Research Article

# Comparison of VT-based and CNN-based Models on Teeth Segmentation

Silan Fidan Vural, Nida Kumbasar*

*Abstract*— **Semantic segmentation is a crucial task in computer vision with a wide array of applications across various fields, especially in medical imaging. One of the most important applications of semantic segmentation is in the field of dentistry, where teeth segmentation plays a significant role in diagnosing and treating oral health issues. Accurate segmentation of teeth in dental images is vital for detecting abnormalities, planning treatments, and monitoring the progress of dental procedures. In this paper, a comprehensive comparative analysis is presented, focusing on the use of Convolutional Neural Network (CNN)-based and Vision Transformer (VT)-based models for image segmentation within the context of dentistry. The paper presents a comparison of eight different models, contributing to the literature on dental image segmentation and showcasing practical applications in clinical dental settings. The research presented in this study uses several state-of-the-art segmentation models, namely U-Net, LinkNet, and Swin U-Net, along with different backbones to perform teeth segmentation on publicly available two datasets: one representing adults and the other children. The experiments were conducted to determine which models and backbones provided the best segmentation performance for each dataset. The study also emphasizes that the segmentation modeling process should be handled separately since the alignment of child and adult teeth is different. The U-Net model with the ResNet101 backbone achieved the best performance on the adults dataset, while for the children dataset, the U-Net model with the same ResNet101 backbone also demonstrated superior results. The highest Dice scores obtained were 0.9543 for the adults dataset and 0.9019 for the children dataset, indicating the effectiveness of these models in accurately segmenting teeth. The findings from this research demonstrate the potential of deep learning techniques in improving the accuracy and efficiency of dental diagnosis and treatment planning. Codes used throughout the study will be publicly available at https://github.com/FidanVural/Teeth-Segmentation-in- Panoramic-Radiography/tree/main**

*Index Terms*—**Convolutional Neural Networks, Panoramic Radiography, Semantic Segmentation, Vision Transformer**

**Silan Fidan Vural**, is with TUBITAK, Informatics and Information Security Research Center (BILGEM), Gebze, Kocaeli, 41470, Turkey,(e-mail: fsilanvural@hotmail.com).

https://orcid.org/0009-0000-9488-3809

**Nida Kumbasar**, is with TÜBİTAK, Informatics and Information Security Research Center (BİLGEM), (e-mail: nida.kumbasar@tubitak.gov.tr).

https://orcid.org/0000-0001-5497-4618

Manuscript received Mar 23, 2024; accepted Feb 24, 2025
DOI: 10.17694/bajece.1457754

## I. Introduction

PANORAMIC RADIOGRAPHY (PR) is the most preferred 2 Dimension (2D) imaging technique in the dental field due to its relatively low radiation rate, fast results and low cost [1]. In PR teeth, gingiva, jaw bones, sinus cavities, temporomandibular joint and surrounding anatomical structures as well as intraoral diseases such as tooth decay, gum diseases, cysts, tumors, lesions, etc. can be analyzed in detail. PR is important for planning and follow-up in tooth extraction, dental implants, prosthetic applications and oral surgical procedures.

Deep Learning (DL) applications in the analysis of dental imaging techniques are becoming increasingly common. Caries detection [2], cyst and jaw tumor differentiation [3], root fracture detection [4], periodontal disease detection [5], tooth segmentation [6], jaw segmentation [7], wisdom teeth analysis [8], [9]dental biometric systems [10], [11] are some of the application areas. The combination of Artificial Intelligence (AI) tools with the dentist's vision improves the diagnostic treatment process by predicting diseases and outcomes for complex cases.

Teeth segmentation provides a visual reference for dentists to evaluate the condition of the teeth and closely follow the diagnosis and treatment process. In addition, teeth segmentation guides the diagnosis and treatment process of oral problems directly related to the teeth such as prosthesis - implant placement, tooth extraction, scaling, braces treatment, tooth decay detection. As various 3D anatomical structures overlap on 2D PR, distortions occur in the image. Moreover, the image quality varies from device to device and the low contrast of the image makes it difficult to perform teeth segmentation manually. Considering these factors, the correct teeth segmentation depends on the experience and availability of dentists [12]. Accurate and fully automated teeth segmentation is important to improve the performance of the clinical process.

This paper focuses its attention on the application of image segmentation in dentistry. The main motivation of the work is to present a comparative study on Vision Transformer (VT)-based and Convolutional Neural Network (CNN)-based tooth segmentation using two separate datasets which are adults and children with different patterns and sizes. Another motivation can be explained that performing these models with different backbone architectures.

The contributions in this paper can be outlined as follows:

- Inter-model and intra-model backbones comparison for tooth segmentation with PR is presented.

- The comparison of CNN-based models with VT-based models was performed on both children dental dataset with limited data and adults dental dataset with larger data volumes.

- Due to the relative scarcity of segmentation on children teeth datasets, experiments were conducted to contribute to the literature. To the best of our knowledge, there are no applications comparing DL and VT algorithms on a pediatric tooth dataset.

The rest of the study is organized as follows: Related works are explained in Section 2. Material and method are defined in Section 3. Experiments and experimental results are presented in Section 4. Section 5 which is discussion includes comparison with the literature. Finally, Section 6 outlines the study for teeth segmentation and mentions future work plans.

## II. RELATED WORK

Semantic segmentation is one of the fundamental Computer Vision (CV) tasks used in a wide range of fields, from autonomous vehicle systems to medical images. It classifies each pixel in the image according to predefined categories.

CNN-based segmentation models were quite popular in medical image segmentation task at first. U-Net [13] is one of the most used CNN-based models in medical image segmentation which has an encoder-decoder architecture with skip connections. There are other CNN-based models like Fully Convolutional Networks (FCN) [14] and LinkNet [15] has also demonstrated significant success in medical data. On the other hand, these models have been employed in teeth segmentation with considerable success. Furthermore, teeth segmentation plays a crucial role in assisting dentists in understanding the state of dental health. Even though CNN-based models have achieved substantial success, novel approaches based on transformers have emerged due to CNNs' lack global contextual understanding of images.

Transformers first gained prominence in the field of Natural Language Processing (NLP) [16], creating a profound impact. Subsequently, transformers entered the field of CV and initially achieved state-of-the-art (SOTA) performance in image classification [17]. After this paper [17], many new VT models emerged and vision transformers become quite popular in the field of CV such as image segmentation and object detection. Many models appeared like Swin U-Net [18], TransUNet [19] and PromptUNet [20] for medical image segmentation.

Many studies in the literature have demonstrated the potential of CNN-based or VT-based teeth segmentation approaches to assist clinicians in dental imaging.

Silva et al. [21] applied the Mask R-CNN technique for automatic teeth segmentation on PR. Koch et al. [22] and Sivagami et al. [23] proposed the use of the U-Net network for teeth segmentation, while Jader et al. [24] utilized Mask R-CNN for segmentation of teeth. Wirtz et al. [25] added a modeling process to the R-CNN mask and manually annotated individual tooth forms on PR. Lee et al. [26] performed teeth segmentation automatically with a fine-tuned Mask R-CNN. Zhao et al. [27] integrated the U-Net architecture into the segmentation branch to improve the segmentation effect in the Mask R-CNN model and presented comparative results with U-Net and Mask R-CNN. Similarly, Silva et al. [28] comparatively analyzed teeth segmentation with Mask R-CNN, PANet, HTC, and ResNeSt. Sheng et al. [29] presented a comparative study of teeth segmentation with U-Net, LinkNet FPN, and Swin U-Net methods on PR images. Arora et al. [30] proposed a novel multimodal CNN architecture in which the encoder part consists of conventional CNN, atrous-CNN and separable CNN, and the decoder part consists of a single stream of deconvolutional layers for segmentation. Kanwal et al. [31] implemented a novel architecture for teeth segmentation on PR images that utilizes a dual-path transformer-based network integrated with a panoptic quality loss function. Dhar et al. [32] proposed a novel approach to teeth segmentation with PR by adding grid-based attention gates to the skip links of FUSegNet. Ghafoor et al. [33] proposed a new teeth segmentation model that combines an M-Net-like structure with swin transformers and teeth attention block. Zhang et al. [34] collected PR images used for different purposes in the literature for teeth segmentation. In addition, they prepared a child-specific PR dataset that was not previously available in the literature and shared it publicly available [34]. They performed teeth segmentation with U-Net, PSPNet, R2 U-Net and DeepLab V3+ using both adults and children's PR data separately and together. Brahmi et al. [35] employed Mask R-CNN for instance segmentation of teeth on PR. There is an increasing number of studies in the literature that create a model by grouping children's PRs separately from adults'. Asci et al. [36] utilized U-Net to perform dental caries segmentation in the PRs of children in primary dentition, mixed dentition and permanent dentition. Wathore et al. [37] introduced a new bilateral symmetry-based enhancement method specifically designed to improve tooth segmentation in PR and evaluated the effectiveness of the proposed method using U-Net, SE U-Net and TransUNet. Altan et al. [38] performed tooth segmentation using PR with Mask R-CNN on ResNet-50 backbones.

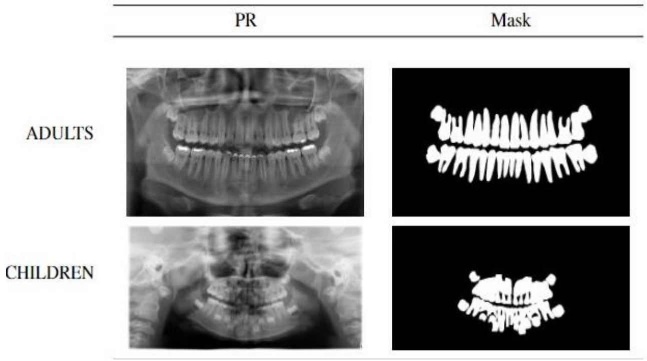

Fig.1. Examples of datasets

## III. MATERIAL METHOD

### A. Material

In this study, the publicly available "Dataset and Code" and "Children's Dental Caries Segmentation Dataset" mentioned in Zhang et al. [34], were used as two different datasets, namely

Adults Dataset and Children Dataset, respectively. The adults dataset consists of 1978 PR teeth images and masks with different image sizes, 1500 images for training, 202 images for validation and 276 images for testing. The children dataset consists of 193 image-mask pairs with non-standard image size obtained from pediatric patients between the ages of 2 and 13. Of these images, 148 were used for training, 15 for validation and 30 for testing. In the datasets, images have distinct sizes but we resized them to a fixed size. The examples of the datasets can be seen in Fig.1. Also, the distribution of the datasets was shown in Table 1.

TABLE I
LIST OF DATASETS USED IN THIS STUDY

| | Train | Validation | Test | Total |
|---|---|---|---|---|
| Adults Dataset | 1500 | 202 | 276 | 1978 |
| Children Dataset | 148 | 15 | 30 | 193 |

## B. *Method*

Our experiments were conducted not only using well-known CNN-based architectures which are U-Net and LinkNet but also using VT architecture which is Swin U-Net.

### *1) U-Net*

U-Net which is developed by Olaf Ronneberger et al. [13] is one of the most important and successful algorithms with an encoder-decoder architecture used in the field of medical image segmentation. U-Net can be represented as a function $f: X \rightarrow Y$, where $X$ denotes the input image and $Y$ represents the output segmentation mask. The encoder part can be mathematically defined as a function $E(x)$ maps the input image $X$ to a latent feature space $z$.

$z$ is calculated using Equation (1),

$$z = f(W * X + b) \qquad (1)$$

where $W$ and $b$ are weights and biases, $*$ denotes the convolution operation, and f is the activation function which is ReLU. The encoder part of the network is responsible for extracting features and learning the representations of input image. The encoder network is usually nothing more than the classification architectures like VGGNet or ResNet. It can be used various networks for the encoder. On the other hand, the decoder part of the network is utilized for generating a segmentation belonging to the input image using encoder representation. In addition to this, there are lots of skip connections between the encoder and decoder networks. These skip connections, also known as shortcut connections, provide information transfer from the encoder to the decoder in order to obtain better segmentation results. Also, U-Net architecture is a modified and extended version of the FCN.

### *2) LinkNet*

LinkNet [15] is also a DL architecture designed for image segmentation tasks, particularly semantic segmentation. The LinkNet architecture, characterized by its encoder-decoder design, is quite similar to the U-Net. Input of each encoder layer is also passed to the output of its corresponding decoder to obtain better results in segmentation by preserving to spatial information. These processes are called skip connections. Moreover, layers in the LinkNet are not concatenated to each other through skip connections like in U-Net; instead, they are summed. The difference is visualized in Fig.2.

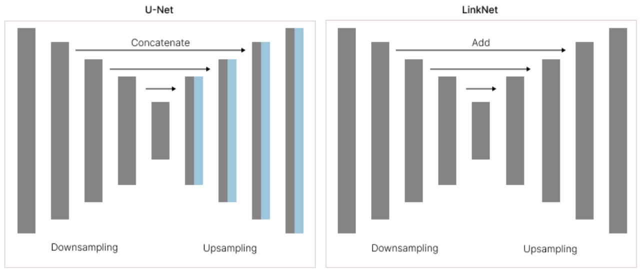

Fig.2. Difference between U-Net and LinkNet

### *3) Swin U-Net*

Within the scope of this research, another architecture which is Swin U-Net was also employed. Swin U-Net model is U-Net shaped encoder-decoder architecture that consists of Swin transformers [18]. Firstly, the image is divided into non-overlapping patches. The number of patches, denoted as $N$, can be calculated using the formula in Equation (2),

$$N = (\frac{H \times W}{P^2}) \qquad (2)$$

where $H$ is the height of the image, $W$ is the width of the image, and $P$ is the patch size. After that, linear embedding is performed to change the channel size of the input. Swin transformer blocks are applied to these token patches. Each Swin transformer block consists of two successive swin transformer modules. While the first module consists of layer normalization (LN), window based multi-head self attention (W-MSA) and Multi Layer Perceptron (MLP), the second module composed of layer normalization, shifted window based multi-head self attention (SW-MSA), and MLP. A special mention can be made for the attention mechanism inside of the W-MSA and the SW-MSA. Self-attention formula is calculated using Equation (3),

$$A = softmax\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V \qquad (3)$$

where $Q, K, V$ are the query, key, and value matrices, and $d_k$ is the dimension of the key vector. The self-attention mechanism allows a model to focus on different parts of the input data when making predictions. Also, residual (skip) connections are applied in each module. The patch expanding layer in the decoder part is utilized to upsample the feature maps. The linear projection layer is performed on these upsampled features in order to generate the pixel-level segmentation. Swin U-Net architecture is shown in Fig.3.

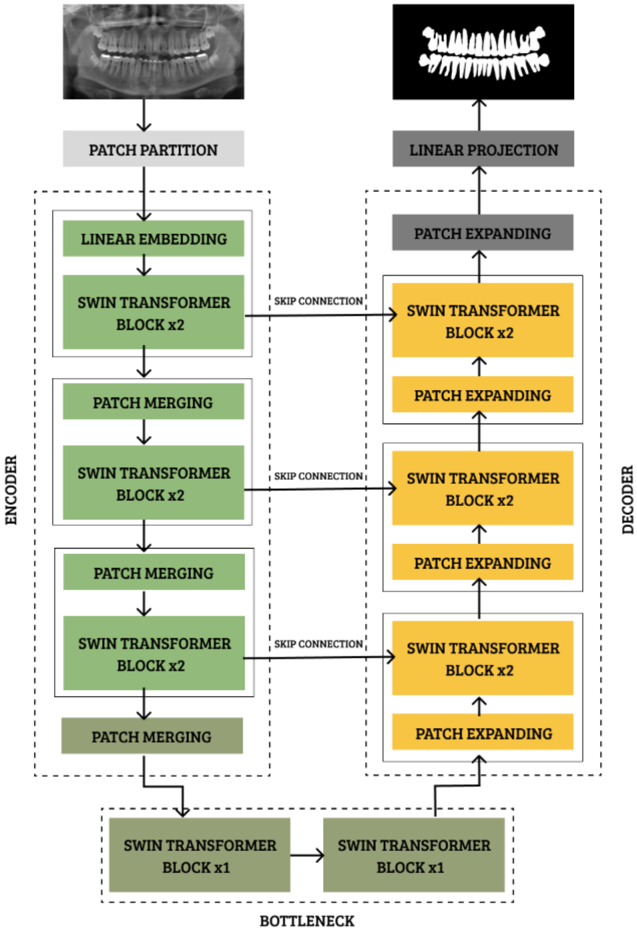

Fig 3. Swin U-Net architecture

## IV. RESULTS

### A. Implementation Details

U-Net, LinkNet and Swin U-Net models trained based on Python 3.8.10 and Pytorch 1.12.1. For U-Net and LinkNet, the input image size is set as 256x256. On the other hand, the input image size for Swin U-Net is 224x224. Moreover, the learning rate, batch size and number of epochs were configured as 1e-4, 8, 100 for all models, respectively. Throughout the training period, Adam was used for optimizer and combination of binary cross-entropy and dice score as utilized for loss function. Our models were trained on four 16GB RAM GPUs, Tesla V100.

### B. Evaluation Criteria

Dice Score (DS) and Intersection over Union (IoU) metrics express the performance of the predicted region in segmentation problems. DS and IoU are often preferred in segmentation problems as they provide sensitive measures of the overlap between the ground truth and the predicted region. DS is calculated two times the intersection between the predicted and ground truth image segmentations divided by the sum of pixels in both images. DS equation presented in Equation (4).

$$DS = \frac{2TP}{2TP + FP + FN} \tag{4}$$

Also, IoU is computed as the intersection of the areas covered by the predicted and ground truth segmentations over the union of these areas. These metrics provide us a quantitative measure of how well the predicted masks align with the ground truth masks. The equation can be seen in Equation (5).

$$IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

### C. Experimental Results

In this study, the experimental studies are presented in two separate sections as Adults Dataset and Children Dataset.

1) Adults Dataset

Table II shows the segmentation performance for adults dataset in terms of DS and IoU by standard deviation values. In the experiments conducted by varying a large number of parameters, all models performed well for the adults dataset. The most successful model is U-Net on ResNest101 backbone by 0.9543 DS and 0.9150 IoU. The model-based comparison shows that the highest performance belongs to U-Net ResNet101 by 0.9543 DS, LinkNet ResNet50 by 0.9542 DS and Swin U-Net T by 0.9529.

TABLE II
TEST RESULTS OF ADULTS DATASET

| Adult Dataset | | |
|---|---|---|
| **Model** | **DS Mean± Std** | **IoU Mean± Std** |
| U-Net ResNet34 | 0.9516 ± 0.0362 | 0.9099 ± 0.0631 |
| U-Net ResNet50 | 0.9536 ± 0.0377 | 0.9137 ± 0.0659 |
| U-Net ResNet101 | **0.9543 ± 0.0374** | **0.9150 ± 0.0654** |
| LinkNet ResNet34 | 0.9491 ± 0.0378 | 0.9055 ± 0.065 |
| LinkNet ResNet50 | 0.9542 ± 0.0350 | 0.9145 ± 0.0613 |
| LinkNet ResNet101 | 0.9515 ± 0.0390 | 0.9100 ± 0.0677 |
| Swin U-Net T | 0.9529 ± 0.0359 | 0.9123 ± 0.0627 |
| Swin U-Net S | 0.9517 ± 0.0378 | 0.9104 ± 0.0658 |

Fig.4. shows a graphical comparison of the adults dataset.

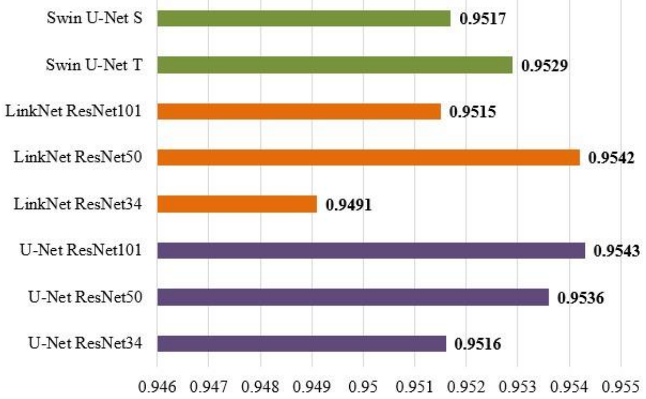

Fig.4. Comparison of DS on adults dataset

### 1) *Children Dataset*

Children dataset segmentation results are presented in Table III by DS, IoU and standard deviation values for eight different models. It has been observed that the highest performance in the children dataset, as in the adults dataset, belongs to U-Net ResNet101 by a DS of 0.9019 and an IoU value of 0.8217. In Fig.5. where the model-based comparison is presented, it is seen that Swin U-Net has the lowest performance and U-Net has the highest performance. The highest DS for Swin U-Net T, LinkNet ResNet50 and U-Net ResNet101 are 0.8189, 0.8914 and 0.9019 respectively.

TABLE III
TEST RESULTS OF CHILDREN DATASET

| Children Dataset | | |
|---|---|---|
| **Model** | **DS Mean± Std** | **IoU Mean± Std** |
| U-Net ResNet34 | 0.8996 ± 0.0127 | 0.8178 ± 0.0208 |
| U-Net ResNet50 | 0.9013 ± 0.0118 | 0.8206 ± 0.0194 |
| U-Net ResNet101 | **0.9019± 0.0123** | **0.8217 ± 0.0203** |
| LinkNet ResNet34 | 0.8889 ± 0.0153 | 0.8004 ± 0.0243 |
| LinkNet ResNet50 | 0.8914 ± 0.0156 | 0.8044 ± 0.0250 |
| LinkNet ResNet101 | 0.8906 ± 0.0145 | 0.8032 ± 0.0232 |
| Swin U-Net T | 0.8189 ± 0.0389 | 0.6951 ± 0.0522 |
| Swin U-Net S | 0.8112 ± 0.0381 | 0.6841 ± 0.0514 |

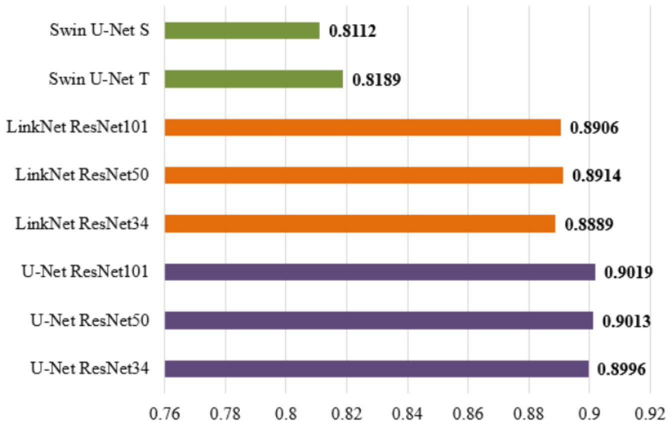

Fig.5. Comparison of DS on children dataset

Fig.6. shows the predicted segmentation results of ground truth and the models on one example each of adult and child PR images.

The experiments were carried out to identify the models and backbones that delivered the most accurate segmentation performance for each dataset. For the adults dataset, the U-Net model with the ResNet101 backbone yielded the best results, while the same U-Net model with the ResNet101 backbone also showed excellent performance on the children dataset. The highest Dice scores recorded were 0.9543 for the adults dataset and 0.9019 for the childrens dataset, highlighting the models' effectiveness in accurately segmenting teeth.

The high accuracy of segmentation in U-Net is ensured by the skip connections between layers so that attributes are not lost and details are preserved. U-Net can be particularly effective when training with limited data because its structure helps to learn general features even in small data sets. This can improve the accuracy of segmentation in small-scale dental images, such as limited dental data. Although LinkNet has deeper networks, it is notable for its more efficient parameterization. This makes it a viable option for better performance on large datasets. In large-scale datasets such as PR, LinkNet processes the data more efficiently and enables fast segmentation. However, LinkNet has not been found to be as effective as U-Net in identifying more complex structures and details, as the structures used are not considered sufficient to capture less fine details. Since Swin is based on the U-Net transformer structure, it is expected to be able to learn local and global relationships better, especially with diversity in the dataset and large-scale images. In PRs, Swin-U-Net's strengths include the relationships between teeth, variability in jaw structure, and the different characteristics of each tooth, which require more contextual information. However, when Swin-U-Net works on smaller datasets, it has been observed that the large number of parameters degrades its performance.

The structure of U-Net does not have many parameters, as it works with an encoder-decoder architecture. However, components such as skip connections and upsampling layers can require high computational power during model training. U-Net generally has a medium level of computational complexity. LinkNet has a lower computational complexity compared to U-Net. This is an important advantage, especially when working on limited hardware. The fact that the model has fewer parameters allows for faster processing time. Swin-U-Net is the model with the highest computational complexity. Transformer-based structures require high computational power, especially for large data sets and high-resolution images. The large number of parameters and more complex computational processes make it necessary to run Swin-U-Net with higher hardware requirements.

## V. Discussions

PR is an important medical imaging tool for the diagnosis of oral diseases. In dental radiology, the automatic segmentation of the teeth structure with the help of PR is important as a first step to improve the performance of the diagnostic treatment process. Automatic analysis of PR images increases the efficiency of dentists in densely populated areas and speeds up the processes of patients waiting for treatment.

By the integration of highly computational hardware into machines, AI-based DL is increasingly preferred for problem solving and decision-making tasks. The advancement of technology has increased the rate of data accumulation, making it easier to access the data that DL algorithms need to be trained. CNN-based and VT-based DL models have become increasingly popular in medical and dental imaging.

Recently, teeth segmentation with PR has been widely used in the literature. Silva et al. [21] obtained an accuracy of 0.9208 with Mask R-CNN on 1500 PR data that they brought to the literature. On the same dataset, Koch et al. [22] achieved 0.936 DS with FCN based U-Net. Zhao et al. [1] validated their Two-Stage Attention Segmentation Network (TSASNet) designed for teeth segmentation with Silva et al. [21] dataset and achieved 0.9272 DS. Hou et al. [39] used 1500 PRs collected by themselves to measure the performance of Teeth U-Net, which they designed with various additional modules between encoder and decoder and at the bottleneck, and obtained a DS of 0.9428. Arora et al. [30] achieved a precision of 0.9501 with a new encoder-decoder architecture based on multimodal feature extraction on 1500 PRs. Ghafoor et al. [33] validated their 540 PRs in their proposed M-Net-like structure with swin transformers teeth attention block cooperating model and obtained a DS of 0.9102. Zhang et al. [34] presented a preliminary study with 1978 and 193 PRs for adults and children, respectively. In the adult data, the DS values of U-Net, R2 U-Net, PSPNet, DeepLab V3+ are 0.9392, 0.9411, 0.9299, 0.9267 respectively.

For children data, the DS values of U-Net, R2 U-Net, PSPNet, DeepLabV3+ are 0.9120, 9027, 0.9083, 0.8961 respectively. Since this study utilizes the datasets presented by Zhang et al. [34] a detailed comparison is presented in Table VI. When Table VI is analyzed, as a result of the experiments, it is observed that the proposed model has an improvement of 1.40% with U-Net ResNet101 for the adults dataset. Unfortunately, the relatively small number of data in the children dataset did not result in a significant increase in segmentation performance results. Due to the low number of data, it was observed that the swin transfomers based approach was inferior in the child dataset compared to the adult dataset. This study is important in terms of evaluating separate models for teeth segmentation for children and adults, comparing CNN and VT based segmentation on small and large datasets, and analyzing a model with respect to backbones of different depths. On the other hand, considering the dental segmentation achievements, the proposed study emphasizes that dental segmentation is suitable for practical application in clinics.

Although teeth segmentation on PR has great potential in clinical settings, some limitations and challenges can be encountered. Since PR usually presents a 2D projection of the teeth and surrounding tissues, lack of depth information and deformations can complicate segmentation processes. Another major challenge is the variety and quality of the data. PR images obtained in the clinical setting can have a wide range of quality depending on the different devices and acquisition techniques. This diversity can complicate the generalization ability of the model and cause preprocessing steps to become more complex to ensure accurate segmentation. Furthermore, the lack of a sufficient number of different patient samples in the datasets can reduce the generalization success of the model and affect the reliability of the results.

When considering teeth segmentation with supervised learning, the labeling part of the data is a major challenge that affects its applicability in a clinical setting. Correct labeling of PR images is a fundamental step in the training of segmentation models; however, this process is time-consuming and labor-intensive. Teeth need to be correctly labeled, the boundaries of each tooth identified, and associated with the appropriate anatomical structures. This is a specialized task and requires a meticulous examination of each image. In clinical settings, the input of dentists or radiologists is often required to perform this labeling process. However, these specialists may not have the time to perform manual labeling for each image. This increases the workload and can make it difficult to efficiently implement tooth segmentation in a clinical setting. Furthermore, labeling errors can also negatively affect the accuracy of the model; mislabeled data can lead to incorrect learning of the model, reducing the reliability of the segmentation results. The clinical use of teeth segmentation can also bring real-time analysis requirements. Fast and accurate results are important, especially in busy clinical environments. However, some DL models can require high computational power and time, which can pose practical challenges.

## VI. Conclusion

In this study, we conducted a detailed comparison between CNN-based architectures, such as U-Net and LinkNet, and transformer-based architectures, specifically Swin U-Net, using dental radiography datasets. These datasets consist of PR images from both adults and children, allowing us to evaluate the models' performance across diverse demographic groups. The evaluation results revealed interesting insights: while CNN-based models achieved superior performance on smaller, limited datasets, transformer-based Swin U-Net was affected by the limited data and did not perform as well on smaller datasets. This highlights the potential advantages of transformer-based models in handling more complex, large-scale datasets, which is a significant consideration in real-world clinical applications. Our findings suggest that teeth segmentation through DL models can provide substantial benefits for dental professionals by offering precise and automated segmentation of teeth in radiographic images. This segmentation can significantly assist in diagnosing dental conditions, planning treatment procedures, and monitoring the progress of treatments over time. By automating this process, dentists can save time, reduce human error, and focus more on patient care rather than manual image analysis. Moreover, the high DS achieved in both adults and children datasets emphasizes the potential of these models in diverse clinical settings, making them a versatile tool for dental practitioners. Furthermore, this research lays the groundwork for future studies in the field of dental image analysis. The results of this study, especially the performance comparison between CNN-based and transformer-based models, can be useful for researchers exploring the integration of DL techniques into dental applications.

The publicly available datasets used in this study also provide an excellent resource for future work in this area, fostering further innovation and improvements in the field of dental radiography analysis. Overall, we believe that the insights gained from this study will contribute significantly to the advancement of automated dental diagnosis and treatment planning.

Future studies are planned to collect larger datasets from individuals with different age groups and ethnic backgrounds. Thus, the performance of the models on various demographic characteristics will be evaluated in more detail and strategies to improve segmentation accuracy can be developed. Such an approach has the potential to provide more effective and generalizable solutions for different patient groups in medical imaging fields such as dental segmentation.

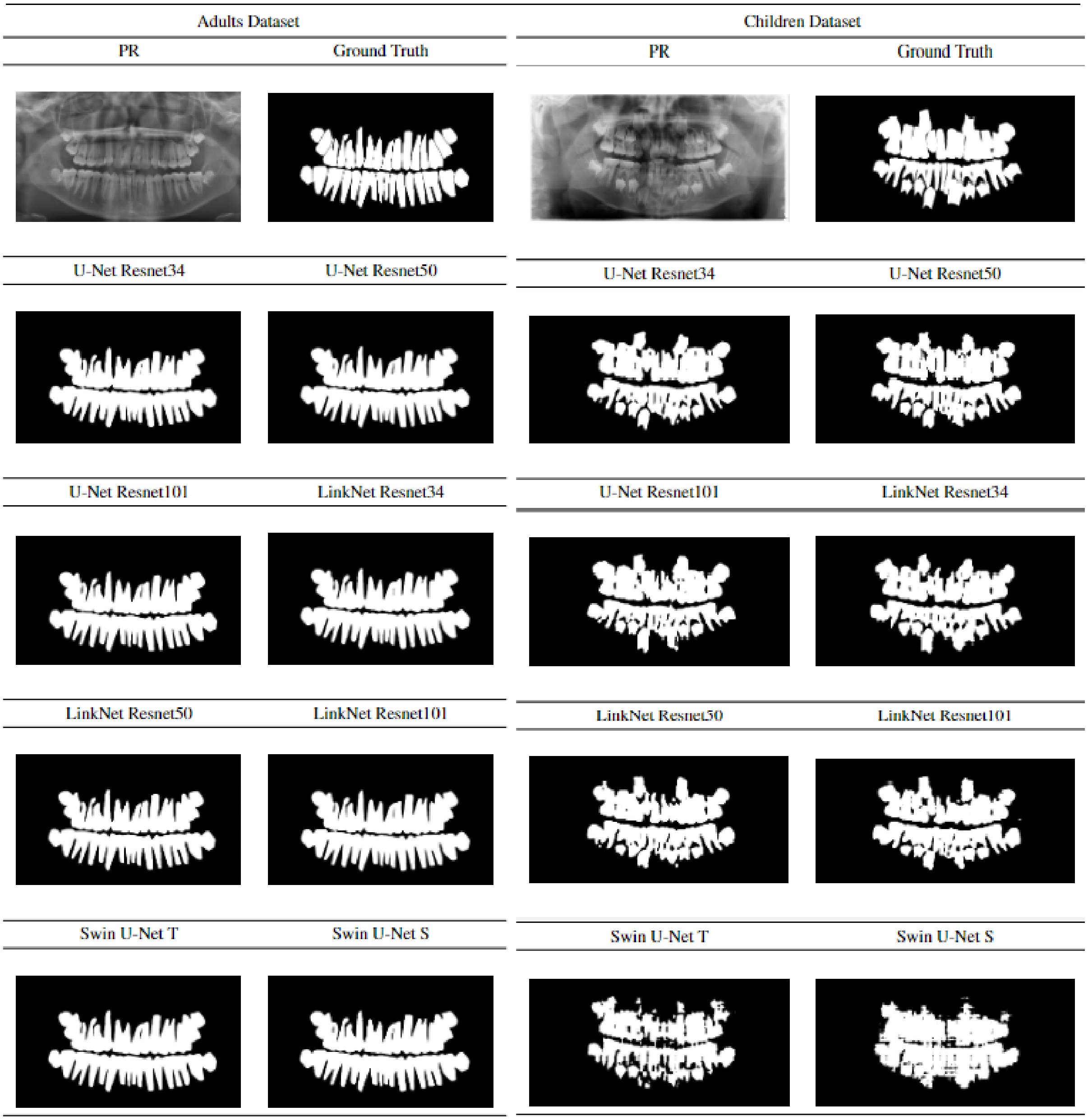

Fig.6 Visualization of the models on adults and children dataset

TABLE VI
COMPARISON OF THIS STUDY AND THE STUDY [34] RESULTS.

| | | Adults | Dataset | Children | Dataset |
|---|---|---|---|---|---|
| | **Models** | **DS** | **IoU** | **DS** | **IoU** |
| **Literature** [34] | U-Net | 0.9392 | 0.8858 | **0.9120** | **0.8387** |
| | R2 U-Net | **0.9411** | **0.8892** | 0.9027 | 0.8247 |
| | PSPNet | 0.9299 | 0.8693 | 0.9083 | 0.8324 |
| | DeepLab V3+ | 0.9267 | 0.8639 | 0.8961 | 0.8121 |
| Proposed Study | U-Net ResNet34 | 0.9516 | 0.9099 | 0.8996 | 0.8178 |
| | U-Net ResNet50 | 0.9536 | 0.9137 | 0.9013 | 0.8206 |
| | U-Net ResNet101 | **0.9543** | **0.9150** | **0.9019** | **0.8217** |
| | LinkNet ResNet34 | 0.9491 | 0.9055 | 0.8889 | 0.8004 |
| | LinkNet ResNet50 | 0.9542 | 0.9145 | 0.8914 | 0.8044 |
| | LinkNet ResNet101 | 0.9515 | 0.9100 | 0.8906 | 0.8032 |
| | Swin U-Net T | 0.9529 | 0.9123 | 0.8189 | 0.6951 |
| | Swin U-Net S | 0.9517 | 0.9104 | 0.8112 | 0.6841 |

## ACKNOWLEDGMENT

This study was conducted in the TÜBİTAK-BİLGEM B3LAB. We would like to express our profound gratitude to Dr. Mehmet HAKLIDIR, Director of the TÜBİTAK BİLGEM Artificial Intelligence Institute for supplying of hardware throughout the work and supporting us to do the research.

## REFERENCES

[1] Y. Zhao *et al.*, “TSASNet: Tooth segmentation on dental panoramic X-ray images by Two-Stage Attention Segmentation Network,” *Knowl Based Syst*, vol. 206, p. 106338, 2020.

[2] A. Haghanifar, M. M. Majdabadi, S. Haghanifar, Y. Choi, and S.-B. Ko, “PaXNet: Tooth segmentation and dental caries detection in panoramic X-ray using ensemble transfer learning and capsule classifier,” *Multimed Tools Appl*, vol. 82, no. 18, pp. 27659–27679, 2023.

[3] Y. Ariji *et al.*, “Automatic detection and classification of radiolucent lesions in the mandible on panoramic radiographs using a deep learning object detection technique,” *Oral Surg Oral Med Oral Pathol Oral Radiol*, vol. 128, no. 4, pp. 424–430, 2019.

[4] M. Fukuda *et al.*, “Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography,” *Oral Radiol*, vol. 36, pp. 337–343, 2020.

[5] B. C. Uzun Saylan *et al.*, “Assessing the effectiveness of artificial intelligence models for detecting alveolar bone loss in periodontal disease: a panoramic radiograph study,” *Diagnostics*, vol. 13, no. 10, p. 1800, 2023.

[6] L. Schneider *et al.*, “Federated vs local vs central deep learning of tooth segmentation on panoramic radiographs,” *J Dent*, vol. 135, p. 104556, 2023.

[7] S. Park *et al.*, “Deep learning-based automatic segmentation of mandible and maxilla in multi-center ct images,” *Applied Sciences*, vol. 12, no. 3, p. 1358, 2022.

[8] N. Kumbasar, M. T. Güller, Ö. Miloğlu, E. A. Oral, and I. Y. Ozbek, “Deep-learning based fusion of spatial relationship classification between mandibular third molar and inferior alveolar nerve using panoramic radiograph images,” *Biomed Signal Process Control*, vol. 100, p. 107059, 2025.

[9] M. T. Güller, N. Kumbasar, and Ö. Miloğlu, “Evaluation of the effectiveness of panoramic radiography in impacted mandibular third molars on deep learning models developed with findings obtained with cone beam computed tomography,” *Oral Radiol*, pp. 1–16, 2024.

[10] A. B. Oktay, Z. Akhtar, and A. Gurses, “Dental biometric systems: a comparative study of conventional descriptors and deep learning-based features,” *Multimed Tools Appl*, vol. 81, no. 20, pp. 28183–28206, 2022.

[11] Ö. Miloğlu, N. Kumbasar, Z. T. Tosun, M. T. Güller, and \.Ibrahim Yücel Özbek, “Gender Classification With Hand-Wrist Radiographs Using the Deep Learning Method,” *Current Research in Dental Sciences*, vol. 35, no. 1, pp. 2–7, 2025.

[12] C.-W. Wang *et al.*, “A benchmark for comparison of dental radiography analysis algorithms,” *Med Image Anal*, vol. 31, pp. 63–76, 2016.

[13] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 2015, pp. 234–241.

[14] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[15] A. Chaurasia and E. Culurciello, “Linknet: Exploiting encoder representations for efficient semantic segmentation,” in *2017 IEEE visual communications and image processing (VCIP)*, 2017, pp. 1–4.

[16] A. Vaswani, “Attention is all you need,” *Adv Neural Inf Process Syst*, 2017.

[17] A. Dosovitskiy, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[18] Z. Liu *et al.*, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[19] J. Chen *et al.*, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.

[20] J. Wu, “Promptunet: Toward interactive medical image segmentation,” *arXiv preprint arXiv:2305.10300*, vol. 2, 2023.

[21] G. Silva, L. Oliveira, and M. Pithon, “Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives,” *Expert Syst Appl*, vol. 107, pp. 15–31, 2018.

[22] T. L. Koch, M. Perslev, C. Igel, and S. S. Brandt, “Accurate segmentation of dental panoramic radiographs with U-Nets,” in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 15–19.

[23] S. Sivagami, P. Chitra, G. S. R. Kailash, and S. R. Muralidharan, “Unet architecture based dental panoramic image segmentation,” in *2020 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 2020, pp. 187–191.

[24] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, and L. Oliveira, “Deep instance segmentation of teeth in panoramic X-ray images,” in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2018, pp. 400–407.

[25] A. Wirtz, S. G. Mirashi, and S. Wesarg, “Automatic teeth segmentation in panoramic X-ray images using a coupled shape model in combination with a neural network,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, 2018, pp. 712–719.

[26] J.-H. Lee, S.-S. Han, Y. H. Kim, C. Lee, and I. Kim, “Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs,” *Oral Surg Oral Med Oral Pathol Oral Radiol*, vol. 129, no. 6, pp. 635–642, 2020.

[27] S. Zhao, Q. Luo, and C. Liu, “Automatic tooth segmentation and classification in dental panoramic X-ray images,” 2020.

[28] B. Silva, L. Pinheiro, L. Oliveira, and M. Pithon, “A study on tooth segmentation and numbering using end-to-end deep neural networks,” in *2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, 2020, pp. 164–171.

[29] C. Sheng *et al.*, “Transformer-based deep learning network for tooth segmentation on panoramic radiographs,” *J Syst Sci Complex*, vol. 36, no. 1, pp. 257–272, 2023.

[30] S. Arora, S. K. Tripathy, R. Gupta, and R. Srivastava, “Exploiting multimodal CNN architecture for automated teeth segmentation on dental panoramic X-ray images,” *Proc Inst Mech Eng H*, vol. 237, no. 3, pp. 395–405, 2023.

[31] M. Kanwal, M. M. Ur Rehman, M. U. Farooq, and D.-K. Chae, “Mask-transformer-based networks for teeth segmentation in panoramic radiographs,” *Bioengineering*, vol. 10, no. 7, p. 843, 2023.

[32] M. K. Dhar, M. Deb, D. Madhab, and Z. Yu, “A Deep Learning Approach to Teeth Segmentation and Orientation from Panoramic X-rays,” *arXiv preprint arXiv:2310.17176*, 2023.

[33] A. Ghafoor, S.-Y. Moon, and B. Lee, “Multiclass Segmentation Using Teeth Attention Modules for Dental X-Ray Images,” *IEEE Access*, vol. 11, pp. 123891–123903, 2023.

[34] Y. Zhang *et al.*, “Children’s dental panoramic radiographs dataset for caries segmentation and dental disease detection,” *Sci Data*, vol. 10, no. 1, p. 380, 2023.

[35] W. Brahmi and I. Jdey, “Automatic tooth instance segmentation and identification from panoramic X-Ray images using deep CNN,” *Multimed Tools Appl*, vol. 83, no. 18, pp. 55565–55585, 2024.

[36] E. Asci *et al.*, “A Deep Learning Approach to Automatic Tooth Caries Segmentation in Panoramic Radiographs of Children in Primary Dentition, Mixed Dentition, and Permanent Dentition,” *Children*, vol. 11, no. 6, p. 690, 2024.

[37] S. Wathore and S. Gorthi, “Bilateral symmetry-based augmentation method for improved tooth segmentation in panoramic X-rays,” *Pattern Recognit Lett*, vol. 188, pp. 1–7, 2025.

[38] G. Altan and A. Al Samar, “Tooth segmentation on dental panoramic X-rays using Mask R-CNN,” in *Mining Biomedical Text, Images and Visual Features for Information Retrieval*, Elsevier, 2025, pp. 481–498.

[39] S. Hou, T. Zhou, Y. Liu, P. Dang, H. Lu, and H. Shi, “Teeth U-Net: A segmentation model of dental panoramic X-ray images for context semantics and contrast enhancement,” *Comput Biol Med*, vol. 152, p. 106296, 2023.

BIOGRAPHIES

**Şilan Fidan Vural** received the bachelor's degree in computer engineering from Yildiz Technical University, Turkey. Vural is a Machine Learning Engineer at Wiro AI, Turkey. Her field of study includes generative AI.

**Nida Kumbasar** received the B.Sc. degree in Computer Engineering Department from Ataturk University, Erzurum, Turkey, in 2015. She received the integrated Ph.D. degree in Electrical and Electronics Engineering Department from Ataturk University, Erzurum, Turkey, in 2024. She completed their PhD as a recipient of the YÖK 100/2000 PhD Scholarship Program, a competitive funding initiative by the Council of Higher Education of Turkey (YÖK) to support doctoral research in priority fields. Throughout this period, she gained professional experience by working in various companies in the private sector within the field of computer science. She is currently a Senior Researcher at TÜBİTAK, Informatics and Information Security Research Center (BİLGEM), Kocaeli, Turkey. Her research interests include medical image processing, remote sensing, data valuation, signal processing, and deep learning.