Journal of Dentistry 156 (2025) 105668

Contents lists available at ScienceDirect

Journal of Dentistry

journal homepage: www.elsevier.com/locate/jdent

Assessment of CNNs, transformers, and hybrid architectures in dental image segmentation

Lisa Schneider [a], Aleksander Krasowski [c], Vinay Pitchika [c], Lisa Bombeck [b], Falk Schwendicke [c,*], Martha Büttner [c]

[a] Department of Oral Diagnostics, Digital Health and Health Services Research, Charité – Universitätsmedizin, Berlin, Germany
[b] Department of Operative, Preventive and Pediatric Dentistry, Charité – Universitätsmedizin, Berlin, Germany
[c] Clinic for Conservative Dentistry and Periodontology, Ludwig-Maximilians-University, Munich, Germany

ARTICLE INFO

Keywords:
Artificial intelligence
Deep learning
Computer vision

ABSTRACT

Objectives: Convolutional Neural Networks (CNNs) have long dominated image analysis in dentistry, reaching remarkable results in a range of different tasks. However, Transformer-based architectures, originally proposed for Natural Language Processing, are also promising for dental image analysis. The present study aimed to compare CNNs with Transformers for different image analysis tasks in dentistry.
Methods: Two CNNs (U-Net, DeepLabV3+), two Hybrids (SwinUNETR, UNETR) and two Transformer-based architectures (TransDeepLab, SwinUnet) were compared on three dental segmentation tasks on different image modalities. Datasets consisted of (1) 1881 panoramic radiographs used for tooth segmentation, (2) 1625 bitewings used for tooth structure segmentation, and (3) 2689 bitewings for caries lesions segmentation. All models were trained and evaluated using 5-fold cross-validation.
Results: CNNs were found to be significantly superior over Hybrids and Transformer-based architectures for all three tasks. (1) Tooth segmentation showed mean±SD F1-Score of 0.89±0.009 for CNNs, 0.86±0.015 for Hybrids and 0.83±0.22 for Transformer-based architectures. (2) In tooth structure segmentation CNNs also outperformed with 0.85±0.008 compared to Hybrids 0.84±0.005 and Transformers 0.83±0.011. (3) Even more pronounced results were found for caries lesions segmentation; 0.49±0.031 for CNNs, 0.39±0.072 for Hybrids and 0.32±0.039 for Transformer-based architectures.
Conclusion: CNNs significantly outperformed Transformer-based architectures and their Hybrids on three segmentation tasks (teeth, tooth structures, caries lesions) on varying dental data modalities (panoramic and bitewing radiographs).
Clinical significance: As deep-learning-based image analysis is part of modern dentistry, practitioners and dental researchers should be aware of strength and limitations of modern model architectures for dental-image analysis. Models that demonstrate optimal performance in other domains do not necessarily constitute the optimal selection for the purpose of dental imaging.

1. Introduction

Convolutional Neural Networks (CNNs) are a popular type of deep learning model architecture, which reached remarkable results in a range of different image-related tasks in dentistry [1] and the broader medical field [2]. Dental tasks range from the segmentation of anatomical structures, such as bone and teeth, as well as tooth structure, which is important for mapping other findings, to clinically relevant pathologies, such as the detection of caries lesions or periodontal bone loss. Segmentation, i.e. outlining pixel clouds affected by a condition or belonging to an anatomical structure, represents a viable approach for CNN-based image analysis, complementing other methodologies such as classification and object detection. Segmentation is particularly advantageous for tasks requiring a high degree of granularity, such as caries diagnosis on bitewing radiographs.

CNNs follow a hierarchical approach comparable to the visual

* Corresponding author at: Clinic for Conservative Dentistry and Periodontology, LMU University Hospital, Ludwig-Maximilians-University Munich, Goethestr. 70, 80336, Munich, Germany.
E-mail address: falk.schwendicke@med.uni-muenchen.de (F. Schwendicke).

https://doi.org/10.1016/j.jdent.2025.105668
Received 20 January 2025; Received in revised form 2 March 2025; Accepted 6 March 2025
Available online 8 March 2025

system of humans, which combines simple structures to recognize more complex ones. Similarly, CNNs interpret images as grids of smaller regions, which are successively scanned pixel-by-pixel by the receptive field of the CNN. First it focuses on finding simple features (e.g., edges, lines, textures), which are later combined to build up more complex patterns or objects (e.g., root canal, tooth, dentin). Learning hierarchies of patterns allows CNNs to efficiently process and interpret images for a range of tasks [3].

Notably, CNNs come with a range of inductive biases related to, for example, the assumption that closely located pixels are more related than those further apart, or to translational invariance, i.e. the detection of objects in an image independent of its location. These biases are on one hand beneficial as they convey basic concepts of images (hierarchies of patterns, local connectivity of pixels, invariance to object location), reducing computation and memory requirements and lowering the amount of data needed for training. On the other hand, inductive biases of CNNs hamper their flexibility: Patterns in data that are not captured under these assumptions may be neglected by the model regardless of their importance. One popular dental example where this behavior may occur is tooth segmentation (and classification) on panoramic radiographs, where pixel-based masks are predicted for each tooth, with this individual tooth then being assigned a class (e.g. along the Universal or FDI scheme). This task is imperative for all automation scenarios, as it facilitates the mapping of other findings (like caries lesions) to teeth and thereby ensuring the clinical benefits of deep learning-based detection of pathologies. For tooth segmentation and classification, the position of a tooth in an image provides important information - also when teeth are far apart. CNNs will not lever this to its full potential but rather rely on the shape of the tooth and potentially neighboring teeth to recognize the class.

While CNNs dominated computer vision for over a decade, they have recently faced strong competition by Transformers, originally a state-of-the-art architecture from Natural Language Processing (NLP) and commonly known as they serve as foundation for chatbots such as ChatGPT (OpenAI, San Francisco, USA). In NLP, it is crucial that information from text data is aggregated from the entire input sequence, also between words that are far apart in a text. Transformers are able to put individual words inside an input sequence into context through a mechanism called self-attention [4]. Notably, Transformers have been adopted for computer vision tasks too, originally under the name Vision Transformer (ViT) [5]. For this, they do not successively scan images, but divide them into small patches and provide these patches simultaneously to the model, which allows the model to capture the whole image at once. Further details can be found in the appendix.

Based on this, Transformer-based architectures may be more flexible and potentially more powerful than CNNs. Notably, Transformers often require large-scale datasets, which are usually scarce in the dental domain [6]. Transformers also come with extensive computational costs and memory consumption. Due to the processing of patches instead of pixels, they may also not be able to capture fine-grained local information.

In short, CNNs and Transformer-based architectures both have strengths and limitations. To combine the strengths of both, while mitigating their limitations, Hybrid architectures between CNNs and Transformers were proposed [7,8]. These are especially promising for medical image analysis as they simultaneously detect fine-grained details and capture global contexts to solve underlying tasks.

Recent studies employed Transformer-based architectures to solve dental use-cases, such as detection of caries and hypomineralization on intraoral photographs or implant positions on cone beam computed tomography [9,10]. However, a comprehensive comparison between Transformers and CNNs as well as Hybrids is missing for dental radiographs. Dental radiographs differ from everyday photographs in terms of contrast and standardization. The efficacy of architectures developed for daily photographs in facilitating image analysis tasks in dentistry, such as diagnosing caries, remains unclear.

In the present study, we aimed to compare two CNN architectures, two Transformer-based architectures and two Hybrid architectures on three different dental image segmentation tasks with varying importance of local and global contexts: (1) A tooth segmentation task on panoramic radiographs based on the FDI notation: It builds the foundation of tooth related image analysis. The significance of global context in this task is high, as the positioning of a tooth in the image is indicative of its identity. (2) A tooth structure segmentation task of bitewing radiographs, which is enhancing the clinical relevance of caries segmentation as it allows the classification of caries depth. (3) A caries lesion segmentation task of bitewing radiographs, the primary task on bitewing analysis where the local image context is relevant once more. We hypothesized that Hybrid architectures would yield superior performances compared with other architectures over all tasks.

2. Material and methods

2.1. Study design

The overview of the study design is represented in Fig. 1. Two CNNs, namely U-Net [11] and DeepLabV3+ [12], two Transformer-based architectures, namely Swin-Unet [13] and TransDeepLab [14] as well as two Hybrids with CNN and Transformer components, namely SwinUNETR [8] and UNETR [7] were trained and compared on three different segmentation tasks. As outlined, the first conducted task was a tooth segmentation task on panoramic radiographs, while the second and third task, a tooth structure segmentation and caries segmentation, respectively, were performed on bitewing radiographs. To provide a fair comparison, an extensive hyperparameter search was conducted for each model architecture on each task. The best hyperparameter configurations were used to train the models on each task. Finally, the results of each network architecture were separately compared and tested for significant differences for all tasks.

2.2. Datasets

Three different datasets were utilized for the underlying study. We deliberately decided to select those datasets and tasks which provide different levels of spatial and positional dependencies within the images to observe strength and weaknesses of the utilized architectures. All data were collected between 2019 and 2020 during routine care até Charité - Universitätsmedizin Berlin with ethical approval (EA4/080/18). Radiographs were recorded with machines from Dürr Dental SE (Bietigheim-Bissingen, Germany) and Sirona Densply Inc. (Bensheim, Germany).

For the tooth segmentation task 1881 panoramic radiographs of patients with a mean±SD age of 44.3 ± 20 years and a sex ratio of 50 % females to males were used. The annotations provided a pixel-wise mask of each individual tooth according to the FDI notation. Thereby, the relative position of each tooth to others plays a crucial role, which may be leveraged by the model architectures for their decision process.

For the tooth structure segmentation task 1625 radiographic bitewings from patients with a mean±SD age of 35.6 ± 15.5 years and a sex ratio of 48 % to 52 % of females and males, respectively, were utilized. Annotations consisted of pixel-wise masks of enamel, dentin, root canal, fillings, and crowns. In this task, local neighborhoods and global, spatial hierarchies play an essential role as the structure of each tooth is consistent.

Finally, for caries segmentation, a dataset of 2689 bitewing radiographs from patients with a mean±SD age of 36.9 ± 13.3 years and a sex ratio of 48 % females and 52 % males was utilized. This task was aimed to be less sensitive to spatial relationships, as caries lesions may be located in varying locations in the radiographs.

For the annotation of the pixel-wise masks for all tasks, one dental expert performed the annotation, and a second dental expert reviewed it regarding its validity and correctness. Training and calibration of the examiners was performed prior to the segmentation. Annotators were

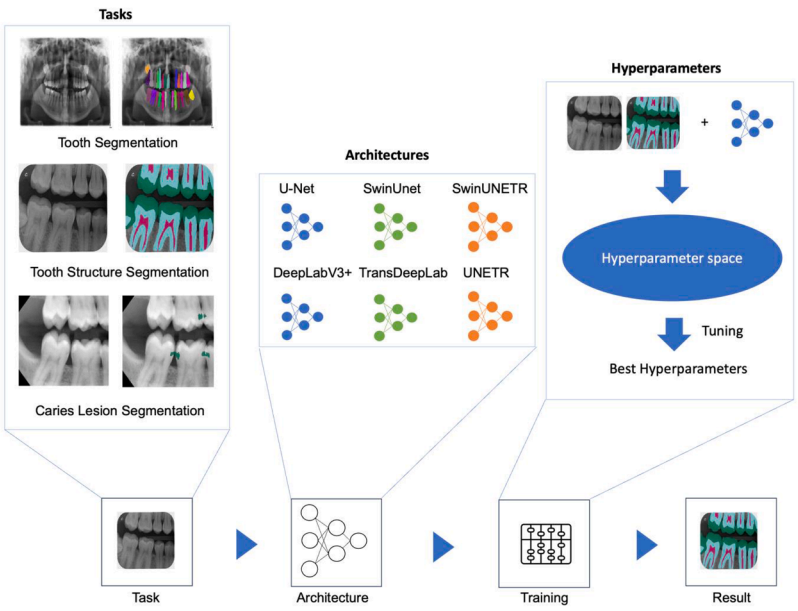

Fig. 1. Overview of study design. Two CNNs (U-Net, DeepLabV3+), two Hybrids (SwinUNETR, UNETR) and two Transformer-based architectures (TransDeepLab, SwinUnet) were compared on three different segmentation tasks of teeth, tooth structures and caries lesions. For the training process of each architecture for each task, hyperparameters were automatically tuned. Note that the images were rescaled to equal height and width for training and testing and are shown in this rescaled format here.

general dentists with more than two years of clinical experience.

All images and masks were downsized to a resolution of 224 × 224 before feeding it to the networks.

2.3. Training

To provide a fair comparison of the architectures, an automatic hyperparameter search was performed to identify appropriate parameters for each model architecture on each task. The detailed description of the hyperparameter space is reported in the Appendix.

Training for all tasks and model architectures was conducted for 300 epochs starting off with pretrained weights from a similar dental task. Training was prematurely stopped upon missing improvement in validation loss for 50 epochs. The objective function employed was a combination of Dice and Focal loss due to its great performance in medical image segmentation [15]. The AdamW optimizer [16] was utilized to steer the learning process of the model, as it has often been utilized for successful training of both CNNs [17,18] and Transformer-based models [19–21]. Further, a cosine learning rate decay was employed, which lowered the initial learning rate over the training period [22]. All other parameters for each architecture and task were taken from the results of the extensive hyperparameter search. The detailed values are reported in Appendix. Training was implemented with the software packages PyTorch 2.0 and MONAI 1.2 and was processed on four NVIDIA A100 40 G GPUs.

2.4. Evaluation and statistical testing

All model performances were evaluated by means of the F1-score, which is the harmonic mean of precision (positive predictive value (PPV)) and sensitivity (recall). Metrics were computed and compared on the independent test sets of each fold. To reach unbiased values, the F1-score was computed from the sum of all true positives, false positives and false negatives [23]. Computed secondary metrics included precision sensitivity and specificity.

The results of CNNs, Hybrids and Transformer-based architectures were compared and tested for statistical significant differences with the Kruskal-Wallis test [24] and a following post-hoc Dunn's test [25]. To address multiple comparisons P-value adjustment using the Benjamini-Hochberg method was performed [26]. The implementation of the statistical testing was performed with statsmodels 0.14, scikit-posthocs 0.7 and SciPy 1.11.

3. Results

The resulting hyperparameters selected by hyperparameter tuning process are reported in the Appendix and were utilized for the training process of the models for the different tasks. The performances reached by the different model architectures on the three segmentation tasks are reported in Fig. 2. For tooth segmentation, the mean±SD F1-Scores were 0.89±0.009 for CNNS, 0.86±0.015 for Hybrids and 0.83±0.22 for Transformer-based architectures. For the tooth structure segmentation, F1-Scores were 0.85±0.008 for CNNS, 0.84±0.005 for Hybrids and 0.83 ±0.011 for Transformers. For caries segmentation F1-Scores were 0.49 ±0.031 for CNNS, 0.39±0.072 for Hybrids and 0.32±0.039 for Transformer-based architectures. Secondary metrics are provided in the appendix in Table S1. CNN architectures were superior over Hybrids and Transformer-based architectures across all tasks with statistical significance. P-values are reported in Table 1.

Fig. 3 shows exemplary segmentation masks predicted by the different architectures for the tooth structure segmentation task. To test our hypothesis, architectures were evaluated within their groups. Statistical differences between the individual architectures were provided in the appendix (Tables S2–S4).

4. Discussion

The present study compared CNNs, Transformer-based architectures and hybrids of these for three exemplary dental tasks, tooth segmentation on panoramic radiographs, tooth structure segmentation and caries segmentation on bitewing radiographs. We hypothesized that Hybrids performed superior over all tasks; we reject this hypothesis based on our results. Instead, CNNs performed superior and outperformed both Hybrids and Transformer-based architectures over the three tasks employed in this study. Our findings raise the assumption that the results from other disciplines cannot simply be transferred to dentistry and

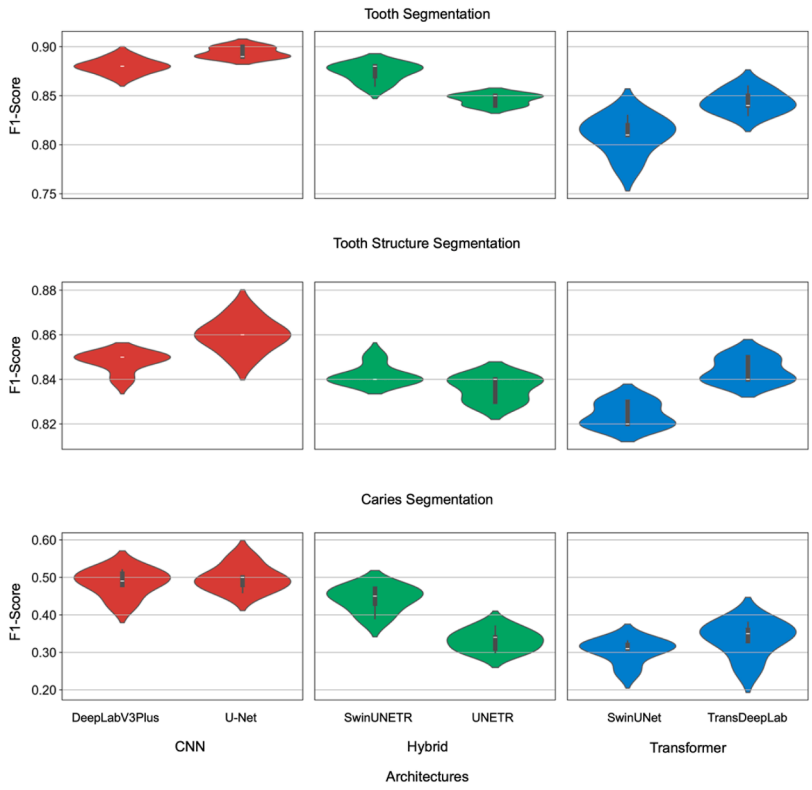

Fig. 2. Violin plot of F1-Scores achieved by the different model architectures for tooth segmentation (top), tooth structure segmentation (mid) and caries lesion segmentation (bottom) over a 5-fold cross-validation. The inner black boxes represent the inner quartile of the distribution with the white line representing the median and the black lines indicating 1.5 times of the inner quartile distribution. Convolutional neural networks (CNN) are illustrated in red, Hybrid architectures in green and Transformer architectures in blue. Note that the y-axis differs between plots for the best illustration for each task.

Table 1
P-values of post-hoc Dunn's test. Multiple comparisons were accounted for with Benjamini-Hochberg p-value correction. Green highlighting indicates statistically significant differences with a significance level of 0.05.

		Hybrids	Transformers
Tooth Segmentation	CNNs	0.023	< 0.001
	Hybrids		0.028
Structure Segmentation	CNNs	0.004	0.001
	Hybrids		0.536
Caries Segmentation	CNNs	0.007	< 0.001
	Hybrids		0.118

require an in-depth discussion.

The inductive biases of CNNs, which provide them with a core understanding of image-related concepts, were likely beneficial to solve the underlying tasks. Processing on pixel-level, concepts of connectivity between neighboring pixels and a strong focus on local contexts may have allowed CNNs to capture the fine-grained details of the class boundaries, which was specifically visible for the tooth structure segmentation task as represented in Fig. 3. Caries lesion detection is a fine-grained task too, as the presence of a caries lesion is determined by a comparison of radiopacity to that of the neighboring pixels or regions. Further, the learning process of CNNs may be useful for tasks such as tooth structure segmentation, where hierarchical structures are directly given by the layers of the tooth anatomy (root canal, dentin and enamel). Further, CNNs most likely benefitted from the initialization with pre-trained model parameters as demonstrated in previous studies for the task of tooth structure segmentation [27,28]. Notably, all three architectures yielded lower performance for caries lesion segmentation than the two other tasks. Humans will similarly find the detection of caries lesions more difficult than segmenting teeth or tooth structures.

Despite its promising performances in other tasks in dentistry, such as caries detection on intraoral photographs, tooth segmentation on dental radiographs or tooth detection on cone-beam computed tomography [9,29–31], we found Transformer-based architectures inferior compared to CNNs and, less so, Hybrids. This may have several reasons. While CNNs process images on pixel level with a local receptive field, Transformers leverage patches of images, as elaborated above. However, for fine-grained segmentation, these patches may not allow to capture the details of the segmentation. Both Transformer-based architectures showed limitations in the fine-grained boundaries of the segmentation classes as visualized in Fig. 3, where the borders of the segmentation masks are coarser than the predictions provided by CNNs or Hybrids. Reducing the patch size to capture features on a smaller scale and provide more fine-grained segmentations is possible but computationally extremely costly, likely limiting its implementation.

Another cause for the inferior performance of Transformer-based models may be our limited dataset size. Transformer-based models are known to be extremely data-hungry as they must learn the core concepts of images from scratch. In the present study sample sizes were limited with 1881 (tooth segmentation), 1625 (tooth structure segmentation) and 2689 (caries segmentation) radiographs – which is, however, not an exception on the lower end, but rather a common dataset size utilized in dentistry: According to a recent systematic review, the median training dataset size used in dentistry was 450 images [1]. We aimed to overcome the limited dataset size by employing a range of methods that reportedly help the Transformer-based architectures to overcome a lack of data. First, all models were initialized by pre-trained parameters to provide a core understanding of image-related concepts from the beginning. In the present study, this may have not been sufficient as pre-training was also performed on a limited dataset size. In other domains, where Transformer-based architectures excelled, pre-training was often performed with extremely large-scale datasets like JFT-300 M (containing

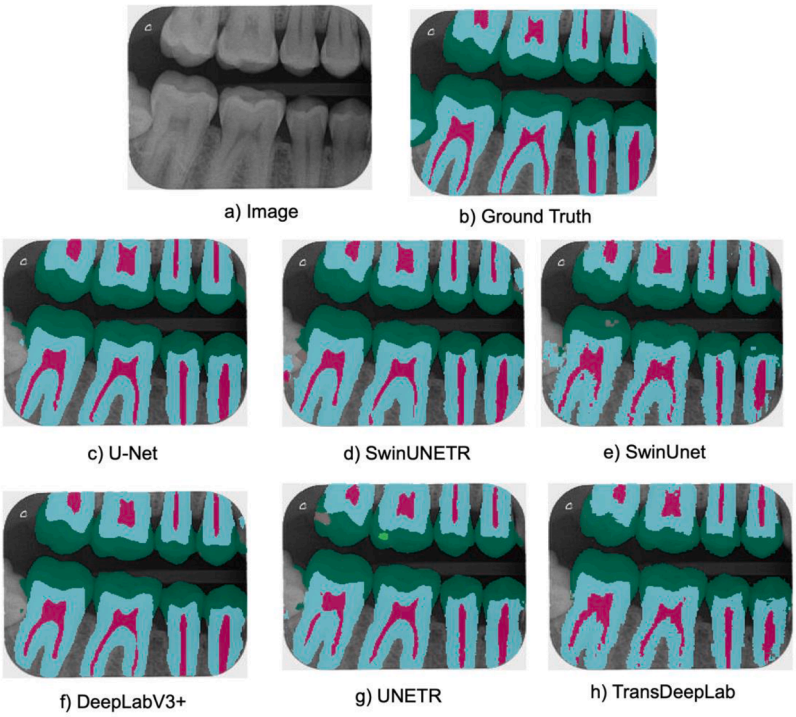

Fig. 3. Bitewing radiograph (a), Ground Truth (b) and predictions of the different model architectures (c-h) for one exemplary sample of the tooth structure segmentation task.

over 300 million labeled images) [32]. Datasets of this scale are, to our knowledge, not available in the medical domain, which may limit the applicability of Transformer-based architectures in this field. Next to pre-training, strong augmentation methods (CutMix, MixUp and RandAugment) were employed in the present study, which have been shown to assist in mitigate the detrimental effects of limited dataset size when using Transformers [33–35]. However, we could not confirm these approaches to overcome the performance disadvantages compared with CNNs. Augmentation techniques have the potential to mimic the variation of real-life images as rotation, saturation or sharpness or visualize objects in different contexts. Radiographs, however, are highly standardized and teeth have always the same context (oral cavity), which might be a reason why augmentation is more beneficial in other domains instead [36,37].

We hypothesized that Hybrids, which combine core components of CNNs and Transformers, lead to superior performance, but failed to accept this hypothesis. We assume that the overall performance of Hybrids may be limited due to the Transformer components, which may face the same challenges as their pure counterparts. Further, Hybrid architectures are relatively new and best practices and techniques for the training process may have not yet been fully optimized.

Based on our findings, for dental use cases with a limited amount of data, time and computation resources available, CNNs may be preferable over the other alternatives, especially for fine-grained segmentations (like segmenting caries lesions). When larger datasets or model parameters from training with large datasets are available, it may be feasible to experiment with Hybrid or Transformer-based architectures to leverage the global context information. As discussed, first approaches to improve training of Transformer-based models on limited data [34,38] are available and current research focuses on enabling fine-grained segmentations with Transformer-based methods, too [39]. The dental community should regularly evaluate the applicability, benefits and detriments of using Transformers combined with such mitigation strategies in case of sparse data.

The setup of our study comes with strengths and limitations. To our knowledge, this is the first holistic evaluation of CNN, Transformer and Hybrid architectures on different tasks in the field of dentistry. The selected tasks included a varying extent of hierarchical and spatial information in their respective dataset, which allowed to explore the performance of the model architectures under these aspects and understand, to some degree, the effects of using different architectures and the reasons behind differential performance. Using 5-fold cross-validation allowed to further gauge the uncertainty around our results, and the conducted hyperparameter search allowed training of each architecture in an appropriate setting. This was specifically essential, as ML research may be biased to configure training in favor of CNNs due to their previous experience. To strengthen this aspect, further methods were applied that specifically have been reported to boost the performance of Transformers. Despite the efforts, we cannot claim generalizability of our finding across other data, tasks or settings. Moreover, and as discussed, limited sample sizes may have influenced our results. Notably, though, it is unlikely that research hubs in dentistry will have millions of labelled images of one type at hand – nor will the technical resources for training models on such scale be regularly available.

5. Conclusion

CNNs significantly outperformed Transformer-based architectures and their Hybrids on three dental segmentation tasks (teeth, tooth structures, caries lesions) on varying dental data modalities (panoramic and bitewing radiographs). The built-in inductive biases of CNNs appear to be beneficial for solving the underlying segmentation tasks, while Transformer-based architectures showed difficulties to provide fine-grained segmentations and likely require (pre-training with) datasets of larger-scale to yield improved results.

CRediT authorship contribution statement

Lisa Schneider: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Aleksander Krasowski:** Writing – review & editing, Resources, Methodology, Data curation, Conceptualization. **Vinay Pitchika:** Writing – review &

editing, Methodology, Formal analysis. **Lisa Bombeck:** Writing – review & editing, Validation, Investigation. **Falk Schwendicke:** Writing – review & editing, Supervision, Resources, Methodology, Data curation. **Martha Büttner:** Writing – review & editing, Visualization, Supervision, Project administration, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Falk Schwendicke is co-founder of the startup dentalXrai GmbH. dentalXrai GmbH did not have any role in conceiving, conducting or reporting this study. The authors are solely responsible for the contents of this paper. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jdent.2025.105668.

References

[1] L.T. Arsiwala-Scheppach, A. Chaurasia, A. Müller, J. Krois, F. Schwendicke, Machine Learning in dentistry: a scoping review, J. Clin. Med. 12 (2023) 937, https://doi.org/10.3390/jcm12030937.

[2] S.K. Zhou, H. Greenspan, C. Davatzikos, J.S. Duncan, B. Van Ginneken, A. Madabhushi, J.L. Prince, D. Rueckert, R.M. Summers, A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises, Proceed. IEEE 109 (2021) 820–838, https://doi.org/10.1109/JPROC.2021.3054390.

[3] I. Goodfellow, A. Courville, Y. Bengio, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. ukasz Kaiser, I. Polosukhin, Attention is all you need. Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. accessed August 26, 2024.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv PreprintarXiv:2010.11929 (2020).

[6] F. Schwendicke, J. Krois, Data dentistry: how Data are changing clinical care and research, J. Dent. Res. 101 (2022) 21–29, https://doi.org/10.1177/00220345211020265.

[7] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, UNETR: transformers for 3D medical image segmentation, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1748–1758, https://doi.org/10.1109/WACV51458.2022.00181.

[8] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H.R. Roth, D. Xu, Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images, in: A. Crimi, S. Bakas (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham, 2022, pp. 272–284, https://doi.org/10.1007/978-3-031-08999-2_22.

[9] M. Felsch, O. Meyer, A. Schlickenrieder, P. Engels, J. Schönewolf, F. Zöllner, R. Heinrich-Weltzien, M. Hesenius, R. Hickel, V. Gruhn, et al., Detection and localization of caries and hypomineralization on dental photographs with a vision transformer model, NPJ. Digit. Med. 6 (2023) 198.

[10] X. Yang, X. Li, X. Li, P. Wu, L. Shen, Y. Deng, ImplantFormer: vision transformer-based implant position regression using dental CBCT data, Neur. Comput. Appl. 36 (2024) 6643–6658, https://doi.org/10.1007/s00521-023-09411-1.

[11] O. Ronneberger, T.Brox P.Fischer, U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2015, pp. 234–241. http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.

[13] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.

[14] R. Azad, M. Heidari, M. Shariatnia, E.K. Aghdam, S. Karimijafarbigloo, E. Adeli, D. Merhof, Transdeeplab: convolution-free transformer-based deeplab v3+ for medical image segmentation. International Workshop on PRedictive Intelligence In MEdicine, Springer, 2022, pp. 91–102.

[15] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, Y. Pan, Rethinking dice loss for medical image segmentation, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 851–860.

[16] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv PreprintarXiv:1711.05101 (2017).

[17] Z. Xue, J. Liang, G. Song, Z. Zong, L. Chen, Y. Liu, P. Luo, Large-batch optimization for dense visual predictions: training faster R-CNN in 4.2 minutes, Adv. Neural. Inf. Process. Syst. 35 (2022) 18694–18706.

[18] R.R. Shivwanshi, N. Nirala, Implementation of an advanced lung nodule classification system using optimized ConvMixer and AdamW-based CNN architecture, in: 2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2023, pp. 54–59, https://doi.org/10.23919/SPA59660.2023.10274463.

[19] Y. Bai, J. Mei, A.L. Yuille, C. Xie, Are transformers more robust than cnns? Adv. Neural. Inf. Process. Syst. 34 (2021) 26831–26843.

[20] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F.E. Tay, J. Feng, S. Yan, Tokens-to-token vit: training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.

[21] X. Chen, C.-J. Hsieh, B. Gong, When vision transformers outperform resnets without pre-training or strong data augmentations, arXiv PreprintarXiv:2106.01548 (2021).

[22] I. Loshchilov, F. Hutter, SGDR: Stochastic Gradient Descent with Warm Restarts, 2017, https://doi.org/10.48550/arXiv.1608.03983.

[23] L. Schneider, P. Dave, L. Arsiwala-Scheppach, F. Schwendicke, J. Krois, Exploring bias in F-score computation methods of multi-class segmentation models, in: Proceedings of the 2021 5th International Conference on Video and Image Processing, Association for Computing Machinery, New York, NY, USA, 2022, pp. 76–84, https://doi.org/10.1145/3511176.3511189.

[24] P.E. McKight, J. Najab, Kruskal-wallis test. The Corsini Encyclopedia of Psychology, 2010, p. 1. –1.

[25] O.J. Dunn, Multiple comparisons using rank sums, Technometrics 6 (1964) 241–252.

[26] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. Roy. Statist. Soc.: Ser. B (Methodolog.) 57 (1995) 289–300.

[27] A. Ke, W. Ellsworth, O. Banerjee, A.Y. Ng, P. Rajpurkar, CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-ray interpretation, in: Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 116–124, https://doi.org/10.1145/3450439.3451867.

[28] L. Schneider, L. Arsiwala-Scheppach, J. Krois, H. Meyer-Lueckel, K.K. Bressem, S. M. Niehues, F. Schwendicke, Benchmarking deep learning models for tooth structure segmentation, J. Dent. Res. 101 (2022) 1343–1349, https://doi.org/10.1177/00220345221100169.

[29] Y. Li, G. Zeng, Y. Zhang, J. Wang, Q. Jin, L. Sun, Q. Zhang, Q. Lian, G. Qian, N. Xia, et al., Agmb-transformer: anatomy-guided multi-branch transformer network for automated evaluation of root canal therapy, IEEE J. Biomed. Health Inform. 26 (2021) 1684–1695.

[30] S. Gao, X. Li, X. Li, Z. Li, Y. Deng, Transformer based tooth classification from cone-beam computed tomography for dental charting, Comput. Biol. Med. 148 (2022) 105880.

[31] C. Sheng, L. Wang, Z. Huang, T. Wang, Y. Guo, W. Hou, L. Xu, J. Wang, X. Yan, Transformer-based deep learning network for tooth segmentation on panoramic radiographs, J. Syst Sci. Compl. 36 (2023) 257–272.

[32] C. Sun, A. Shrivastava, S. Singh, A. Gupta, Revisiting unreasonable effectiveness of data in deep learning era, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 843–852.

[33] E.D. Cubuk, B. Zoph, J. Shlens, Q.V. Le, Randaugment: practical data augmentation with no separate search, arXiv PreprintarXiv:1909.13719 2 (2019) 7.

[34] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6023–6032.

[35] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: beyond empirical risk minimization, arXiv PreprintarXiv:1710.09412 (2017).

[36] A. Ke, W. Ellsworth, O. Banerjee, A.Y. Ng, P. Rajpurkar, CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-ray interpretation, in: Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 116–124.

[37] L. Schneider, L. Arsiwala, J. Krois, H. Meyer-Lueckel, K. Bressem, S. Niehues, F. Schwendicke, Benchmarking deep learning models for tooth structure segmentation, J. Dent. Res. (2022) 002203452211001, https://doi.org/10.1177/00220345221100169.

[38] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond Empirical Risk Minimization, 2018, https://doi.org/10.48550/arXiv.1710.09412.

[39] Y. Yi, Y. Jiang, B. Zhou, N. Zhang, J. Dai, X. Huang, Q. Zeng, W. Zhou, C2FTFNet: coarse-to-fine transformer network for joint optic disc and cup segmentation, Comput. Biol. Med. 164 (2023) 107215, https://doi.org/10.1016/j.compbiomed.2023.107215.