Check for updates

Journal of Dental Research
1–7
© International Association for Dental Research and American Association for Dental, Oral, and Craniofacial Research 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00220345221100169
journals.sagepub.com/home/jdr

Benchmarking Deep Learning Models for Tooth Structure Segmentation

L. Schneider[1,2], L. Arsiwala-Scheppach[1,2], J. Krois[1,2], H. Meyer-Lueckel[3], K.K. Bressem[4,5], S.M. Niehues[4], and F. Schwendicke[1,2]

Abstract

A wide range of deep learning (DL) architectures with varying depths are available, with developers usually choosing one or a few of them for their specific task in a nonsystematic way. Benchmarking (i.e., the systematic comparison of state-of-the art architectures on a specific task) may provide guidance in the model development process and may allow developers to make better decisions. However, comprehensive benchmarking has not been performed in dentistry yet. We aimed to benchmark a range of architecture designs for 1 specific, exemplary case: tooth structure segmentation on dental bitewing radiographs. We built 72 models for tooth structure (enamel, dentin, pulp, fillings, crowns) segmentation by combining 6 different DL network architectures (U-Net, U-Net++, Feature Pyramid Networks, LinkNet, Pyramid Scene Parsing Network, Mask Attention Network) with 12 encoders from 3 different encoder families (ResNet, VGG, DenseNet) of varying depth (e.g., VGG13, VGG16, VGG19). On each model design, 3 initialization strategies (ImageNet, CheXpert, random initialization) were applied, resulting overall into 216 trained models, which were trained up to 200 epochs with the Adam optimizer (learning rate = 0.0001) and a batch size of 32. Our data set consisted of 1,625 human-annotated dental bitewing radiographs. We used a 5-fold cross-validation scheme and quantified model performances primarily by the F1-score. Initialization with ImageNet or CheXpert weights significantly outperformed random initialization ($P < 0.05$). Deeper and more complex models did not necessarily perform better than less complex alternatives. VGG-based models were more robust across model configurations, while more complex models (e.g., from the ResNet family) achieved peak performances. In conclusion, initializing models with pretrained weights may be recommended when training models for dental radiographic analysis. Less complex model architectures may be competitive alternatives if computational resources and training time are restricting factors. Models developed and found superior on nondental data sets may not show this behavior for dental domain-specific tasks.

Keywords: computer vision, artificial intelligence, segmentation, tooth structures, transfer learning, neural networks

Introduction

Deep learning (DL) has been widely employed for image analytics in dermatology (skin photographs) (Jafari et al. 2016), ophthalmology (retina imagery) (Son et al. 2020), or pathology (histological specimens) (Kather et al. 2019). Also in dentistry, DL classification models have been employed to predict the modality of radiographs (Cejudo et al. 2021), the presence of caries lesions (Lee et al. 2018), periodontal bone loss (Krois et al. 2019), and apical lesions (Ekert et al. 2019) on dental radiographs. DL segmentation models, which perform a classification task at the pixel level, were used for the segmentation of anatomical structures in panoramic images (Cha et al. 2021), apical lesions on cone beam computed tomography scans (Orhan et al. 2020), periodontal bone loss on panoramic radiographs (Kim et al. 2019), and caries lesions on bitewings (Cantu et al. 2020).

Recent guidelines in the field call for rigorous and comprehensive planning, conducting, and reporting of DL studies in dentistry (Schwendicke et al. 2021). One key element in those guidelines is a hypothesis-driven selection of the DL model configuration, which includes, among others, its architecture, its complexity, and the initialization strategy for the model weights (e.g., via transfer learning). (1) **Architecture:** The basic unit of an artificial neural network is a neuron, which is a nonlinear mathematical model inspired by the biological neuron (McCulloch and Pitts 1943). These units are stacked to build layers that are connected via mathematical operations

[1]Department of Oral Diagnostics, Digital Health and Health Services Research, Charité–Universitätsmedizin, Berlin, Germany
[2]ITU/WHO Focus Group on AI for Health, Topic Group Dental Diagnostics and Digital Dentistry, Geneva, Switzerland
[3]Department of Restorative, Preventive and Pediatric Dentistry, Zahnmedizinische Kliniken der Universität Bern, University of Bern, Bern, Switzerland
[4]Charité–Universitätsmedizin Berlin, Klinik für Radiologie, Berlin, Germany
[5]Berlin Institute of Health at Charité–Universitätsmedizin Berlin, Berlin, Germany

A supplemental appendix to this article is available online.

Corresponding Author:
F. Schwendicke, Department of Oral Diagnostics, Digital Health and Health Services Research, Charité–Universitätsmedizin Berlin, Aßmannshauser Str. 4-6, Berlin, 14197, Germany.
Email: falk.schwendicke@charite.de

source: https://doi.org/10.48350/170584 | downloaded: 12.10.2024

with other layers of neurons. The arrangement of these layers and operations defines the model architecture. Model architectures such as ResNet (He et al. 2016) or VGG (Simonyan and Zisserman 2015) are widely used in the field of machine learning. For image segmentation, specialized layers extend the basic model architectures, which in such a setting are referred to as backbone. This allows one to plug in different backbones and benchmark them for image segmentation tasks. (2) **Complexity:** Most model architectures are available in different degrees of complexities, which reflects the depth of the neural network (i.e., the number of layers included and the number of neurons and connections between them). Deeper models are more complex as they consist of more parameters (i.e., connections between neurons). (3) **Initialization:** The connections between neurons and layers of neurons, which are also referred to as model weights, are basically digits that correspond to the strength of the connection. During model training, these weights are adjusted to find a set of values that are most suitable to solve the underlying task. Starting with a predefined setting of these weights enhances the efficiency of the training process and improves model convergence. Using a predefined setting of weights that stem from a previously trained neural network provides a meaningful starting point for the training process. This technique is referred to as transfer learning (Tan et al. 2018).

The sheer number of possible configurations of model architecture, including backbones, complexity, and initialization strategies, impedes systematic and comprehensive comparisons of existing study findings (Schwendicke et al. 2019). One strategy to overcome this issue is to perform benchmarking, which involves the systematic comparison of different model architectures and model configurations on an identical data set. Such benchmarking studies provide guidance for researchers in the model design process, which improves research efficiency by enabling the development of high-performing models in a shorter time at lower development costs. However, in the medical domain and, more so, dentistry, benchmarking initiatives are scarce, owing to limited data availability and high costs for establishing solid and accepted ground truth labels and annotations. To cope with these difficulties, the ITU/WHO Focus Group Artificial Intelligence for Health (FG-AI4H) is developing a standard evaluation process and benchmarking framework for artificial intelligence (AI) models in health. The present study will inform this initiative.

In a recent benchmarking study, Bressem et al. (2020) benchmarked 16 different model architectures for classification tasks on 2 openly available chest radiograph data sets: CheXpert (Irvin et al. 2019) and the COVID-19 Image Data Collection. They showed that complex and deep models do not necessary outperform simpler architectures. Similarly, Ke et al. (2021) addressed the assumption that model architectures that perform better on the ImageNet data set (Deng et al. 2009), a popular open-source benchmark data set containing millions of labeled images, also generally perform better on CheXpert. This assumption was not found to be valid based on the comparison of 16 convolutional architectures on 5 classification tasks.

In the present study, we aim to expand the studies of Bressem et al. (2020) and Ke et al. (2021) to a dental segmentation task. We benchmarked 216 DL models defined by their architecture, complexity, and initialization strategy. We evaluated these model configurations for a specific dental task: tooth structure (enamel, dentin, pulpal cavity, fillings, and crowns) segmentation on dental bitewing radiographs. We deliberately decided to use this application since first, there is evidence that segmentation models perform well on this task (Ronneberger et al. 2015a) and, second, there is less ambiguity about the establishment of the ground truth for this task, with tooth structures being easily discriminated even by nonsenior clinicians. We expect our results to inform dental researchers about suitable model configurations for their experiments and aim to contribute to evidence-guided DL model selection in dental research.

Materials and Methods

Benchmarking Tasks

This analysis is based on a segmentation task for tooth structures on dental bitewing radiographs. Several model development aspects were benchmarked. (1) **Architecture:** First, we assessed different DL model architectures, since to date, most neural networks have mainly been benchmarked on openly available data sets such as ImageNet. However, it is not yet determined whether the best-performing networks on ImageNet will also perform best for dental radiographic images. Hence, we benchmarked architectures such as U-Net (Ronneberger et al. 2015b), U-Net++ (Zhou et al. 2018), Feature Pyramid Networks (FPN) (Kirillov et al. 2019), LinkNet (Chaurasia and Culurciello 2017), Pyramid Scene Parsing Network (PSPNet) (Zhao et al. 2017), and Mask Attention Network (MAnet) (Fan et al. 2020), among others. These networks were selected, as they all allow to employ the same established backbones of varying depths of model layers (ResNet50 [He et al. 2016], VGG13 [Simonyan and Zisserman 2015], DenseNet121 [Huang et al. 2017]). The depth of the encoder is conventionally represented by the digits behind the name of the architecture (e.g., ResNet18, ResNet34). All model implementations were taken from the same software package (Yakubovskiy 2020). (2) **Complexity:** Second, we investigated the model performances emanating from model complexity. Supposedly, deeper DL models, which have more trainable parameters, outperform shallower alternatives if enough data and computational resources are available. However, deeper models are more likely to overfit training data, and model convergence may not be reached. Furthermore, limited computational resources imply restrictions regarding image resolution or batch size; both may negatively affect the model performance. (3) **Initialization:** Third, we analyzed different initialization strategies, such as random weights initialization or initialization based on pretrained weights from the ImageNet as well as the CheXpert data set. The latter strategies are referred to as transfer learning. Thereby, features learned on large, open data sets are directly transferred to a new task and hence do not have to be learned from scratch. This technique speeds up

model convergence and improves model performance. Initialization with ImageNet is one of the most popular transfer learning strategies. Even for tasks on medical radiographs, transferring knowledge from models trained on ImageNet yields a boost in performance (Ke et al. 2021). However, the feature space learned on ImageNet differs fundamentally from medical features of radiographs. ImageNet consists of natural RGB color images that are classified into more than 20,000 classes, while radiographic images contain grayscale images and are usually classified in only a few categories. Hence, an initialization with pretrained models on radiographic images such as the CheXpert data set (Irvin et al. 2019) may potentially be more suitable for medical segmentation tasks of, for instance, dental radiographs.

Ethics Statement

This study was ethically approved by the ethics committee of the Charité (EA4/102/14 and EA4/080/18).

Study Design

In the present study, 72 models were built from a combination of varying architectures and encoder backbones and were each trained with 3 different initialization strategies on a tooth structure segmentation task. Each model was trained with 5-fold cross-validation with varying train, validation, and test sets for each fold. Hence, for each model run, the data were randomly split into training, validation, and test data with proportions of 60% (3 folds), 20% (1 fold), and 20% (1 fold), respectively. We additionally applied a sensitivity analysis and assessed model performances on underrepresented classes (in our case, fillings and crowns), as in real life, medical data set class imbalance is likely the rule and not the exception. Reporting of this study follows the Standards for Reporting Diagnostic Accuracy guideline (STARD) (Bossuyt et al. 2015) and the Checklist for Artificial Intelligence in Dental Research (Schwendicke et al. 2021).

Performance Metrics

Model performances were primarily quantified by the F1-score, which captures the harmonic mean of recall (specificity) and precision (positive predictive value [PPV]). F1-scores are computed from the sum of true positives, false positives, and false negatives over all channels of segmentation masks and cross-validation folds. This method was described by Forman and Scholz (2010) and results in unbiased F-scores in cross-validation schemes. Secondary metrics were accuracy, sensitivity, precision, and intersection of union (IoU). Based on the distribution of the results, the median was chosen as a descriptive statistic.

Data Set, Sample Size, and Reference Test

The available data set consisted of 1,625 dental bitewing radiographs with a maximum of 8 to 9 teeth per image and is described in detail in the Appendix. Tooth structures visible on bitewing radiographs (namely, enamel, dentin, the pulp cavity, and nonnatural "structures" like fillings and crowns) were annotated in a pixel-wise fashion (as masks) by 1 dental expert. These masks represent the ground truth for each data sample. In a second iteration, those annotations were reviewed by another dental expert for validity and correctness. Each annotator independently assessed each image using an in-house custom-built annotation tool described in Ekert et al. (2019). All examiners were calibrated and advised on how to perform the segmentation. Images with implants, bridges, or root canal fillings were very rare (<1%) and therefore excluded.

Notably, enamel, dentin, and pulpal areas were present in every radiograph, while fillings and crowns were only available in 80% and 20% of images, respectively. Images and segmentation masks were resized to a resolution of 224 × 224 to provide a fixed input size of the images as required by the model architectures.

Models and Training

As represented in Figure 1, models were built by combining different model architectures (U-Net, U-Net++, FPN, LinkNet, PSPNet, MAnet) with backbones from 3 different families (ResNet, VGG, DenseNet) of different depths (ResNet18, ResNet34, ResNet50, ResNet101, ResNet152, VGG13, VGG16, VGG19, DenseNet121, DenseNet161, DenseNet169, DenseNet201). This led to a total of 72 model designs, which were each initialized with 3 different strategies (random, ImageNet, CheXpert), resulting into 216 trained models in total. All models were trained under a 5-fold cross-validation scheme, where the combination of samples in training, validation, and test set was varied for each fold to achieve a reasonable estimate of the model performance independent from the data split. Details on training are described in the Appendix.

Statistical Analysis

Model configurations with respect to initialization strategies and architectures were ranked according to their median F1-score and formally tested for differences between configurations with the nonparametric Wilcoxon rank-sum test. The nonparametric Spearman's rank-order correlation was estimated to determine the relationship between complexity and model performance (F1-score). To account for multiple comparisons, we adjusted the P values using the Benjamini–Hochberg method (Benjamini and Hochberg 1995). P values below 0.05 were considered statistically significant. The number of pairwise comparisons C of conditions k was computed via equation (1).

$$C = \frac{k(k-1)}{2} \tag{1}$$

Results

Figure 2 presents an overview of segmentation outputs generated by different model architectures in comparison to the ground truth. Figure 3 shows the F1-scores of different model

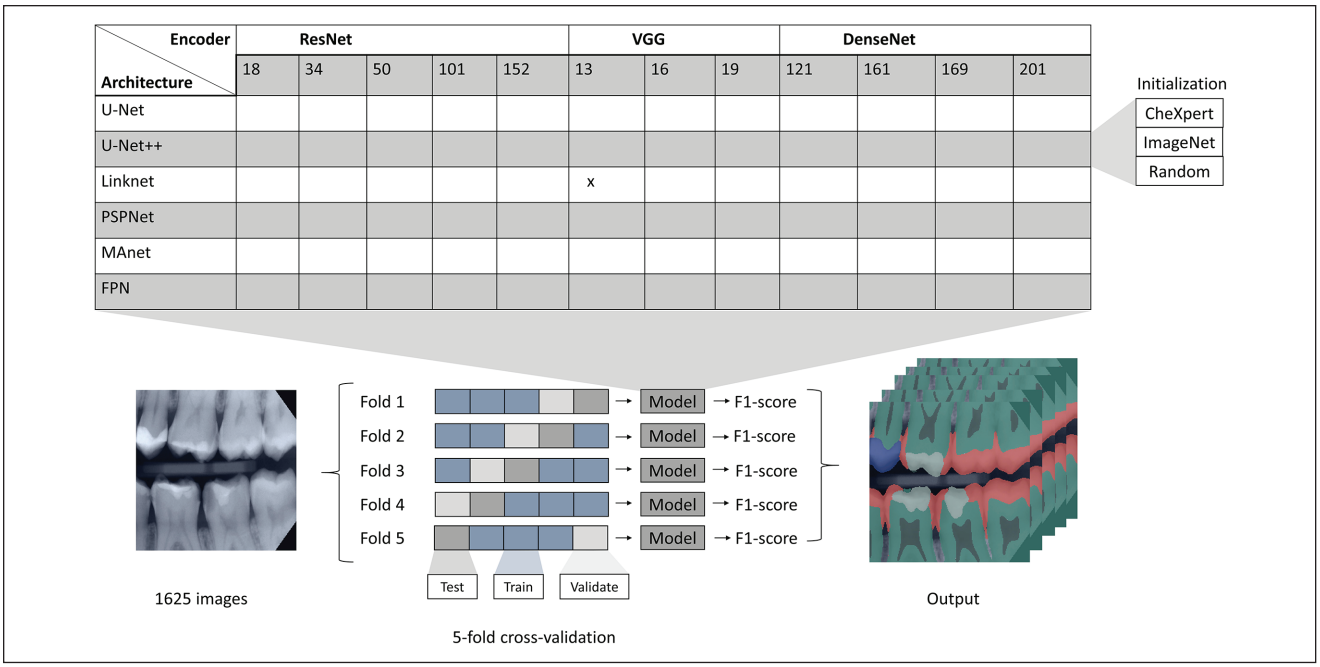

Figure 1. Illustration of the study design. Model setups were based on different architectures, encoder backbones, and initialization strategies (top) and 5-fold cross-validation with varying train, validation, and test sets for each fold (bottom). Exemplary bitewing radiograph (left) and tooth structure components overlaid on an input image (right).

configurations grouped by architecture, backbone family, and initialization strategy.

(1) **Architecture:** Out of 15 pairwise comparisons of model architectures, 14 turned out to be statistically significantly different. U-Net++, U-Net, and LinkNet achieved a median (interquartile range [IQR]) F1-score of 0.86 (0.85, 0.87), (0.84, 0.86), and (0.85, 0.88), respectively, and outperformed MAnet, PSPNet, and FPN with statistical significance. Backbones from the VGG and DenseNet group reached a median (IQR) of 0.85 (0.83, 0.86) and (0.81, 0.86), respectively, while the ResNet group reached a median (IQR) F1-score of 0.84 (0.81, 0.86). Models with backbones from the VGG group outperformed models with backbones of the ResNet group with statistical significance.

(2) **Complexity:** We found a statistically significant weak positive monotonic relationship between the network size and its performance with $r = 0.32$ ($P < 0.001$).

(3) **Initialization:** Different initialization strategies computed over all architectures and backbones achieved F1-scores of 0.86 (0.83, 0.87) (ImageNet), 0.86 (0.83, 0.87) (CheXpert), and 0.83 (0.77, 0.84) (random initialization). Models initialized with ImageNet or CheXpert outperformed models initialized with random weights ($P_{\text{ImageNet}} < 0.001$, $P_{\text{CheXpert}} < 0.001$). No significant difference was observed between ImageNet and CheXpert ($P = 0.85$).

(4) **Class imbalances:** In a sensitivity analysis, the model performance was evaluated on the minority classes of filling (80%) and crown (20%). In general, models' performance was inversely related to class frequencies (Fig. 4).

(4.1) **Architecture:** Models based on a VGG backbone outperformed models with a ResNet backbone on the minority classes of filling ($P = 0.009$) and crown ($P = 0.013$). Notably, there was no statistical difference between the 3 backbones on the majority classes of pulpal cavity and dentin.

(4.2) **Complexity:** We found a statistically significant weak positive monotonic relationship between the network size and its performance for class dentin ($r = 0.245$, $P < 0.001$), enamel ($r = 0.239$, $P < 0.001$), filling ($r = 0.195$, $P = 0.004$), pulpa ($r = 0.218$, $P < 0.001$), and class crown ($r = 0.154$, $P < 0.023$).

(4.3) **Initialization:** Models with ImageNet and CheXpert initialization consistently outperformed models with random initialization. There was no statistically significant difference between ImageNet and CheXpert initializations.

Discussion

We benchmarked 216 models defined by their architecture, complexity, and initialization strategy on a tooth structure segmentation task of dental bitewing radiographs. Several findings require a more detailed discussion.

First, we aimed to evaluate whether there are superior model architectures for the tooth segmentation task at hand. We discovered a performance advantage of models with backbones from the VGG family over models with backbones from the ResNet family. Our findings are consistent with those from Ke et al. (2021), who reported that architecture improvements reported on ImageNet may not always be translated to performances on medical imaging tasks. New model architectures and model improvements seem to be prone to overfitting on ImageNet data sets. Hence, transferability of newest AI research results into other domains, here the dental domain, may not be guaranteed.

The statistically significant performance advantage of models with VGG encoder backbones plead for the usage of VGG encoders, when solid baseline models are required, which perform reasonably well across different model configurations and settings. This may be relevant for the implementation of proof of concepts, for example. The top 10 performing models on the tooth structure segmentation task were built with backbones from the ResNet and DenseNet family. Consequently, if the focus is on model performance, it seems warranted to invest time to find an optimal model configuration based on more complex models (e.g., from the ResNet family). If, however, the validation of general concepts or benchmarking is the focus of the study, VGG-based models seem a reasonable choice as they are more robust across model configurations.

Second, one of our objectives evolved around the effect of the model complexity on the model performance. One of the key findings was a weak positive relationship between model depth and model performance. Therefore, we accept our hypothesis. Notably, however, the number of parameters increased in large steps, with only incremental improvements of model performance. Hence, the performance improvement was oftentimes disproportionate to the increasing demands for computational resources, training time, or the need to reduce image resolutions. The largest network in the present study was MAnet combined with a ResNet152 backbone, which reached an F1-score of 0.85 (0.85, 0.85) over all folds (ImageNet initialization). LinkNet in combination with a ResNet50 backbone was 5 times smaller but reached an F-score of 0.88 (0.88, 0.88) in comparison. It should be highlighted that lower computational costs allow for input imagery of higher resolution, which may be relevant for many dental applications.

Our third objective, aimed to give insights whether initializing with ImageNet or CheXpert, is consistently superior even when there is a difference in performance between both initialization strategies. We found statistically significant performance boosts for models initialized with ImageNet or CheXpert weights in comparison to a random initialization. These findings are consistent with those from Ke et al. (2021), who reported that 12 of 16 architectures benefited from an initialization with ImageNet weights for a classification task of chest radiographs. The comparison of ImageNet and CheXpert initialization showed no significant differences.

Fourth, we additionally found predictions on the minority class of filling (80%) to be generally more stable over different model configurations than predictions on class crowns (20%). Our results showed that there are superior architectures for segmenting minority classes (e.g., U-Net, U-Net++, LinkNet), but choosing a reasonable architecture may not be sufficient to

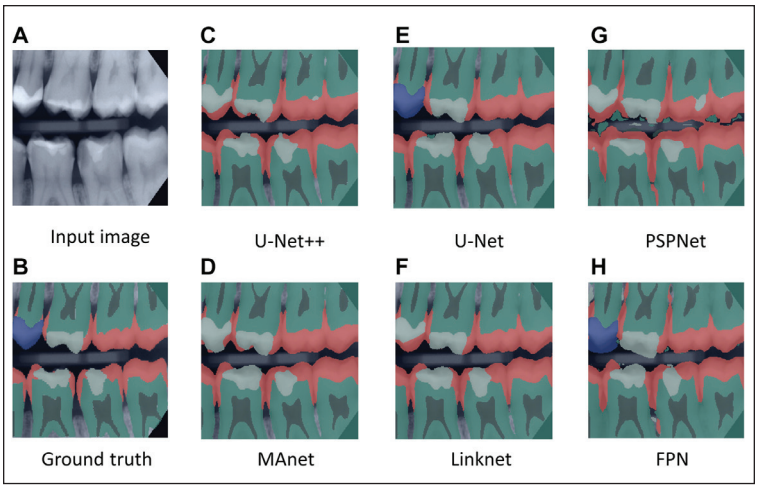

Figure 2. Examples of segmented bitewing radiographs. (**A**) Naive input image. (**B**) Ground truth and (**C–H**) output of tooth structure segmentation by different model architectures. The red, dark green, light green, gray, and blue colors indicate enamel, pulp cavity and root canals, dentin, filling, and crown classes, respectively. All models in this example were built with a ResNet50 backbone and initialized with pretrained CheXpert weights. This figure is available in color online.

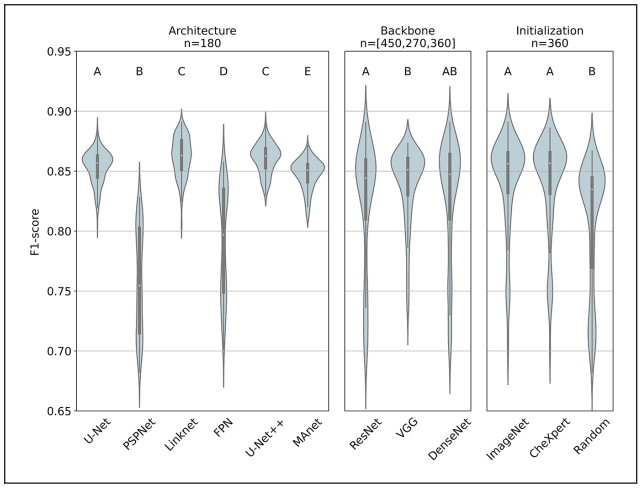

Figure 3. F1-scores stratified by initialization strategy, architecture, and backbone family based on sample sizes *n*. Median, interquartile range, and 95% confidence interval are represented by the white dot, the black box, and the black line, respectively. Different superscript letters indicate statistically significant difference (e.g., between U-Net and LinkNet), while the same superscript letters represent no significant difference (e.g., between LinkNet and U-Net++) (see Appendix for more details).

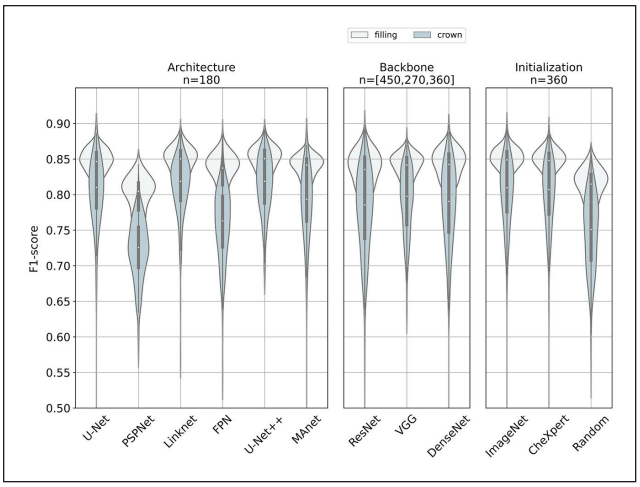

Figure 4. F1-scores of different models in the minority classes, filling (white) and crown (steel blue), respectively. We stratified the analyses by initialization strategy, architecture, and backbone family. Median, interquartile range, and 95% confidence interval are represented by the white dot, the black box, and the black line, respectively. Results are based on a sample size *n*. This figure is available in color online.

overcome class imbalance. Hence, it could be recommended to address this problem with weighted loss functions (Guerrero-Penã et al. 2018) or oversampling (Buda et al. 2018).

This study comes with several limitations. First, our results were based on 1 specific DL task, a tooth structure segmentation on bitewing radiographs, and are limited to the examined model architectures. Hence, we do not claim generalizability of our findings across other segmentation tasks or over all existing model architectures. Second, images of our data set originate from varying machines, which may lead to different behavior of the models. Furthermore, radiographs with bridges, implants, and root canal fillings were not considered in the present study as they were very rare. We accept this as our aim was to benchmark models and not to build clinically useful ones in this study. In line with this, we were only aiming at a model comparison instead of proposing a high-precision model. Hence, we did not take any actions against the existing class imbalance and did not perform an extensive hyperparameter search. Finally, we based our analysis of the relationship between model performances and model complexity exclusively on the number of model parameters. It may be the case that model architectures with more parameters require less computational power through more efficient structures of layers. Furthermore, we did not evaluate the effect of minor differences in performance within the dental environment or how computational resources are affected by differences in the number of parameters of the models.

Conclusion

We benchmarked different configurations of DL models based on their architecture, backbone, and initialization strategy regarding their performance on a tooth structure segmentation task of dental bitewing radiographs to provide guidance for researchers in their DL model selection process. Regarding the superiority of certain model architectures, we found that VGG backbones provided solid baseline models across different model configurations, while peak performances were reached through combinations of U-Net++, LinkNet, and ResNet or DenseNet encoders. Superior architectures did not overcome class imbalance. Models known to perform better than others on a nondental data set like ImageNet did not demonstrate such superiority on our dental imaging task. The analysis of the relationship between model complexity and performance showed that deeper models did not necessarily perform better than shallow alternatives with lower demands in computational resources. Finally, we found that transfer learning boosts model performance, independent of the origin of transferred knowledge.

Author Contributions

L. Schneider, contributed to conception, design, data analysis, and interpretation, drafted and critically revised the manuscript; L. Arsiwala-Scheppach, contributed to analysis, critically revised the manuscript; J. Krois, contributed to conception, design, and data analysis, drafted and critically revised the manuscript; H. Meyer-Lueckel, contributed to interpretation, critically revised the manuscript; K.K. Bressem, contributed to acquisition and interpretation, critically revised the manuscript; S.M. Niehues, contributed to acquisition, critically revised the manuscript; F. Schwendicke, contributed to conception, design, data acquisition, and interpretation, drafted and critically revised the manuscript. All authors gave final approval and agree to be accountable for all aspects of the work.

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: F. Schwendicke and J. Krois are cofounders of the dentalXrai Ltd., a startup. dentalXrai Ltd. did not have any role in conceiving, conducting, or reporting this study.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

L. Schneider https://orcid.org/0000-0002-4431-2669
L. T. Arsiwala https://orcid.org/0000-0002-1428-6543
J. Krois https://orcid.org/0000-0002-6010-8940
F. Schwendicke https://orcid.org/0000-0003-1223-1669

References

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Method. 57(1):289–300.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, De Vet HC, et al. 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin Chem. 61(12):1446–1452.

Bressem KK, Adams LC, Erxleben C, Hamm B, Niehues SM, Vahldiek JL. 2020. Comparing different deep learning architectures for classification of chest radiographs. Sci Rep. 10(1):13590.

Buda M, Maki A, Mazurowski MA. 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 106:249–259.

Cantu AG, Gehrung S, Krois J, Chaurasia A, Rossi JG, Gaudin R, Elhennawy K, Schwendicke F. 2020. Detecting caries lesions of different radiographic extension on bitewings using deep learning. J Dent. 100:103425.

Cejudo JE, Chaurasia A, Feldberg B, Krois J, Schwendicke F. 2021. Classification of dental radiographs using deep learning. J Clin Med. 10(7):1496.

Cha JY, Yoon HI, Yeo IS, Huh KH, Han JS. 2021. Panoptic segmentation on panoramic radiographs: deep learning–based segmentation of various structures including maxillary sinus and mandibular canal. J Clin Med. 10(12):2577.

Chaurasia A, Culurciello E. 2017. Linknet: exploiting encoder representations for efficient semantic segmentation. In: IEEE Visual Communications and Image Processing (VCIP); December 10–13, 2017; St. Petersburg, FL. doi:10.1109/vcip.2017.8305148

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition; June 20–25, 2009; Miami, FL. p. 248–255. doi:10.1109/CVPR.2009.5206848

Ekert T, Krois J, Meinhold L, Elhennawy K, Emara R, Golla T, Schwendicke F. 2019. Deep learning for the radiographic detection of apical lesions. J Endod. 45(7):917–922.

Fan T, Wang G, Li Y, Wang H. 2020. Ma-net: a multi-scale attention network for liver and tumor segmentation. IEEE Access. 8:179656–179665.

Forman G, Scholz M. 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explor Newsl. 12(1):49–57.

Guerrero-Penã FA, Marrero Fernandez PD, Ing Ren T, Yui M, Rothenberg E, Cunha A. 2018. Multiclass weighted loss for instance segmentation of cluttered cells. In: 25th IEEE International Conference on Image Processing (ICIP); October 7–10, 2018; Athens, Greece. p. 2451–2455. doi:10.1109/ICIP.2018.8451187

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27–30, 2016; Las Vegas, NV. p. 770–778. doi:10.1109/CVPR.2016.90

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. 2017. Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21–26, 2017; Honolulu, HI. p. 2261–2269. doi:10.1109/CVPR.2017.243

Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, et al. 2019. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc Conf AAAI Artif Intell. 33(1):590–597.

Jafari MH, Karimi N, Nasr-Esfahani E, Samavi S, Soroushmehr SMR, Ward K, Najarian K. 2016. Skin lesion segmentation in clinical images using deep learning. Presented at: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE; Cancún, Mexico. p. 337–342. doi: 10.1109/ICPR.2016.7899656.

Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, Marx A, Boor P, Tacke F, Neumann UP, Grabsch HI, Yoshikawa T, Brenner H, Chang-Claude J, Hoffmeister M, Trautwein C, Luedde T. 2019. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med. 25(7):1054–1056.

Ke A, Ellsworth W, Banerjee O, Ng AY, Rajpurkar P. 2021. CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-ray interpretation. In: Proceedings of the Conference on Health, Inference, and Learning; April 8, 2021. New York (NY): Association for Computing Machinery. p. 116–124. doi:10.1145/3450439.3451867

Kim J, Lee HS, Song IS, Jung KH. 2019. DeNTNet: deep neural transfer network for the detection of periodontal bone loss using panoramic dental radiographs. Sci Rep. 9(1):1–9.

Kirillov A, Girshick R, He K, Dollár P. 2019. Panoptic feature pyramid networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 6392–6401. doi:10.1109/CVPR. 2019.00656. https://arxiv.org/abs/1901.02446.

Krois J, Ekert T, Meinhold L, Golla T, Kharbot B, Wittemeier A, Dörfer C, Schwendicke F. 2019. Deep learning for the radiographic detection of periodontal bone loss. Sci Rep. 9(1):1–6.

Lee JH, Kim DH, Jeong SN, Choi SH. 2018. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent. 77:106–111.

McCulloch WS, Pitts W. 1943. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 5(4):115–133.

Orhan K, Bayrakdar I, Ezhov M, Kravtsov A, Ozyürek T. 2020. Evaluation of artificial intelligence for detecting periapical pathosis on cone-beam computed tomography scans. Int Endod J. 53(5):680–689.

Ronneberger O, Fischer P, Brox T. 2015a. Dental X-ray image segmentation using a U-shaped deep convolutional network. In: International Symposium on Biomedical Imaging; Brooklyn, NY. Vol. 1. p. 3. http://www-o.ntust.edu.tw/~cweiwang/ISBI2015/challenge2/isbi2015_Ronneberger.pdf

Ronneberger O, Fischer P, Brox T. 2015b. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical Image Computing and Computer-Assisted Intervention (MICCAI); Lecture Notes in Computer Science. Vol. 9351. Cham (Switzerland): Springer. p. 234–241.

Schwendicke F, Golla T, Dreher M, Krois J. 2019. Convolutional neural networks for dental image diagnostics: a scoping review. J Dent. 91:103226.

Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, Uribe S, Krois J; IADR e-Oral Health Network and the ITU WHO Focus Group AI for Health. 2021. Artificial intelligence in dental research: checklist for authors, reviewers, readers. J Dent. 107:103610.

Simonyan K, Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. arXiv preprint. https://arxiv.org/abs/1409.1556.

Son J, Shin JY, Kim HD, Jung KH, Park KH, Park SJ. 2020. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. Ophthalmology. 127(1):85–94.

Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. 2018. A survey on deep transfer learning. In: Kurková V, Manolopoulos Y, Hammer B, Iliadis L, Maglogiannis I, editors. Artificial Neural Networks and Machine Learning–ICANN 2018. Cham (Switzerland): Springer International Publishing. p. 270–279.

Yakubovskiy P. 2020. Segmentation models pytorch. pytorch [accessed 2021 April 22]. https://segmentation-modelspytorch.readthedocs.io/en/latest/.

Zhao H, Shi J, Qi X, Wang X, Jia J. 2017. Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21–26, 2017; Honolulu, HI. p. 6230–6239. doi:10.1109/CVPR.2017.660

Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. 2018. Unet++: a nested U-net architecture for medical image segmentation. In: Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T, Martel A, Maier-Hein L, Tavares JMR, Bradley A, Papa JP, Belagiannis V, et al., editors. Deep learning in medical image analysis and multimodal learning for clinical decision support. Cham (Switzerland): Springer International. p. 3–11. doi:10.1007/978-3-030-00889-51.