Akgün, and A. Akdağ, "Effect of data size on tooth numbering performance via artificial intelligence using panoramic radiographs," *Oral Radiol.*, vol. 39, no. 4, pp. 715–721, Oct. 2023.
[20] Y. Zhao et al., "Detrs beat yolos on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16965–16974.
[21] K. Kokomoto, R. Okawa, K. Nakano, and K. Nozaki, "Tooth development prediction using a generative machine learning approach," *IEEE Access*, vol. 12, pp. 87645–87652, 2024.

ZHONGBO TANG (Member, IEEE) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2015, and the M.E. degree from Osaka University, Osaka, Japan, in 2024. He is currently pursuing the Ph.D. degree with the Graduate School of Information Science and Technology. His research field is computer science. His research interests include image processing and artificial intelligence. He is a member of IPSJ.

TOMOAKI MAMENO received the Ph.D. degree from the Department of Prosthodontics, Gerodontology, and Oral Rehabilitation, Graduate School of Dentistry, Osaka University, in 2019. He has been an Assistant Professor with the Department of Prosthodontics, Gerodontology, and Oral Rehabilitation, Graduate School of Dentistry, Osaka University, since 2019. His research focuses on the impact of oral health on overall systemic health in older adults.

TOMONORI HAYAMI received the Ph.D. degree from Osaka University, in 2023. He is currently a specially-appointed Assistant Professor with the D3 Center, Osaka University. His current research interests include artificial intelligence, bioinformatics, and high-performance computing. He is a member of IPSJ.

CHONHO LEE received the Bachelor of Computer Science degree from the University of California, Irvine, and the Ph.D. degree in computer science from the University of Massachusetts, Boston. He is currently a Professor with the Department of Information and Computer Engineering, Okayama University of Science. His research interests include optimization and self-adaptation for large-scale problems using machine learning and bio-inspired approaches.

SUSUMU DATE (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Osaka University, in 1997, 2000, and 2002, respectively. He was an Assistant Professor with the Graduate School of Information Science and Technology, Osaka University, from 2002 to 2005. He was a Visiting Scholar with the University of California at San Diego, in 2005. From 2005 to 2008, he was a specially-appointed Associate Professor of the internationalization of education with the Graduate School of Information Science and Technology, Osaka University. Since 2008, he has been an Associate Professor with the Cybermedia Center, Osaka University. His research field is computer science. His current research interests include cloud, cluster, grid, and high-performance computing and their applications. He is a member of IPSJ.

KAZUNORI IKEBE received the Ph.D. degree from the Graduate School of Dentistry, Osaka University, in 1991. He has been a Professor and the Chair of the Department of Prosthodontics, Gerodontology, and Oral Rehabilitation, Graduate School of Dentistry, Osaka University, since 2018. He is currently an experienced Dental Researcher in geriatric dentistry and prosthodontics. He was honored with the International Association for Dental Research, a Distinguished Scientist Award, in 2015.

KAZUNORI NOZAKI received the degree from the Faculty of Dental Medicine, Hokkaido University, in March 2001, the Ph.D. degree in dentistry from Osaka University, in 2004, and the Ph.D. degree in information science, in 2009. He was a Visiting Professor with GIPSA-Laboratory, Joseph Fourier University, in 2011, and joined Osaka University as a specially-appointed Lecturer, in 2011. In 2013, he became an Assistant Professor with Osaka University Dental Hospital, an Associate Professor, in 2019, and the Head of Oral Dental Informatics, in 2024. He is currently the Chair of the Dental Informatics Study Group, Japan. He specializes in integrating medical informatics and computer science. He is an Expert Member of ISO committees. In 2024, he joined the Editorial Board of *Journal of Dental Research*.