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

Identification of Dental Lesions Using Self-Supervised Vision Transformer in Radiographic X-ray Images

Yilin Li[1], Huadong Zhao[2], Daiwei Yang[3], Simeng Du[4], Xiaobin Cui[5] and Jun Zhang[6]*

[1,2,3,4,5,6]Department of Orthodontics, School and Hospital of Stomatology, Cheeloo College of Medicine, Department and Organization, Shandong University & Shandong Key Laboratory of Oral Tissue Regeneration & Shandong Engineering Laboratory for Dental Materials and Oral Tissue Regeneration, Jinan, China
[1]leeyling@126.com, [2]hdzhao0509@163.com, [3]ydw1020117316@163.com,
[4]dsmkathryn@163.com, [5]2571166300@qq.com, [6]zhangj@sdu.edu.cn

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

Abstract. The early identification of dental problems via the employment of imaging equipment has the potential to significantly enhance the accuracy of clinical diagnosis as well as the outcome of treatment. One way to do this is by using imaging technologies. When it comes to imaging technology, X-rays are mostly considered to be the instrument that is used to discover dental diseases. Manual examination of dental images, on the other hand, is a difficult operation that will take a significant amount of time and is prone to mistakes throughout the process. As a result of this, the automated approach to dental diagnosis has been receiving a growing amount of attention, especially in light of the introduction of algorithms for machine learning. In order to address the difficulty of dental picture classification, a great number of deep learning models, in particular convolutional neural networks, have been released. These models have been used to tackle the problem. The outcomes that these models have achieved in a range of benchmarks, including the handling of dental X-ray photos in clinical procedures, have been favorable. The models that were mentioned before, on the other hand, will not display the traits of flexibility and stability when they are applied to situations that are taken from the actual world. The fundamental reason for these problems is that convolutional neural networks have a limited capacity to capture the global linkages between distant pixels. This ability is a key drawback of these networks, which can be ascribed to the fact that it is the primary reason for these problems. Taking this into consideration, the objective of this research is to develop a dual-channel vision transformer model that is only centered on the challenge of dental image categorization. The self-attention module is the key component that is employed by the approach that has been proposed, while the convolutional operators are completely eliminated from the process. Furthermore, the method of transfer learning is used in order to adjust the weighting parameters of the visual transformer that is being shown. This is done in order to get optimal results. The results of the studies provide evidence that the suggested technique has higher performance in comparison to the most powerful deep learning algorithms that are presently available.

Keywords: Dental Image; X-ray; Vision Transformer; Machine Vision; Medical Image Analysis.
DOI: https://doi.org/10.14733/cadaps.2024.S23.332-342

1 INTRODUCTION

In the discipline of dentistry, dental informatics is a relatively young branch of study that has made a significant contribution to the diagnosis and treatment of dental diseases in clinical settings. This is a widely held belief. Reducing the amount of work that dentists are required to undertake, has the potential to reduce the amount of pain that patients experience and to make the whole process of diagnosis more efficient [1] . In dentistry, there are at least the following fields, including restorative dentistry, endodontics, orthodontics, dental surgery, and periodontology [20] . During this time, a wide variety of imaging methods have been applied in the area of dental informatics in order to catch different types of dental diseases that may be found in these locations, such as cavities, dental crowns, fillings, and root canal surgeries [19] . These approaches have been utilized to capture these dental illnesses.

The dentist is responsible for interpreting the information that is included in dental photos, such as CT and X-ray images [20] . Radiographic X-rays that can pass through dental regions may be obtained via the use of X-ray imaging technology. It is important to keep this particular matter in mind. As a result of the efforts of dentists all around the globe, the vast majority of dental conditions may be identified via the inspection of dental pictures. However, manual operation is hard, time-consuming, and prone to mistakes, which significantly hinders the deployment of imaging methods in the area of dentistry. This is because the manual operation is tedious. Several alternative machine-learning methods for dental pictures have been proposed to overcome the challenges that are connected with human evaluation. These algorithms have been the subject of a great deal of research. These computerized algorithms for dental image categorization have the potential to contribute to the detection of dental problems and may even contribute to the prevention of tooth loss. They are geared toward the classification of dental images. Furthermore, they have the potential to be exploited in order to eliminate the problems that were brought about by manual interpretation.

Currently, deep learning models, such as convolutional neural networks (CNN), have been steadily applied in clinical procedures [11] [17] . The convolutional neural network is a strong approach to machine learning that has been created specifically for this purpose. The neural network in question is capable of carrying out tasks such as the identification of images, the segmentation of images, and the classification of images with a high degree of accuracy. To identify deterioration, periapical periodontitis, and periodontal diseases of mild, moderate, and severe severity, deep convolutional neural network (CNN) algorithms were developed and used for clinical dental periapical radiographs. Furthermore, in recent years, a great variety of algorithms that are based on deep learning have been created for dental photos [9] [15] [16] [25] . These algorithms have been successfully implemented. CNN-based models have the potential to be an excellent diagnostic tool for dental images, as shown by the results of the trials that were conducted. This type of model, on the other hand, is afflicted by the limits of local receptive fields, which leads them to ignore the global correlations that exist between long-range pixels in dental photographs. This is a problem since these models are restricted by the local receptive fields.

The goal of this research is to analyze the employment of vision transformers in dental image processing. This evaluation will take into account the analysis that was stated before. To be more specific, a new version of the vision transformer is shown as a way of executing the process of picture classification for X-ray dental pictures. This is done in order to make the procedure more transparent. To be more specific, the self-supervised approach is applied in the vision transformer that is offered in order to make the training process more convenient. In order to accomplish the goal of pretraining the vision transformer model, this is of utmost significance. When compared to other methods, the recommended technique performs much better in the dental photo dataset that was generated manually, as shown by the outcomes of the experiment.

The following is a list of the key contributions that this work makes:

- They generated a dataset that was compiled by hand with great care and attention to detail. A dentist and two assistants played an important role in the process of labeling the manual dataset and were crucial in the procedure. The datasets are each made up of four different groups of classes that are unique from one another. Image-enhancing methods such as magnification, rotation, horizontal flipping, and vertical flipping were performed on the images in order to get the intended effects. These approaches brought about the anticipated outcomes. Long overdue, the process of picture annotation was finally carried out with the assistance of the LabelMe applications.
- The vision transformer-based model that has been built is an early attempt in the area of dental X-ray image categorization, according to our present knowledge of the situation.
- In contrast to the deep learning models that are deemed to be state-of-the-art, the experimental data demonstrate that the technique that has been offered presents a higher performance.

2 RELATED WORK

In the field of dental informatics, several teeth categorization algorithms have been established by using radiography pictures. The periapical radiographs were the primary focus of the research that Hassan and his colleagues [13] conducted. They looked at automated feature recognition, segmentation, and quantification, in that particular sequence. Compared to the Xnet and SegNet models that were used in this scientific investigation, the U-Net-oriented models were able to achieve greater performance in terms of the mean intersection over union (mIoU) and dice coefficient. This was the case because the U-Net models were able to accomplish superior performance. The research conducted by [1] resulted in the development of a method that is capable of identifying and counting teeth in periapical videos. Through the use of a faster region convolutional neural network, sometimes referred to as a faster R-CNN, this method is effectively implemented. According to the data, the model that was provided attained a level of accuracy and recall that was ninety percent higher than the standard benchmark, in addition to a level of mIoU that was ninety-one percent. This was determined by comparing the model to the standard benchmark. Through the use of deep convolutional neural networks (CNNs) in their research, Ekert et al. were able to recognize apical lesions in panoramic dental radiographs [6]. For the purpose of this particular experiment, a total of 2,001 panoramic radiographs were made available, and these radiographs were used in the process of training a CNN model that consists of seven layers. The authors of [12] make use of two different models of multi-sized CNN in order to identify and categorize teeth that are visible in dental panoramic radiographs. This allows them to automatically organize the filing of dental charts. Through the use of a four-fold cross-validation technique, the testing data set was able to reach a high level of accuracy in the object detection network. Convolutional neural networks (CNNs) are used by Fukuda et al. in their research to identify vertical root fractures (VRF) in panoramic radiographs [8] . The CNN that was used was developed using the utilization of Detect Net five-fold cross-validation, and DIGITS version 5.0 was utilized in order to enhance the dependability of the model. One hundred and fifty percent accuracy was attained by the neural network. The evaluation of radiographic bone loss and the generation of image-based periodontal diagnostics might be beneficially accomplished with the help of neural networks [21] . To reveal dental restorations, artificial intelligence may be used.

The findings of the research [1] indicate that artificial intelligence might be used in the field of restorative dentistry to recognize and classify dental restorations. 93.6% of dental restorations were detected by the algorithms that were applied in their investigation, which was conducted on 83 panoramic photographs. In addition, the distribution and shape of the grey values were used in order to classify the restorations into eleven distinct categories. Panoramic pictures and convolutional neural networks were used by Chang et al. [3] in order to identify periodontal bone level (PBL), cemento enamel junction level (CEJL), and teeth in order to arrive at a diagnosis of periodontitis stage. During the research, an automated method was used to calculate and classify the percentage of bone loss [2] . In research [14] , convolutional neural networks were applied to make a prediction about whether or not the extraction of the third molar may eventually lead to paresthesia of the inferior alveolar nerve. The extraction of the lower third molar is one of the most frequent operations performed in the field of dentistry. It is possible for the nerve to feel paresthesia after the removal of a wisdom tooth from the jaw. A series of panoramic photographs were obtained before the extraction, and CNN took advantage of the connection that exists between the nerve canal and the tooth roots in order to provide a prediction about the possibility of nerve paresthesia. It is necessary to do more studies since, according to the findings of scientists, the use of two-dimensional images in panoramic radiography may result in a greater number of false positive and false negative results.

3 METHODOLOGY

Within this section, you will find a comprehensive presentation of facts about the strategy that has been prescribed. The first stage is the initialization of the dataset, the second step is the augmentation of the dataset, the third step is the manual labeling, and the last step is the description of the recommended vision transformer model. All are steps in the process.

3.1 Dataset Preparation

We began by collecting panoramic X-ray pictures of the oral cavity from clinical patients as part of our inquiry. These photos were taken throughout the investigation. Every single one of the shots was obtained from the digital platform, and the resolution of these pictures is much higher than average. During the time that the procedure of initializing the picture dataset was being carried out, a total of 1,418 images were gathered.

3.2 Data Augmentation

In order to expand the size of the training set that is required for the proposed vision transformer, a range of data augmentation approaches are applied at this part of the process. Since there were only 1,418 photos captured at the beginning of the procedure, the augmentation approaches were applied in order to create a greater number of images for the training set. When everything is taken into consideration, there are a total of 2,836 samples that were designed using the primary dataset. Regarding the overall number of training examples, there are 4,254 of them now available. In this specific approach, the following augmentation processes were utilized: shearing, rotation, scaling, horizontal flip, and vertical flip. For the sake of greater precision, these processes were carried out.

3.3 Image Labelling

It is essential to bear in mind that the annotation of the training set serves as a crucial aspect in the optimization of deep learning models. This is something that should be paid attention to. In the current investigation, which is currently being carried out, a vision transformer network is being suggested for application in the categorization of medical images. Consequently, in order to engage in the annotation of the 1,418 raw photographs, it was requested that three dentists who had a great deal of expertise be involved. The implementation of a voting system that is based on a majority vote was done in preparation for the possibility of a dispute occurring throughout this procedure. In addition to that, the procedure of annotation was carried out with the aid of a labeling tool that is known as LabelMe [24] .

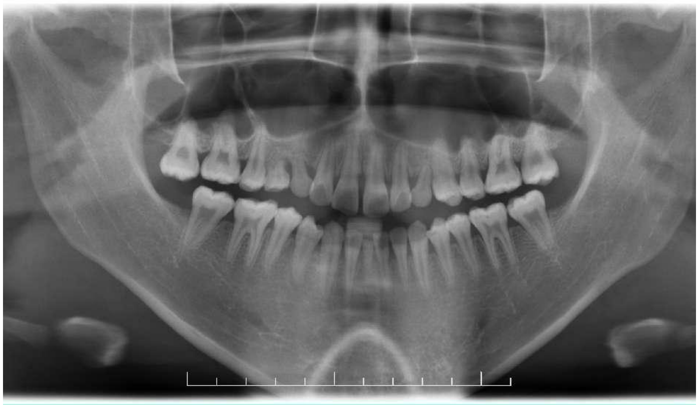

Figure 1: This is an example of one of the dental pictures that were acquired by manual collection (The photographs have been edited to remove any personal information that may have been associated with the patients).

The resolution of all of the photographs that were gathered and improved was then scaled to 600 by 400 pixels in line with the specifications of the vision transformer's input. This was done after the previous step. In addition, the whole collection of picture samples was divided into three unique sets: the training set, which made up seventy percent of the total, the testing set, which made up twenty percent of the total, and the evaluation set, which made up ten percent of the total.

3.4 The proposed Deep Learning Model

The model of the vision transformer that has been provided adheres to a standard framework for vision transformers, such as the PViT [7] , UViT [10] , and Swin transformers [18] . This framework is included in the list of vision transformers. Within the framework of the pre-training procedure, the model that has been suggested takes into consideration two different points of view for every single input picture. Image patches that are 16 pixels by 16 pixels in size (position) and visual tokens (feature) are included in these viewpoints. The procedure of tokenization was used in order to convert the original photographs into visual tokens. A collection of the picture patches was masked and then fed into the backbone component of the vision transformer that was presented. This was in addition to the previous point. It is anticipated that the pre-training will be able to extract the visual tokens that have been specified in line with the masked picture patches. Adding the classification layer to the pre-trained encoder in the vision transformer was essential in order to accomplish the classification job using the vision transformer properly. This was important in order to get the desired results.

This inquiry encompasses the presentation of the self-supervised vision transformer model, which was generated from the masked image modeling scheme and was inspired by BERT [4] . This model was developed within the scope of this investigation. Through the process of separating the input photos into individual patches, a grid of picture patches was produced. These picture patches depict the concept of the vision transformer. In addition to this, the photographs were concurrently translated into discrete tokens by the use of the discrete model [22] . As part of the pre-training operation, a section of the image patches was masked, and then they were supplied into the transformer. The visual tokens, as opposed to the bits that had been hidden, were retrieved as a consequence of this consequence. Following that, the picture classification job was exposed to both self-supervised learning and fine-tuning. Through the process of fine-tuning using the ImageNet labels, it is possible to enhance the overall performance of the classification system.

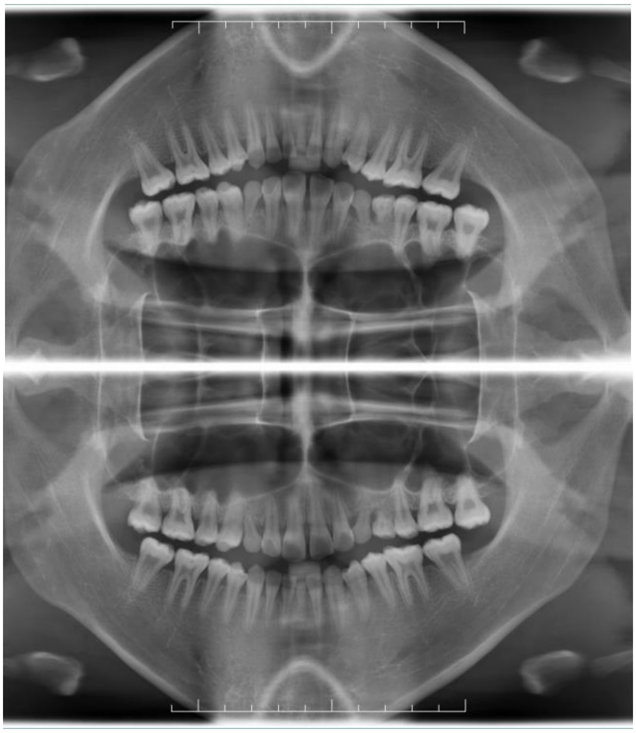

Figure 2: In the manual dataset, two photographs have been enhanced.

Tokens that are capable of being trained are included in the input in line with the architecture of the vision transformer. It is being taken into account that the output of the vision transformer that is being given is being used as a representation of the input. In addition, the positional information is employed in a manner that is consistent with the pictures that are supplied, as shown in Equation (1):

$$\mathrm{O} = [x_{class}; x_p^1 E; x_p^2 E; ...; x_p^N E] + E_{pos} \ , \tag{1}$$

where O symbolizes the output of the vision transformer, x represents the class of the input, and E_{pos} holds the positional information of each input are all included in this expression.

After being presented with an image, the suggested model can encode it into vector representations, as seen in Figure 3. Following the division of the two-dimensional pictures into a collection of image patches, any vision transformer can take these patches as input. Specifically, a picture with a resolution of $\mathrm{H} \times \mathrm{W} \times \mathrm{C}$ is scaled into $N = HW / P \times P$ patches, where C is the number of channels, $\mathrm{H} \times \mathrm{W}$ is the input resolution and $\mathrm{P} \times \mathrm{P}$ is the resolution of each image patch. This provides a clearer understanding of the process. The work of BERT [4] serves as the source of inspiration for the process of flattening into vectors and linearly projecting a series of picture patches that are obtained from the same image. H = 16, W = 16, and P = 16 are the starting parameters of the input pictures. As seen in Figure 3, the process of visual token learning is comprised of two modules, namely the tokenizer and the decoder. It is important to keep in mind that the decoder is not used throughout the process of picture categorization. The tokenizer can convert picture pixels into distinct tokens by adhering to a language. It is the responsibility of the decoder to retrieve the input picture following the visual tokens at hand. Since the visual tokens are not continuous, the training is not comparable to other methods. After that, around forty percent of the picture patches are randomly masked. The masked image patches also have a learnable embedding added to them as an additional feature. In order to make a prediction about the visual tokens for each masked patch, a softmax classifier is used. Last but not least, the objective of the pre-training is to maximize the log-likelihood of the visual tokens given the pictures that have been masked.

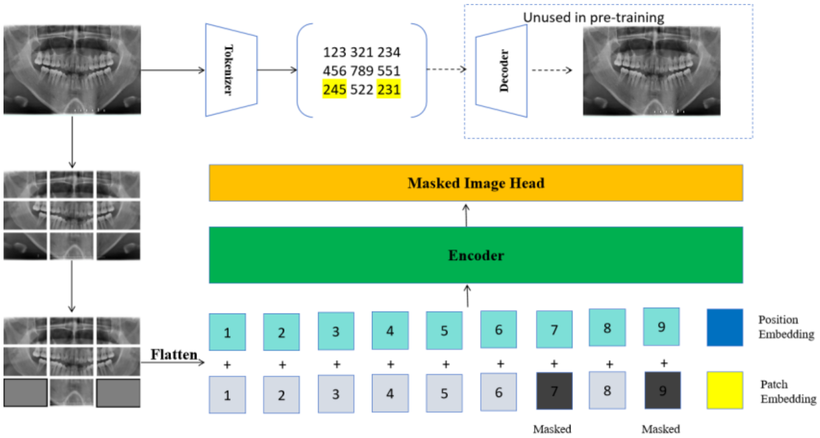

Figure 3: A look at the overall structure of the vision transformer concept that has been presented. A tokenizer for images is used in order to produce the tokens. Each picture is shown in two different perspectives during the pre-training phase. These views include image patches and visual tokens. Some picture patches, often known as gray patches, have been disguised.

Additionally, the encoder shown in Figure 3 incorporates two essential modules, namely the multi-head self-attention (MSA) and the multi-layer perception unit (MLP). Furthermore, the proposed encoder design incorporates two additional modules, namely the layer normalization module and the Gaussian error linear unit (GELU) as the activation module. Both of these modules are used in the process of encoding data. The formation of the weighting value is accomplished by the use of the similarity that exists between the query and the key. In terms of layer normalization, the mathematical formulas for MSA and MLP are provided by Equation (2) and Equation (3), respectively. Listed below are the formulas in question:

$$\mathrm{Z}_L^{'} = MSA(LayerNorm(Z_{L-1})) + Z_{L-1}, \tag{2}$$

$$\mathrm{Z}_\mathrm{L} = MLP(LayerNorm(Z_L^{'})) + Z_L^{'}, \tag{3}$$

where L denotes the layer.

It is possible to mathematically define the matrices K, Q, and V that are used in the encoder module using Equation (4):

$$[K,Q,V] = ZW_{KQV}^{'}, \tag{4}$$

where $\mathrm{W}_{\mathrm{KQV}}$ represents the matrix used for weighting. Furthermore, the output of the encoder may be expressed using Equation (5) and Equation (6), which are as follows:

$$\mathrm{O}(Z) = P.V, \tag{5}$$

where:

$$\mathrm{P{=}softmax}(\frac{QK^T}{\sqrt{V}}), \tag{6}$$

As an additional point of interest, the GELU module acts as the classification head, which finally leads to the production of the output classification result. Additionally, in the last phase of the dual-channel vision transformer, the linear layer is used in order to integrate the output embeddings that are received from the two channels. This is done in order to meet the requirements of the final phase.

4 EXPERIMENTAL RESULTS

As an initial point of departure, the recommended model was trained with the use of the image samples that were obtained from ImageNet-ISLVRC [23] . During the pre-training phase, there are 800 epochs, and the batch size is 2k. Adam is used for optimization purposes, and the values are β_1 = 0.9 and β_2 = 0.99. In addition to a warmup period of ten epochs and a cosine learning rate decline, the learning rate is given a value of 1.0e-3. There is a 0.04 percent weight loss. In order to complete the training procedure, which requires sixteen Nvidia Telsa V100 32GB GPU cards, it takes roughly four days.

During the assessment, the impacts of two hyper-parameters, namely the number of layers (L) and the number of heads (H), on the suggested vision transformer were the primary focus of attention. In addition, comparative tests were carried out in order to assess the proposed model in comparison to the approaches that are considered to be state-of-the-art. Through the combination of spatial and temporal loss functions, a one-of-a-kind loss function was developed for the CNN-Transformer mathematical model. To add insult to injury, the performance of the comparison methods was evaluated by using the following evaluation metrics: sensitivity, specificity, accuracy, and F1 score.

4.1 Ablation Study

In order to determine which of the two hyper-parameters for the proposed vision transformer would be the most successful combination, a series of comparison tests were carried out. Particularly noteworthy is the fact that throughout the ablation investigation, just two hyper-parameters were taken into account. An additional set of hyper-parameters would add an additional load to this investigation. As a result of the results of the proposal on twenty percent of the information that was manually gathered (as shown in Figure 4), the VT_L16_H16 model, which has sixteen layers and sixteen heads, was selected as the best model.

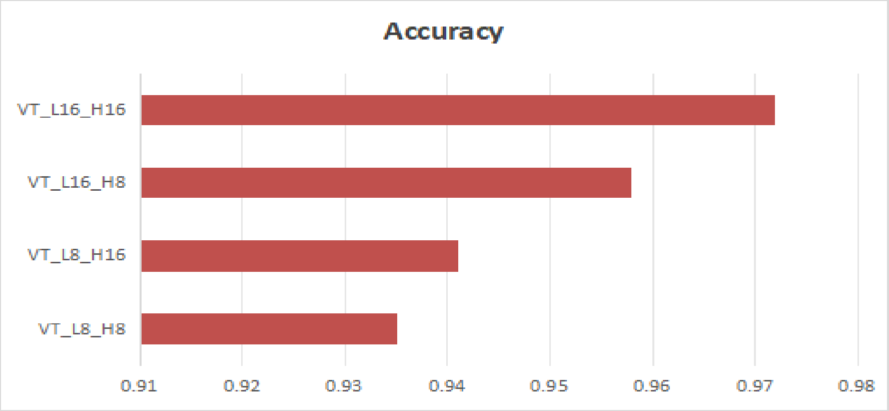

Figure 4: The Accuracy of the proposed model on part of the dataset using different combinations of the hyper-parameters.

4.2 Comparison Experiments

In addition, we conducted comparative research between our approach and the most cutting-edge methodologies, making use of the whole dataset that was acquired by hand. The comparison makes use of the following vision transformer-based techniques, which are as follows: Vision Transformer [5] , PViT [7] , and UViT [10] , as well as Swin transformer [18] , as shown in Table 1.

The data are shown in Table 1, and they make it abundantly evident that the proposed model does better than the models that are considered to be state-of-the-art in terms of sensitivity, specificity, accuracy, and F1 score.

Models	*Sensitivity (%)*	*Specificity (%)*	*Accuracy (%)*	*F1 score (%)*
Vision Transformer	*92.1*	*91.9*	*92.5*	*92.8*
PViT	*93.5*	*94.0*	*94.5*	*93.3*
UViT	*93.2*	*94.2*	*93.9*	*94.7*
Swin Transformer	*92.7*	*93.1*	*91.9*	*91.6*
The proposed	*96.8*	*95.9*	*97.2*	*97.1*

Table 1: The comparison between the most advanced models and the proposed approach.

5 CONCLUSIONS

In order to train the suggested deep learning model efficiently, it is important to have a significant amount of data samples in order to carry out the optimization of the weighting parameters. Additionally, a bigger number of samples would guarantee that the deep learning network is capable of being efficiently adapted to a certain job. Note that incorporating extra photographs will result in an inefficient use of resources and an increased danger of over-fitting. This is something that should be taken into consideration.

An innovative transformer model that makes use of a self-supervised process is proposed in this paper. In order to recognize dental X-ray pictures, the model that was provided was used. It is abundantly obvious that the benefits of the suggested method are established by the outcomes of the experiment. As a result, one may draw the conclusion that the approach that was presented has the potential to be an effective instrument in clinical practice.

Yilin Li, http:// orcid.org/0009-0008-9565-1303
Huadong Zhao, http://orcid.org/0009-0000-9467-1689
Daiwei Yang, http://orcid.org/0009-0001-7501-7347
Simeng Du, http://orcid.org/0009-0009-8457-8689
Xiaobin Cui, http://orcid.org/0009-0008-1690-0018
Jun Zhang, http://orcid.org/0000-0002-6068-2504

REFERENCES

[1] Abdalla-Aslan, R.; Yeshua, T.; Kabla, D.; Leichter, I.; Nadler, C.: An artificial intelligence system using machine-learning for automatic detection and classification of dental restorations in panoramic radiography, Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology, 130, 2020, 593-602. https://doi.org/10.1016/j.oooo.2020.05.012

[2] Cha, J.-Y.; Yoon, H.-I.; Yeo, I.-S.; Huh, K.-H.; Han, J.-S.: Peri-Implant Bone Loss Measurement Using a Region-Based Convolutional Neural Network on Dental Periapical Radiographs, Journal of Clinical Medicine, 10, 2021, 1009. https://doi.org/10.3390/jcm10051009

[3] Chang, H.-J.; Lee, S.-J.; Yong, T.-H.; Shin, N.-Y.; Jang, B.-G.; Kim, J.-E.; Huh, K.-H.; Lee, S.-S.; Heo, M.-S.; Choi, S.-C.; et al.: Deep Learning Hybrid Method to Automatically Diagnose Periodontal Bone Loss and Stage Periodontitis, Scientific Reports, 10, 2020, 753. https://doi.org/10.1038/s41598-020-64509-z

[4] Devlin J.; Chang M.; Lee K.; Toutanova K.: BERT: pretraining of deep bidirectional transformers for language understanding, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2019, 4171-4186. https://doi.org/10.18653/v1/N19-1423

[5] 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

[6] 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

[7] 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

[8] Fukuda, M.; Inamoto, K.; Shibata, N.; Ariji, Y.; Yanashita, Y.; Kutsuna, S.; Nakata, K.; Katsumata, A.; Fujita, H.; Ariji, E.: Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography, Oral Radiology, 36, 2020, 337–343. https://doi.org/10.1007/s11282-019-00409-x

[9] Geetha, V.; Aprameya, K.S.; Hinduja, D.M.: Dental caries diagnosis in digital radiographs using back-propagation neural network, Health Information Science and Systems, 8, 2020, 8. https://doi.org/10.1007/s13755-019-0096-y

[10] 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

[11] Imangaliyev, S.; Veen, M.H.; Volgenant, C.; Keijser, B.J.; Crielaard, W.; Levin, E.: Deep learning for classification of dental plaque images, In International Workshop on Machine Learning, Optimization, and Big Data; Springer: Cham, Switzerland, 2016, 407-410. http://doi.org/10.1007/978-3-319-51469-7_34

[12] Johari, M.; Esmaeili, F.; Andalib, A.; Garjani, S.; Saberkari, H.: Detection of vertical root fractures in intact and endodontically treated premolar teeth by designing a probabilistic neural network: An ex vivo study, Dento maxillo facial Radiology, 47, 2017, 20160107. https://doi.org/10.1259/dmfr.20160107

[13] Khan, H.A.; Haider, M.A.; Ansari, H.A.; Ishaq, H.; Kiyani, A.; Sohail, K.; Muhammad, M.; Khurram, S.A.: Automated feature detection in dental periapical radiographs by using deep learning, Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology, 131, 2021, 711-720. https://doi.org/10.1016/j.oooo.2020.08.024

[14] Kim, B.S.; Yeom, H.G.; Lee, J.H.; Shin, W.S.; Yun, J.P.; Jeong, S.H.; Kang, J.H.; Kim, S.W.; Kim, B.C.: Deep Learning-Based Prediction of Paresthesia after Third Molar Extraction: A Preliminary Study, Diagnostics, 11, 2021, 1572. https://doi.org/10.3390/diagnostics11091572

[15] Lee, C.T.; Kabir, T.; Nelson, J.; Sheng, S.; Meng, H.W.; Van Dyke, T.E.; Walji, M.F.; Jiang, X.; Shams, S.: Use of the deep learning approach to measure alveolar bone level. Journal of Clinical Periodontology, 49, 2022, 260-269. https://doi.org/10.1111/jcpe.13574

[16] Lee, J.-H.; Han, S.-S.; Kim, Y.H.; Lee, C.; Kim, I.: Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. Oral

Surgery, Oral Medicine, Oral Pathology and Oral Radiology, 129, 2019, 635-642. https://doi.org/10.1016/j.oooo.2019.11.007

[17] Liu, L.; Xu, J.; Huan, Y.; Zou, Z.; Yeh, S.C.; Zheng, L.R.: A smart dental health-IoT platform based on intelligent hardware, deep learning, and mobile terminal. IEEE Journal of Biomedical and Health Informatics, 24, 2019, 898-906. http://doi.org/10.1109/JBHI.2019.2919916

[18] 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

[19] Oprea, S.; Marinescu, C.; Lita, I.; Jurianu, M.; Visan, D.A.; Cioc, I.B.: Image processing techniques used for dental X-ray image analysis, In Proceedings of the 2008 31st International Spring Seminar on Electronics Technology, Budapest, Hungary, 7–11, May, 2008, 125–129. http://doi.org/10.1109/ISSE.2008.5276424

[20] Ossowska, A.; Kusiak, A.; Świetlik, D.: Artificial Intelligence in Dentistry—Narrative Review, International Journal of Environmental Research and Public Health, 19, 2022, 3449. http://doi.org/10.3390/ijerph19063449

[21] Pakbaznejad Esmaeili, E.; Pakkala, T.; Haukka, J.; Siukosaari, P.: Low reproducibility between oral radiologists and general dentists with regards to radiographic diagnosis of caries, Acta Odontol. Scand, 76, 2018, 346–350. https://doi.org/10.1080/00016357.2018.1460490

[22] Ramesh A.; Pavlov M.; Goh G.; Gray S.; Voss C.; Radford A.; Chen M.; Sutskever I.: Zero-shot text-to-image generation, ArXiv, abs/2102.12092, 2021. https://doi.org/10.48550/arXiv.2102.12092

[23] 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. https://doi:10.1007/s11263-015-0816-y

[24] 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

[25] Silva, G.; Oliveira, L.; Pithon, M.: Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives, Expert Systems with Applications, 107, 2018, 15-31. https://doi.org/10.1016/j.eswa.2018.04.001