Oral Radiology
https://doi.org/10.1007/s11282-023-00689-4

ORIGINAL ARTICLE

Effect of data size on tooth numbering performance via artificial intelligence using panoramic radiographs

Semih Gülüm[1] · Seçilay Kutal[2] · Kader Cesur Aydin[3] · Gazi Akgün[4] · Aleyna Akdağ[5]

Received: 15 December 2022 / Accepted: 29 May 2023

Abstract

Objective This study aims to investigate the effect of number of data on model performance, for the detection of tooth numbering problem on dental panoramic radiographs, with the help of image processing and deep learning algorithms.

Study Design The data set consists of 3000 anonymous dental panoramic X-rays of adult individuals. Panoramic X-rays were labeled on the basis of 32 classes in line with the FDI tooth numbering system. In order to examine the relationship between the number of data used in image processing algorithms and model performance, four different datasets which include 1000, 1500, 2000 and 2500 panoramic X-rays, were used. The training of the models was carried out with the YOLOv4 algorithm and trained models were tested on a fixed test dataset with 500 data and compared based on F1 score, mAP, sensitivity, precision and recall metrics.

Results The performance of the model increased as the number of data used during the training of the model increased. Therefore, the last model trained with 2500 data showed the highest success among all the trained models.

Conclusion Dataset size is important for dental enumeration, and large samples should be considered as more reliable.

Keywords Artificial intelligence · Image processing · Health technologies · Panoramic x-ray · Tooth numbering

Introduction

Artificial intelligence (AI), a branch of computer science that can analyze complex medical data [1], is expanding its medical applications day by day with the latest developments in digitized data collection, machine learning and computing infrastructure. In this way, it became involved in many medical fields that were previously considered only the domain of human specialists [2]. In addition, the increase in the number and complexity of data in the medical field means that artificial intelligence (AI) will be applied more actively in the coming days [3]. Popular AI techniques include the classical approach of support vector machine (SVM) and machine learning (ML) methods for structured data such as neural networks. Oncology, neurology and cardiology are among the main areas treated using AI [4]. However, at this point, supervised learning is very costly in the medical field, because it is difficult to obtain field-specific information and ground truth labels [5]. Therefore, as the

Semih Gülüm and Seçilay Kutal have contributed equally to this work and share first authorship.

Kader Cesur Aydin and Gazi Akgün have contributed equally to this work and share senior authorship.

✉ Kader Cesur Aydin
kadercesur@yahoo.com

Semih Gülüm
semih.gulum@avl.com

Seçilay Kutal
secilay.kutal@areal.ai

Gazi Akgün
gazi.akgun@marmara.edu

Aleyna Akdağ
aleyna.akdag@std.medipol.edu.tr

[1] AVL Research and Engineering, Abdurrahmangazi Mah. Ataturk Cad. No: 22 11/22 Kat: 6 Sultanbeyli, Istanbul 34920, Turkey

[2] AREAL.ai, San Francisco, USA

[3] School of Dentistry, Head of Department Dentomaxillofacial Radiology, Istanbul Medipol University, Ataturk Bulvarı No: 27 Unkapanı, Istanbul 34083, Turkey

[4] Technology Faculty RTE Campus, Marmara University, T1/203, Maltepe, Istanbul 34854, Turkey

[5] School of Dentistry, Istanbul Medipol University, Goztepe Mah, Ataturk Cad, Beykoz, Istanbul 34810, Turkey

Published online: 05 July 2023

medical field to be studied (e.g. dental health, the relevant field in this study) becomes specialized, the effort and cost increase.

As the medical field becomes more specialized, which is desired to be solved with deep learning methods, the specified processes become more and more difficult. This also applies to oral diseases, which are among the most common diseases in humans. Since only the crowns of the erupted teeth can be visualized in the mouth and the roots cannot be detected by inspection, it is difficult for dentists to manually diagnose dental anomalies and diseases. Therefore, dental x-ray imaging methods are the most popular and most preferred auxiliary examination method for diagnosing dental anomalies and diseases before dental treatments [6]. Evaluation and interpretation of x-rays and making the correct diagnosis is one of the most important processes in the initial phase of treatment. Automatic tooth recognition applications have started to be developed by using panoramic x-ray images, which is the most frequently used dental imaging method today.

Jaideep Sur et al. conducted a survey among dentists in India about artificial intelligence and its possible contributions to radiological diagnosis [7]. On contrary to the popular classical belief, their study proves that the use of artificial intelligence in the field of health is supported by physicians. When the results of the study are examined, it is seen that dentists mostly gave positive answers about the preference and widespread use of artificial intelligence in radiological detection. At the same time, it is predicted that the use of artificial intelligence in this field may be useful for dentists, especially in examining complex X-rays.

Newly released study by Gunec et al. stated that 83.7% of the laypeople think that "artificial intelligence applications can be effective in dental diagnosis" and 93.8% confirm that "dentist and artificial intelligence can work together" [8]. This survey reveals that artificial intelligence-based dental health applications should be used in dentistry practice and therefore they can play role in the benefit of the society in both preventive medicine and dental health tourism.

In this study, the relationship between the problem of tooth detection and numbering, which is one of the main topics of dentistry, and the number of data and the success rate, which is one of the main topics of deep learning problems, are discussed. The number of data was considered important in this study because the number of trained classes were high (32 different classes for all teeth) and different from studies using few classifications, we assume that the success rate of the trainings that have higher number of images during testing will be elevated. In the problem that was tried to be solved with the You Only Look Once v4 (YOLOv4) algorithm -which is an object detection algorithm with high accuracy and speed- it was aimed to measure the role of the number of data in the success of the model, and in this direction, help from specialist dentists was taken. The development of tooth number detection may be used by dentists, dental students and even by laypeople, for initial acquiring of the relevant tooth and related dental pathologies may be addressed easier.

Methods and materials

Dataset

To create the dataset, only adult panoramic X-rays obtained through the databases of various dental clinics were included, both genders were selected, and care was taken to include X-rays taken at different time intervals. All X-rays were anonymized to protect patient confidentiality before the data was labeled and made meaningful for the model. In order to avoid performance loss that may occur due to data differences during training, all data were analyzed and separated according to image quality, color difference, etc. After the analysis of the data, 3000 panoramic X-rays and 58,575 labeled teeth became usable in the study.

Oral physiology was taken into consideration and therefore the maximum number of adult teeth classes were determined as 32, for the classification problem, although there may be fewer numbers of teeth. Support was received from dentomaxillofacial radiology specialist dentists during the labeling process carried out in line with the preferred FDI (Federation Dentaire Internationale) numbering system [9] for class names. Labeling was carried out with the LabelImg graphic labeling tool [10] and the resulting label files were brought into a suitable format so that they can be given as input to the model.

During the labeling process, large variations such as fillings, crowns and implants that disrupt the morphological structure of the tooth were not labeled in order to learn the tooth numbers by the model. Tooth samples with the mentioned variations are given in Fig. 1.

The data distribution according to the tooth numbers in the data set, which has an average of 19.5 teeth labeled in each x-ray, is shown in Fig. 2. Considering this distribution, it can be concluded that teeth numbered 18, 28, 38 and 48 are less common in an adult human mouth compared to other teeth. In addition, the distribution of teeth belonging to other classes on the data is uneven. However, the fact that the data set was collected from different time intervals, clinics, genders and patients proves that the data obtained has a high power to represent the real world. Therefore, no augmentation was performed for the teeth in the data set and care was taken to preserve the observed distribution.

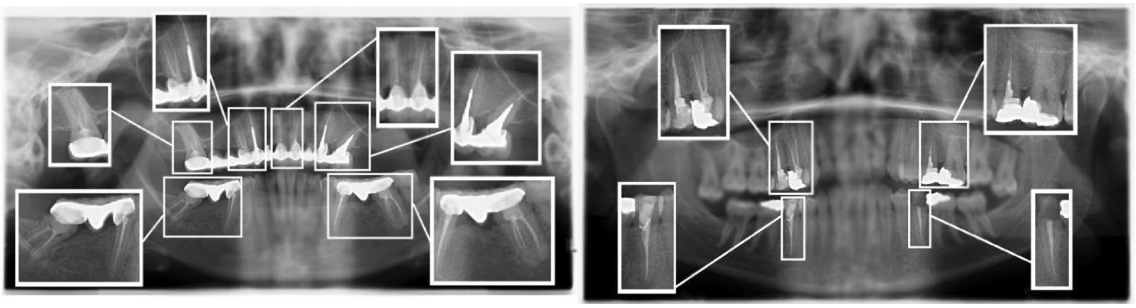

Fig. 1 Unlabeled teeth with variations

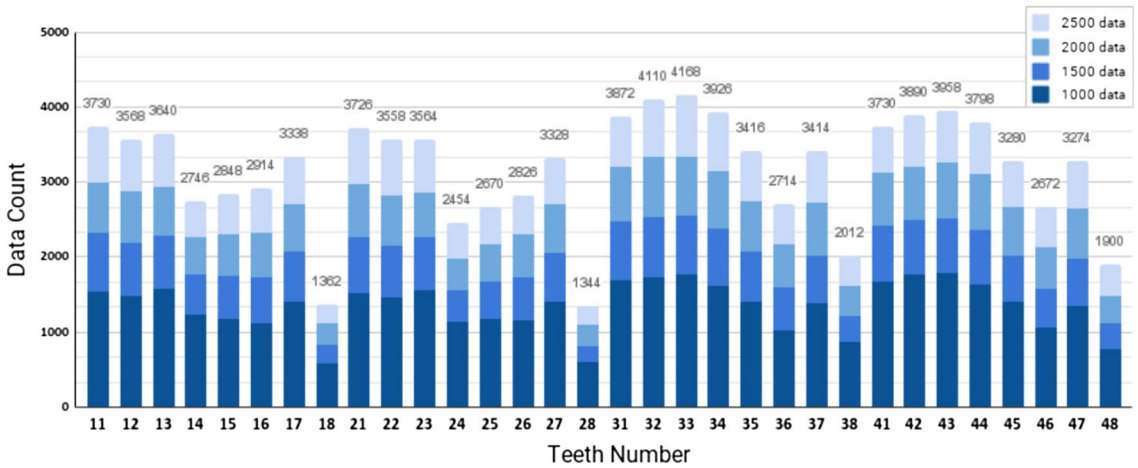

Fig. 2 Data distribution according to the data sets used in the models

YOLOv4 method

The deep learning model chosen for training in the study is YOLOv4 [11], which has a CNN-based architecture developed by Bochkovskiy et al. and is a real-time object recognition algorithm that can present multiple objects in a single frame. YOLOv4, which has 162 layers in total, contains more than 64 million parameters. This architecture, which can detect more than one object on an image, draws bounding boxes around each object to indicate the area of the object it predicts. The model needs an image-specific label file containing the coordinate information and class names of the labeled objects, as well as the image on which the labeling is made, as input during the training phase.

By dividing the data set consisting of 3000 X-rays in total, four different model trainings were carried out with 1000, 1500, 2000 and 2500 data numbers to be used in training. The remaining 500 X-rays were separated from the training dataset and treated as a fixed test dataset to measure the performance of each model. The four models trained in 12,000 iterations were tested by recording their weights every 3000 iterations. Model parameters determined for the trainings carried out with the GPU support provided on Google Colab Notebook are given in Table 1.

Table 1 Model parameters

Image size (width × height)	Batch size	Subdivision	Learning rate
608 × 608	32	16	0.001
Confidence	Momentum threshold	Decay	Iteration count
0.3	0.949	0.0005	12,000

Results

Precision, recall, mean average precision (mAP), sensitivity and F1 scores, which are frequently preferred metrics, were taken into account during the testing of deep learning models and comparison of their performance. The importance of the metrics varies according to various studies. The reason why the F1 score is preferred is that it can provide

a class-specific performance evaluation in classification problems.

The performance graphs observed as a result of the tests performed at the end of each epoch on the test data set separated during the training of the model are given in Fig. 3. The highest performance values obtained on the test data sets of each trained model are given in Table 2.

In the four different trainings conducted, a directly proportional correlation was found between the number of data and performance metrics, and it was observed that the model performance increased as the number of data used during the training increased. However, in cases where the number of data is insufficient, underfitting is observed since it cannot learn enough to explain the data. By increasing the number of the dataset, the tendency of underfitting was overcome. In this case, if the model was trained a lot, the model showed a tendency to overfit by memorizing the data. Overfitting was observed after 9 epochs in 2500 data training with the highest training data, as expected, showed itself in earlier epochs in models with less data numbers. In addition, as expected, underfitting was observed in 1000 data training with the lowest training data.

The outputs obtained as a result of the test performed with the most successful epochs of each model are shown in Fig. 4. When this figure is examined, it is observed that as a result of the model trainings carried out with 1000 and 1500 data, it is not possible to distinguish between the presence of the tooth in neighbouring quadrants at the same horizontal axis. While this problem was overcome in the 2000 data training, it is observed that the problems of estimating more than one class on a single tooth that emerged and especially in the determination of the wisdom teeth were overcome as a result of the training made with 2500 data. The study is mainly based on re-testing data and cut-off for the study was set at 3000 radiographs (2500 labelling and 500 for testing) because the testing for this data size revealed overfitting. The model learns about the noise in the training data because its complexity is high. In overfitting problems due to high education classes, increasing data size may reveal limited or no significance for success.

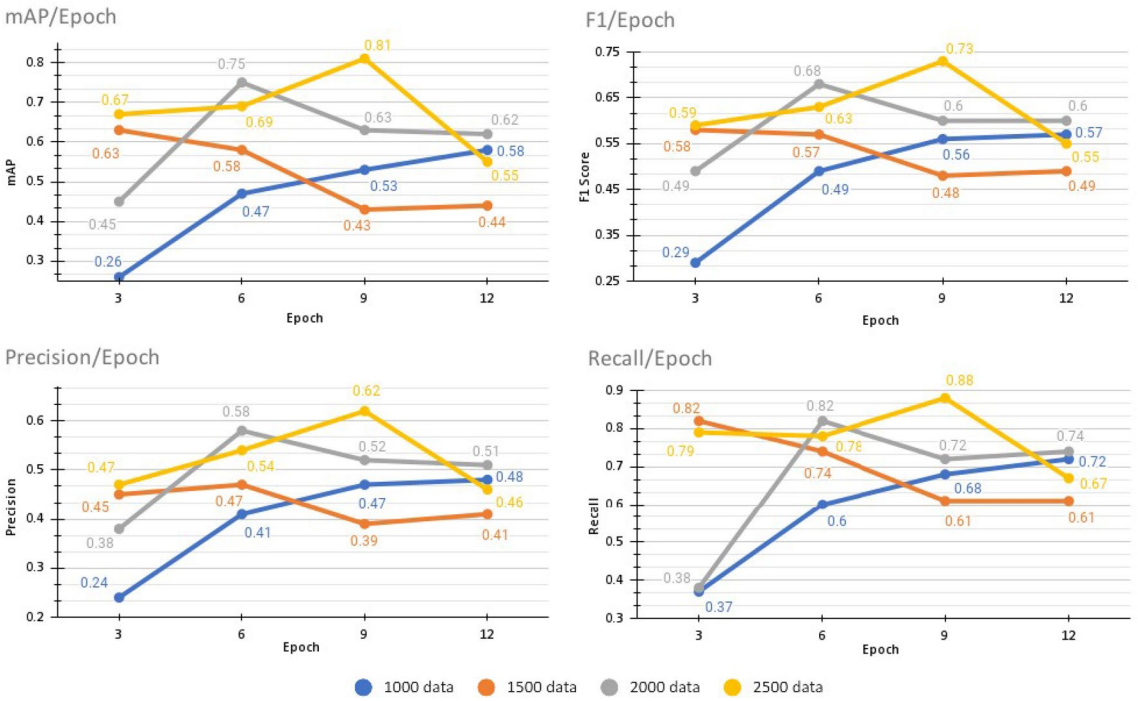

Fig. 3 Test accuracy graphics of models

Table 2 Model-specific highest test performance metrics and related epoch information

	Precision	Recall	F1-score	Sensitivity	mAP	Best epoch
1000 data	0.48	0.72	0.57	0.72	0.58	12
1500 data	0.45	0.82	0.58	0.82	0.63	3
2000 data	0.58	0.68	0.68	0.83	0.75	6
2500 data	0.62	0.73	0.83	0.88	0.81	9

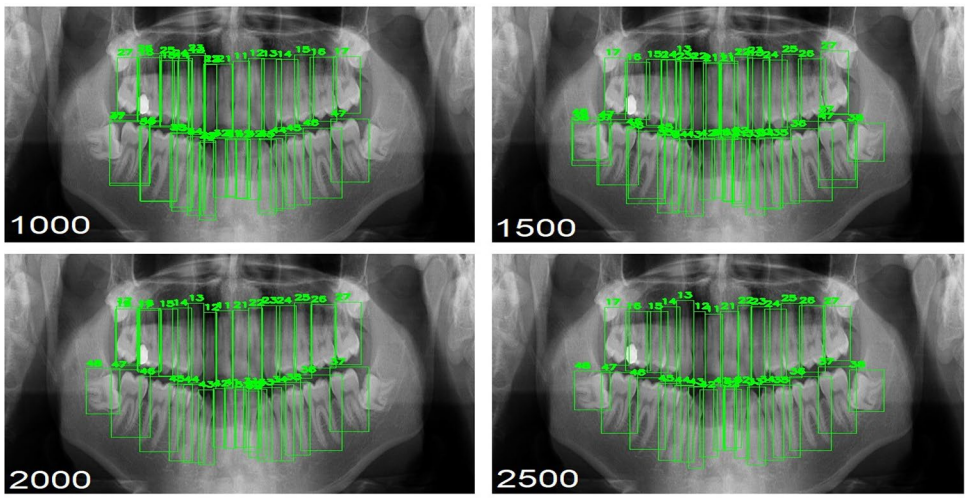

Fig. 4 Predicted test samples

Among the models trained in line with the findings, the highest performance model was the model in which 9 epoch trainings were performed with 2500 data. In order to examine the effect of the intersection over union (IoU) parameter on the model results, this model was also tested with different IoU values. As the IoU value falls below 0.5, an increase is observed in mAP values, while a decrease is observed when it is above 0.5.

Discussion

Tuzoff et al. used Faster Region Based Convolutional Neural Networks (R-CNN) for the detection of the teeth in the method they applied, and the detected teeth were cropped and then given to the Visual Geometry Group-16 (VGG-16) algorithm for numbering [12]. Then, the X-ray was post-processed in order for the X-ray to be interpretable and the cropped teeth were reassembled.

Mahdi et al. preferred Residual Neural Network-50 (ResNet-50) as a backbone in the Faster R-CNN architecture they used to solve the tooth numbering problem [13]. During the data labeling of their study with 900 X-Rays, teeth with large restorations, such as teeth with large-area fillings, were not labeled. As a result of the study, a value of 0.942 mAP was reached.

Muramatsu et al. in their study, the problems of determining the teeth, discussed the problems of identifying teeth, determining the types of teeth, and classifying teeth according to their restorations [14]. In the study, which was carried out with 1000 X-ray images, it was aimed to minimize the performance problem that may arise with the fourfold cross validation method due to the insufficient number of data. With the Convolutional Neural Networkbased (CNN-based) GoogleNet architecture used for the detection of teeth, a sensitivity value of 96.4% has been reached. With the ResNet architecture used for classification problems, accuracy values of 93.2% and 98% were achieved, respectively, in the classification of teeth according to their types and restorations.

In their study, Kim et al., in addition to detecting teeth and implants, also worked on the numbering of detected teeth [15]. In the study, R-CNN model was preferred and a total of 303 X-rays, 253 trains and 50 test data, were used. It has reached an accuracy of 77.4% in tooth numbering. The biggest problem in numbering was that when there were dental interventions such as fillings and crowns on the tooth, the shape of the tooth could not be preserved because it could not be identified.

Muresan et al. used the Semantic Segmentation method to detect fourteen different dental diseases [16]. Labeling has been done for fourteen different diseases on grayscale X-rays, and also for the background. Since the background class is related to pixel values, it can take values between 0 and 255. In the study, the Efficient Residual Factorized Convolutional Network (ERFNet) model, which is the first 16-layer encoder and the last 7-layer decoder, was preferred. Among the methods used, the highest F1 score was 0.93.

Cho et al. performed tests using the Hospital Picture Archiving and Communication System (PACS) Dataset with the help of Convolutional Neural Network (CNN) algorithm in order to observe the effect of the data set size on solving medical classification problems. As a result of the tests, it was seen that there was a directly proportional correlation between the data set size and the success rates [17].

Yüksel et al. performed enumeration tests on panoramic radiograps, initially by segmenting to 4 quadrants, then by numerating the 8 teeth on all quadrants, with the resultant classification of 32 FDI notations [18]. By decreasing the number of classes in each quadrant, object detection problem was overcome. They have evaluated average precision scores for success evaluation and it resulted as 89.4% with an overall of 600 labeled images. Average precision score in our 2500 data testing resulted as 62%, regarding all 32 classes were

determined on the same images during testing. The difference in the average precision between the two studies is considered about difference of the number of classes (8 vs. 32). On the other hand, different evaluation parameters can reveal high results, as we have performed higher results for mAP (0.81) and recall values (0.73) than Yüksel et.al.

In this study, the effect of data number on model success on tooth numbering problem with YOLOv4 architecture, which is one of the most state of-the-art object detection algorithms, was investigated in tooth detection and numbering tasks. The most successful model obtained as a result of the trainings carried out was the model with the highest number of data with an F1 score of 0.83. With this model, which was trained, high-performance results were obtained by numbering the teeth at the radiologist level. The mAp value outcomes are lower than Mahdi et al. [13], on the other hand this finding may be in relation with the test size, and if the labelled images were used for testing as well, this may have caused a false positive outcome. Also the classification achievement by Muramatsu et al. was higher [14], with a tooth detection sensitivity of 96.4%, detailed analysis of the data has shown that their classification was based on incisor, canine, premolar and molar labelling (4 classes) and 32 classes of the FDI notation was not included. Regarding the great number of classes trained, our results with 88% sensitivity reveal a great role.

At the stage of measuring the performance of the models, each X-ray was analyzed independently by the trained model and a radiologist who is an expert in the field. In the analyzed X-rays, it was seen that most of the predictions that the model interpreted as incorrect were not wrong, but teeth that were largely deformed in their structure and were not labeled were found to be correct by the model and increased the False Positive (FP) value.

There are limitations to this study due to labeling exclusions of largely variated tooth shapes like fillings, crowns and implants. In future studies, teeth with large restorations can be numbered and included in the model training, thus increasing the success of the model. As a different solution method, in addition to the detection of teeth with object detection algorithms in the prediction phase of the model, location-based information can also be included. Thus, the tooth number prediction that can be made by looking at the tooth shape can be strengthened according to the latitudinal region of the tooth and the model performance can be increased.

Conclusion

In our study, the problem of detecting tooth numbering in panoramic X-rays with artificial intelligence and image processing algorithms is emphasized. In addition, the importance of the size of the data set in object detection algorithms for medical images was tried to be emphasized. In case the number of data is low, usage of various altering augmentation solutions can be used to increase the number of data artificially and the level of performance obtained can be increased.

With the spread of algorithms based on artificial intelligence and deep learning and their introduction into daily life, an assistant can be developed that facilitates dentists to make faster and more reliable clinical decisions for clinicians and dental students.

Acknowledgements The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions KCA: conceptualization, methodology, ınvestigation, writing—original draft preparation, writing—reviewing and editing, visualization. SK: methodology, ınvestigation, software, validation, data curation, ınvestigation, writing—original draft preparation. SG: methodology, ınvestigation, software, validation, data curation, ınvestigation, writing—original draft preparation, resources. GA: methodology, project administration, supervision. AA: software, validation, resources.

Funding Not applicable

Data availibility The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Conflicts of interest All authors declare that they have no conflict of interest.

Ethics approval Not applicable. This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed consent Not applicable

References

1. Ramesh AN, Kambhampati C, Monson JRT, Drew PJ. Artificial intelligence in medicine. Ann R Coll Surg Engl. 2004;86(5):334–8.
2. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2(10):719–31.
3. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94–8.
4. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43.
5. Holzinger A, Langs G, Denk H, Zatloukal K, Muller H. Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev Data Min Knowl Discov. 2019;9(4):e1312.
6. Whyte A, Matias MATJ. Imaging of orofacial pain. J Oral Pathol Med. 2020;49(6):490–8.
7. Sur J, Bose S, Khan F, Dewangan D, Sawriya E, Roul A. Knowledge, attitudes, and perceptions regarding the future of artificial intelligence in oral radiology in india: a survey. Imaging Sci Dent. 2020;50(3):193.

8. Günec HG, Gökyay SS, Kaya E, Cesur-Aydın K. Toplum Yapay Zeka Ile Dental Tani Konmasina Hazir Mı? Selcuk Dental Journal. 2022;9:200–7. https://doi.org/10.15311/selcukdentj.915522.
9. Keiser-Nielsen S. Fédération Dentaire Internationale two-digit system of designating teeth. Int Dent J. 1971;21:104–6.
10. Tzutalin, Labelimg, Gitcode https://github.com/tzutalin/labelImg, [accessed 20 Oct 2021] (2015)
11. A. Bochkovskiy, C. Wang, H. M. Liao, Yolov4: Optimal speed and accuracy of object detection, CoRR abs/2004.10934 (2020). arXiv:2004.10934. Accessed on 23 Apr 2020
12. Tuzoff DV, Tuzova LN, Bornstein MM, Krasnov AS, Kharchenko MA, Nikolenko SI, Sveshnikov MM, Bednenko GB. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. Dentomaxillofacial Radiology. 2019;48(4):20180051.
13. Mahdi FP, Yagi N, Kobashi S. Automatic teeth recognition in dental x-ray images using transfer learning based faster r-cnn, in, IEEE 50th International Symposium on Multiple-Valued Logic (ISMVL). IEEE. 2020;2020:16–21.
14. Muramatsu C, Morishita T, Takahashi R, Hayashi T, Nishiyama W, Ariji Y, Zhou X, Hara T, Katsumata A, Ariji E, et al. Tooth detection and classification on panoramic radiographs for automatic dental chart filing: improved classification by multi-sized input data. Oral Radiol. 2021;37(1):13–9.
15. Kim C, Kim D, Jeong H, Yoon S-J, Youm S. Automatic tooth detection and numbering using a combination of a cnn and heuristic algorithm. Appl Sci. 2020;10(16):5624.
16. Muresan MP, Barbura AR, Nedevschi S (2020) Teeth detection and dental problem classification in panoramic x-ray images using deep learning and image processing techniques. In: 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, 2020, pp. 457–463.
17. Cho J, Lee K, Shin E, Choy G, Do S (2015) How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?, arXiv preprint arXiv:1511.06348
18. Yüksel AE, Gültekin S, Simsar E, Özdemir ŞD, Gündoğar M, Tokgöz SB, Hamamcı İE. Dental enumeration and multiple treatment detection on panoramic X-rays using deep learning. Sci Rep. 2021;11(1):1–10.