www.cad-journal.net
Computer-Aided Design
AND APPLICATIONS

# Multi-Label Dental Image Classification via Vision Transformer for Orthopantomography X-ray Images

Yilin Li[1] and Jun Zhang[2]

[1]Department of Orthodontics, School, and Hospital of Stomatology, Cheeloo College of Medicine,
Jinan, China, leeyling@126.com
[2]Department of Orthodontics, School, and Hospital of Stomatology, Cheeloo College of Medicine,
Jinan, China, zhangj@sdu.edu.cn

Corresponding author: Jun Zhang, zhangj@sdu.edu.cn

**Abstract.** Dental disease is extensively taken as one of the most important healthcare issues globally. Early detection of dental diseases using imagery-based data can significantly improve clinical diagnosis and treatment, hence decreasing the risk of serious health ailments. To address the task of dental image classification, a large number of deep learning models, especially convolutional neural networks, have been presented and achieved promising outcomes in a variety of benchmarks, including dealing with dental X-ray images in clinical practices. Nevertheless, the aforementioned models lack acceptable adaptability and stability when applied to practical scenarios. A major drawback of convolutional neural networks is their limited ability to capture the global relationships between distant pixels in medical pictures due to their small receptive field. This paper presents a new vision transformer model to address this disparity, specifically focusing on the task of dental picture classification. The proposed method utilizes the multi-head self-attention module while excluding the convolution operators. Furthermore, the process of transfer learning is utilized to optimize the weighting parameters of the visual transformer being presented. The experimental results provide evidence that the suggested method outperforms the current cutting-edge deep learning techniques.

**Keywords:** Medical image analysis; Dental X-ray Image; Vision Transformer.
**DOI:** https://doi.org/10.14733/cadaps.2024.S21.198-207

## 1 INTRODUCTION

Dental disease is an important public health issue that has a high incidence rate globally. Dental imaging or dental radiography is taken as a useful instrument for medical examination and treatment employed for the improvement of oral health. In general, a set of dental imaging techniques have been leveraged in clinical practices, including X-ray, cone-beam computed tomography (CBCT), and optical coherence tomography (OCT). These modalities have been widely exploited in object detection, image classification, and segmentation for the diagnosis and treatment of various dental diseases.  However, the analysis and interpretation of X-ray images in a manual fashion might be time-consuming, labor-tedious, and error-prone. In addition, it may also incur misdiagnosis since the dentists may experience fatigue, tension, or lack of experience [1]. These problems can be resolved after automated X-ray image analysis tools are presented to assist dentists. Accordingly, automatic teeth recognition, dental anomaly detection, and dental disease classification can be addressed intelligently.

Different machine learning methods (ML) have been utilized in dental image classification, including Bayesian algorithm [3], linear algorithm [1], and support vector machine (SVM) [16]. Meanwhile, at least the following hand-crafted features have also been employed in dental image classification as the input of the above-mentioned ML methods, including Fourier descriptors and various contours [12]. To note that elaborate engineering of feature extraction is difficult to realize and the performance of the ML methods rely on the extracted features. Therefore, these ML methods have not provided promising outcomes. While some ML methods did achieve excellent results, the leveraged dataset usually is insufficient in quantity. It could be attributed to the limitations of ML model capabilities. Recently, deep learning (DL) based models have gradually been applied in dental imaging analysis mainly using convolutional neural network (CNN) [9]. DL has also been exploited to identify dental lesions in dental images [3,6,8,10,18]. As shown by the experimental results, the DL methods can be a better diagnosis instrument for dental images. However, the CNN-based models suffer from the limitations of local receptive fields, which might neglect the global relationship between long-range pixels.

Bearing the above-mentioned analysis in mind, this study investigates the employment of vision transformers in dental image analysis. To be specific, a vision transformer is proposed to implement the image classification for Orthopantomography X-ray OPG Images. In the presented vision transformer, the multi-head self-attention mechanism is also leveraged. Experimental results on the manually collected dental image dataset demonstrate the superior performance of the proposed approach.

Considering the analysis given above, this work explores the use of a vision transformer in dental image processing. More precisely, a vision transformer is suggested as a means to carry out image categorization for Orthopantomography X-ray OPG Images. The multi-head self-attention method is utilized in the given vision transformer. The proposed approach exhibits superior performance, as evidenced by the experimental findings obtained from the manually acquired dental image collection.

The primary contributions of this work are as follows:

- This appears to be one of the first attempts at classifying OPG images using a vision transformer, based on our current understanding.
- A paradigm based on vision transformers is proposed for the purpose of dental image classification.
- Experimental assessments confirm the advantages of this study compared to the current state-of-the-art approaches.

## 2 METHODOLOGY

In this section, the details of the proposed method are provided. The first procedure is dataset initialization, the second procedure is dataset augmentation, the third procedure is manual labeling, and the last step includes the proposed vision transformer structure.

### 2.1 Dental Image Dataset Collection

In this study, we first collected the OPG dental panoramic X-ray images from clinical cases. All of the images were obtained from the digital platform, and the resolution of these images is high. In total, 1,418 images were collected in the image dataset initialization process, as shown in Figure 1.

## 2.2 Data Augmentation

To enlarge the training set for the proposed vision transformer, a group of data augmentation operations are employed in this process. Initially, only 1,418 images were captured, and the augmentation methods were used to generate more images for the training set. Finally, 2,836 samples were generated from the initial dataset. And the number of total training samples is 4,254. To be specific, the following augmentation operations were exploited in this procedure, including horizontal flip, vertical flip, rotation, scaling, and shearing, as shown in Figure 2.

## 2.3 Image Labelling

Note that the annotation of the training set plays a vital role in the optimization of deep learning models. Meanwhile, this study proposes a vision transformer network for medical image classification. Accordingly, three experienced dentists were invited to annotate the 1,418 raw images. A majority voting mechanism was adopted in case there is disagreement in this process. In addition, the labeling tool LabelMe [16] (it can be downloaded at https://github.com/wkentaro/labelme) was employed in the annotation procedure.

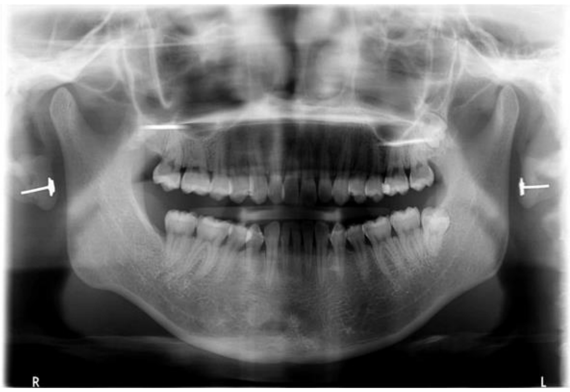

**Figure 1**: An example of the dental image dataset collected in this study (R and L denote right and left, respectively).

Afterward, the resolution of all the collected and augmented images was resized to 600*400 following the requirement of the input of the vision transformer. Moreover, all the image samples were divided into training sets (70%), testing sets (20%), and evaluation sets (10%).

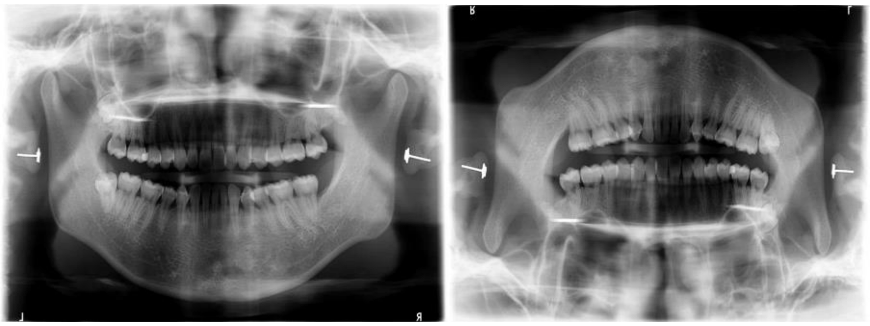

**Figure 2**: A pair of augmented dental X-ray images.

## 2.4 The proposed Deep Learning Model

### *2.4.1 CNN sub-Module*

Plenty of investigations have proved the feasibility of incorporating convolutional modules into transformer models. To note this strategy can improve the classification performance through extracting appropriate features from the local receptive field. In general, the proposed transformer model employs a CNN-based sub-module to generate embeddings from the images, which are then fed into the following vision transformer module. This allows for incorporating local information from the images into the vision transformer. We assumed that the combination of CNN and vision transformer can enhance the adaptability of each other. Therefore, the performance of the integrated model can be guaranteed. To be specific, the incorporated CNN was inspired by the work of [19], which can be used to extract binary descriptors from the images while the structure of extracted features can be regulated. The overall structure of the proposed model is provided in Figure 3.

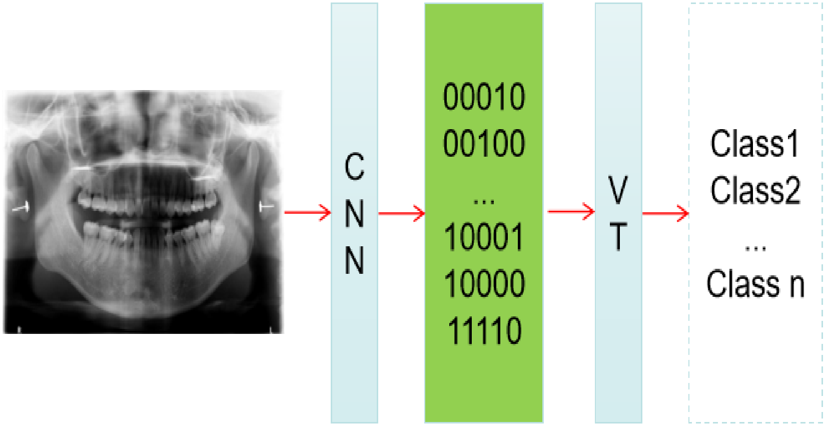

**Figure 3**: The overall structure of the proposed deep learning model.

As depicted in Figure 3, the proposed deep learning pipeline is designed for classifying different types of orthopantomography X-ray images. Initially, it can extract the binary features using the binary descriptor CNN. In addition, the optimal hyper-parameters can be achieved by pre-training. Different from the prior study conducted by Lin et. al [11], which aimed at generating binary descriptors rather than image classification, the proposed approach focuses on semi-supervised learning. To optimally implement a deep learning network for dental image classification, it is necessary to prepare both a set of pre-training samples and a group of dental X-ray images for fine-tuning. The constraints of the binary descriptor include at least the following: distribution uniformity, scale-invariance, and rotation-invariance, as shown in Figure 4.

The constrained embeddings can then be produced using the deep model shown in Figure 4. The corresponding mathematical expression could be formulated as Equation (1):

$$\mathrm{F}(i;W) = f_k(\ldots f_2(f_1(i;w_1);w_2)\ldots;w_k)\,, \tag{1}$$

where $\mathrm{F}(.)$ denotes the introduced CNN model, $\mathrm{f}_k$ represents the operation of each layer in the presented CNN model, and $i$ is the input; $W$ and $w$ are the weighting values of the general CNN model and each layer of the CNN, respectively. The model's weighting values are trained by using the collected and augmented X-ray images. In addition, the utilization of the quantization loss function is exploited for classifying the image samples. The loss function is mathematically expressed as Equation (2).

$$Loss = \alpha \sum_{k=1}^{K} \left\| \mu_k - 0.5 \right\|^2 + \beta \sum_{n=1}^{N} \sum_{\theta} P(\theta) \left\| b_{n,\theta} - b_n \right\|^2 + \gamma \sum_{n=1}^{N} \left\| b_n - F(i_n;W) - 0.5 \right\|^2 , \quad (2)$$

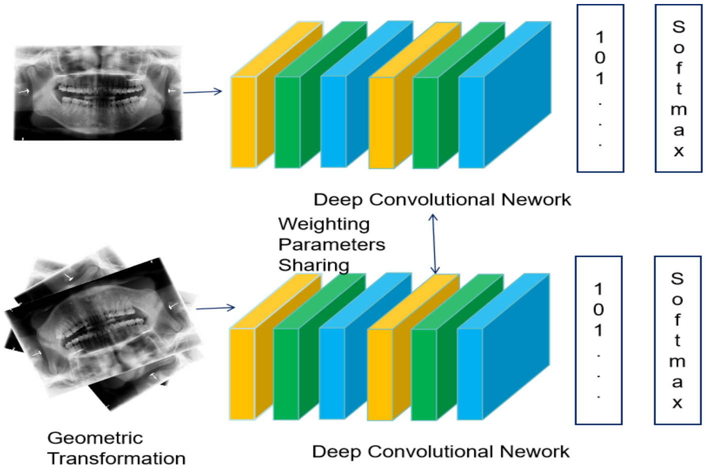

**Figure 4**: The architecture of the introduced CNN model for binary descriptor extraction.

where $K$ denotes the length of extracted features, $\mu_k$ is the mean of the k-th output feature, $N$ denotes the number of input images, and $b$ represents the extracted feature; $P(\theta)$ denotes the penalty function used to regularize the training result according to the scaling and rotating extent $\theta$. In addition, $\alpha$, $\beta$, and $\gamma$ are the regulation terms used to implement the balance between three parts of the constraints. Specifically, the details of the $P(\theta)$ is provided in Equation (3):

$$\mathrm{P}(\theta) = \exp(-\frac{(\theta - \mu)^2}{2\sigma^2}) , \quad (3)$$

### *2.4.2 Vision Transformer Model*

The input of the proposed vision transformer is obtained from the output of the preceding CNN module. To feed the vision transformer, a batch of 16 features extracted from the CNN is taken as the input. To note it follows the input requirement of vision transformers, including Swin transformer [13], PViT [5], and UViT [7]. The vision transformer under consideration is composed of two channels without weighting parameter sharing between each other. The separate channels are employed to address the spatial and temporal information of the input image samples, respectively. The size of the proposed dual-channel vision transformer's input is 16*16, and the RGB color space is adopted. Each channel of the proposed transformer model is introduced from the original vision transformer [3]. Furthermore, the initial input is then flattened into the vectors of length $D$.

According to the fashion of the vision transformer, a trainable class token is integrated into the input of 16*16. Then, the output of the presented vision transformer is taken as the representation of the input. In addition, the positional information is utilized according to the input images, as formulated in Equation (4):

$$\mathrm{Z} = [x_{class}; x_p^1 E; x_p^2 E; ...; x_p^N E] + E_{pos} , \quad (4)$$

where $\mathrm{Z}$ denotes the output of the vision transformer, $x_{class}$ represents the class of the input, and $\mathrm{E}_{pos}$ contains the positional information of each input.

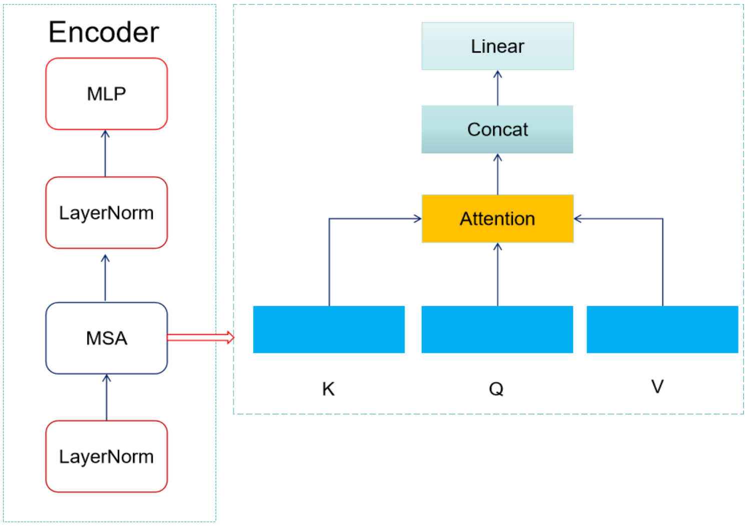

Figure 5: The inner structure of the encoder module in the proposed vision transformer model.

The self-attention mechanism [17], which is used as the encoder in the vision transformer model, is a crucial element in revealing the global connections among distant pixels in the context of vision processing. Figure 5 illustrates the encoder in the presented vision transformer. An encoder utilizes all three components of the input: the key, query, and value. It should be noted that K, Q, and V correspond to key, query, and value, respectively.

Figure 5 illustrates the presence of two main modules: the multi-head self-attention (MSA) module and the multi-layer perception (MLP) module. In addition to these two modules, the proposed encoder design also utilizes two other types of modules: layer normalization module and Gaussian error linear unit (GELU) as the activation module. The weighting value is generated by utilizing the similarity between the query and key. The mathematical formulations for MSA and MLP with layer normalization are presented in Equation (5) and Equation (6) respectively:

$$\mathrm{Z}_L^{'} = MSA(LayerNorm(Z_{L-1})) + Z_{L-1}, \tag{5}$$

$$\mathrm{Z}_{\mathrm{L}} = MLP(LayerNorm(Z_L^{'})) + Z_L^{'} , \tag{6}$$

where $\mathrm{L}$ denotes the layer.

As shown in Figure 5, the $K$ , $Q$ , and $V$ matrices can be mathematically formulated in Equation (7):

$$[K,Q,V] = ZW_{KQV}^{'} , \tag{7}$$

where $\mathrm{W}_{\mathrm{KQV}}$ denotes the weighting matrix. In addition, the output of the encoder is formulated in Equation (8) and Equation (9):

$$\mathrm{O}(Z) = P.V , \tag{8}$$

where:

$$\mathrm{P{=}softmax}(\frac{QK^T}{\sqrt{V}}) , \tag{9}$$

Finally, the output classification result is yielded with the GELU module as the classification head. Moreover, in the last phase of the dual-channel vision transformer, the linear layer is used to integrate the output embeddings derived from the two channels.

## 3 EXPERIMENTAL RESULTS

### 3.1 Implementation Details

The suggested model was trained using the picture samples from ImageNet-ISLVRC [15] as an initial starting point. Furthermore, the dental photos that were acquired manually were utilized to refine and optimize the suggested method. Consequently, the RMSprop algorithm was used as the optimizer, with a learning rate of 0.002 that was reduced by 0.5. Additionally, a batch size of 8 photos was utilized. PyTorch and two NVIDIA Tesla V100 GPUs were employed for the implementation.

The evaluation focused on the effects of two hyper-parameters, namely the number of layers (L) and the number of heads (H), on the proposed vision transformer. Furthermore, comparison tests were carried out to evaluate the suggested model against the state-of-the-art methodologies.

A unique loss function for the CNN-Transformer model was created by combining spatial and temporal loss functions. Furthermore, the performance of the comparison approaches was assessed using the following assessment metrics: Sensitivity, Specificity, Accuracy, and F1 score.

### 3.2 Ablation Study

To identify the most effective combination of the two hyper-parameters for the proposed vision transformer, a series of comparative experiments were conducted. These tests aimed to assess the performance of the models offered under various combinations of hyper-parameters, as illustrated in Table 1.

| *Combination* | *Layer (L)* | *No. of Heads (H)* |
|---|---|---|
| *CV_L4_H4* | *4* | *4* |
| *CV_L4_H8* | *4* | *8* |
| *CV_L8_H4* | *8* | *4* |
| *CV_L8_H8* | *8* | *8* |

**Table 1**: The choices of two hyper-parameters for the suggested model.

Notably, only two hyper-parameters were taken into consideration during the ablation study. And more hyper-parameters would bring more burden to this study. According to the outcome of the proposed on 30% of the manually collected dataset (as shown in Figure 6), the CV_L8_H4 (as shown in Table 1) with eight layers and four heads was chosen as the optimal model.

Furthermore, we performed comparison studies between our method and the most advanced techniques using the complete hand-gathered dataset. The comparison employs cutting-edge approaches, namely Vision Transformer [4], Swin Transformer [13], PViT [5], and UViT [7], as indicated in Table 2.

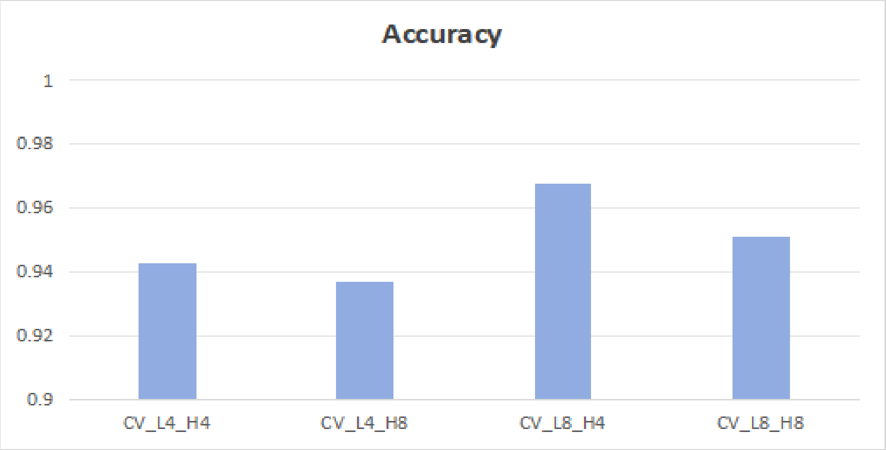

**Figure 6**: The Accuracy outcomes of the proposed model on 30% of the entire dataset.

| *Models* | *Sensitivity* | *Specificity* | *Accuracy* | *F1 score* |
|---|---|---|---|---|
| *Vision Transformer* | *0.921* | *0.919* | *0.925* | *0.928* |
| *Swin Transformer* | *0.927* | *0.931* | *0.919* | *0.916* |
| *PViT* | *0.935* | *0.940* | *0.945* | *0.933* |
| *UViT* | *0.932* | *0.942* | *0.939* | *0.947* |
| *The proposed* | *0.942* | *0.951* | *0.968* | *0.957* |

**Table 2**: The comparison between the state-of-the-art algorithms and the proposed model.

Table 2 presents the findings, which clearly indicate that the suggested model outperforms the state-of-the-art models in terms of Sensitivity, Specificity, Accuracy, and F1 score.

## 4 CONCLUSIONS

To effectively train the proposed deep learning model, a substantial number of picture samples are necessary to carry out the optimization of the weighting parameters. Furthermore, a greater number of samples would ensure that the deep learning network can be effectively tailored to a particular task. It should be noted that including additional photos would lead to the inefficient use of resources and an increased risk of over-fitting. Insufficient visuals, conversely, can reduce the accessibility of the suggested model.

The study combined a dual-channel convolutional neural network (CNN) with a dual-channel vision to create a CNN-Transformer model. This model was used to identify dental X-ray pictures. The experimental findings clearly establish the advantages of the proposed technique. It can thus be inferred that the suggested method could serve as a valuable tool in clinical practice.

*Yilin Li*, http://orcid.org/0009-0008-9565-1303
*Jun Zhang*, http://orcid.org/0000-0002-6068-2504

## REFERENCES

[1] Aeini, F.; Mahmoudi, F.: Classification and numbering of posterior teeth in bitewing dental images, in 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 6, 2010, V6-66-V6-72. http://doi.org/10.1109/ICACTE.2010.5579369

[2] Chen J.; et al.: A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films, Scientific Reports, 9(1), 2019, 1-11. https://doi.org/10.1038/s41598-019-40414-y

[3] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale, 2021 ICLR, 2021. https://doi.org/10.48550/arXiv.2010.11929

[4] Ekert, T.; et al.: Deep learning for the radiographic detection of apical lesions, Journal of Endodontics, 45(7), 2019, 917-922. https://doi.org/10.1016/j.joen.2019.03.016

[5] Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C.: Multiscale vision transformers, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021. https://doi.org/10.48550/arXiv.2104.11227

[6] Gan, Y.; Xia, Z.; Xiong, J.; Li, G.; Zhao, Q.: Tooth and alveolar bone segmentation from dental computed tomography images, IEEE Journal of Biomedical and Health Informatics, 22(1), 2017, 196-204. https://doi.org/10.1109/JBHI.2017.2709406

[7] Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Joon Oh, C.: Rethinking spatial dimensions of vision transformers, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021. https://doi.org/10.48550/arXiv.2103.16302

[8] Krois, J.; et al.: Deep learning for the radiographic detection of periodontal bone loss, Scientific Reports, 9(1), 2019, 1-6. https://doi.org/10.1111/jcpe.12946

[9] Lee, J.-H.; Kim, D.-H.; Jeong, S.-N.; Choi, S.-H.: Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm, Journal of Dentistry, 77, 2018, 106-111. https://doi.org/10.1016/j.jdent.2018.07.015

[10] Lee, J.-H.; Kim, D.-h.; Jeong, S.-N.; Choi, S.-H.: Diagnosis and prediction of periodontally compromised teeth using a deep learningbased convolutional neural network algorithm, Journal of Periodontal & Implant Science, 48(2), 2018, 114. https://doi.org/10.5051/jpis.2018.48.2.114

[11] Lin, K.; Lu, J.; Chen, C. S.; et al.: Learning compact binary descriptors with unsupervised deep neural networks, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE: Piscataway, 2016. https://doi.org/10.1109/CVPR.2016.133

[12] Lin, P.-L.; Lai, Y.-H.; Huang, P.-W.: An effective classification and numbering system for dental bitewing radiographs using teeth region and contour information, Pattern Recognition, 43(4), 2010, 1380-1392. https://doi.org/10.1016/j.patcog.2009.10.005

[13] Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021. https://doi.org/10.48550/arXiv.2103.14030

[14] Mahoor, M. H.; Abdel-Mottaleb, M.: Classification and numbering of teeth in dental bitewing images, Pattern Recognition, 38(4), 2005, 577-586. https://doi.org/10.1016/j.patcog.2004.08.012

[15] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; Li, F.; ImageNet large scale visual recognition challenge, International Journal of Computer Vision (IJCV), 2015, 115(3), 211–252. doi:10.1007/s11263-015-0816-y. https://doi.org/10.48550/arXiv.1409.0575

[16] Russell, B.; Torralba, A.; Murphy, K.; Freeman, W. T.: LabelMe: a database and web-based tool for image annotation, International Journal of Computer Vision, 77(1-3), 2008, 157-173. https://doi.org/10.1007/s11263-007-0090-8

[17] Vaswani, A.; et al.: Attention is All You Need, ArXiv Preprint, 2017, arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

[18] Wu, C.-H.; Tsai, W.-H.; Chen, Y.-H.; Liu, J.-K.; Sun, Y.-N.: Modelbased orthodontic assessments for dental panoramic radiographs, IEEE Journal of Biomedical and Health Informatics, 22(2), 2017, 545-551. https://doi.org/10.1109/JBHI.2017.2660527

[19] Xiao, B.; Hu, Y.; Liu, B.; Bi, X.; Li, W.; G, X.: DLBD: A self-supervised direct-learned binary descriptor, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE: Piscataway, 2023. https://doi.org/10.1109/CVPR52729.2023.01521

[20] Yuniarti, A.; Nugroho, A. S.; Amaliah, B.; Arifin, A. Z.: Classification and numbering of dental radiographs for an automated human identification system, Telkomnika, 10(1), 2012, 137. https://doi.org/10.12928/telkomnika.v10i1.771