Journal of Dentistry 135 (2023) 104556

Contents lists available at ScienceDirect

Journal of Dentistry

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

Federated vs Local vs Central Deep Learning of Tooth Segmentation on Panoramic Radiographs

Lisa Schneider [a,b], Roman Rischke [c], Joachim Krois [b], Aleksander Krasowski [a,b], Martha Büttner [a,b], Hossein Mohammad-Rahimi [b,d], Akhilanand Chaurasia [b,e], Nielsen S Pereira [b,f], Jae-Hong Lee [b,g], Sergio E. Uribe [b,h,i,j], Shahriar Shahab [b,k], Revan Birke Koca-Ünsal [b,l], Gürkan Ünsal [b,m], Yolanda Martinez-Beneyto [n], Janet Brinz [b,o], Olga Tryfonos [b,p], Falk Schwendicke [a,b,*]

[a] Department of Oral Diagnostics, Digital Health, and Health Services Research, Charité — University Medicine Berlin, Berlin, Germany
[b] ITU/WHO Focus Group on AI for Health, Topic Group Dental Diagnostics and Digital Dentistry, Geneva, Switzerland
[c] Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[d] Shahid Beheshti University of Medical Sciences, Tehran, Iran Dental school, Iran
[e] Department of Oral Medicine and Radiology, Faculty of Dental Sciences, King George's Medical University, Lucknow, India
[f] Private Practice in Oral and Maxillofacial Radiology, Rio de Janeiro, Brazil
[g] Department of Periodontology, College of Dentistry and Institute of Oral Bioscience, Jeonbuk National University, Jeonju, Korea
[h] Department of Conservative Dentistry Oral Health, Riga Stradins University, Riga, Latvia
[i] School of Dentistry, Universidad Austral de Chile, Valdivia, Chile
[j] Baltic Biomaterials Centre of Excellence, Headquarters at Riga Technical University, Riga, Latvia
[k] Department of Oral and Maxillofacial Radiology, School of Dentistry, Shahed University of Medical Sciences, Tehran, Iran
[l] Department of Periodontology, Faculty of Dentistry, University of Kyrenia, Kyrenia, Cyprus
[m] Department of Dentomaxillofacial Radiology, Faculty of Dentistry, Near East University, Nicosia, Cyprus
[n] Department of Preventive Dentistry, University of Murcia, Murcia, Spain
[o] Department of Conservative Dentistry and Periodontology, University Hospital, LMU Munich, Munich, Germany
[p] Department of Periodontology and Oral Biochemistry, Academic Centre for Dentistry Amsterdam, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Artificial intelligence
Big data
Computer vision
Deep learning
Informatics
Mathematical models

ABSTRACT

Objective: Federated Learning (FL) enables collaborative training of artificial intelligence (AI) models from multiple data sources without directly sharing data. Due to the large amount of sensitive data in dentistry, FL may be particularly relevant for oral and dental research and applications. This study, for the first time, employed FL for a dental task, automated tooth segmentation on panoramic radiographs.
Methods: We employed a dataset of 4,177 panoramic radiographs collected from nine different centers ($n = 143$ to $n = 1881$ per center) across the globe and used FL to train a machine learning model for tooth segmentation. FL performance was compared against Local Learning (LL), i.e., training models on isolated data from each center (assuming data sharing not to be an option). Further, the performance gap to Central Learning (CL), i.e., training on centrally pooled data (based on data sharing agreements) was quantified. Generalizability of models was evaluated on a pooled test dataset from all centers.
Results: For 8 out of 9 centers, FL outperformed LL with statistical significance ($p<0.05$); only the center providing the largest amount of data FL did not have such an advantage. For generalizability, FL outperformed LL across all centers. CL surpassed both FL and LL for performance and generalizability.
Conclusion: If data pooling (for CL) is not feasible, FL is shown to be a useful alternative to train performant and, more importantly, generalizable deep learning models in dentistry, where data protection barriers are high.
Clinical Significance: This study proves the validity and utility of FL in the field of dentistry, which encourages researchers to adopt this method to improve the generalizability of dental AI models and ease their transition to the clinical environment.

* Corresponding author at: Charité – Universitätsmedizin Berlin, Department of Oral Diagnostics, Digital Health and Health Services Research, Charité - Universitätsmedizin Berlin, Germany, Aßmannshauser Str. 4-6, 14197 Berlin, Germany.
E-mail address: falk.schwendicke@charite.de (F. Schwendicke).

https://doi.org/10.1016/j.jdent.2023.104556
Received 12 April 2023; Received in revised form 16 May 2023; Accepted 17 May 2023
Available online 18 May 2023

1. Introduction

Artificial Intelligence (AI) has shown great potential to transform dentistry; analyzing a wealth of dental data with AI and using it to support diagnostics, treatment planning and actual treatment has been demonstrated to be feasible across all dental disciplines [1].

Most dental AI employs machine learning, where mathematical models are utilized to identify the inherent structure of a training dataset to allow inference (prediction) on unseen test data. Usually, this involves labeling of data by experts e.g., the classification of an image as showing caries lesions, or detecting the location of a certain pathology on an image etc. [2].

The translation of developed AI models from the research stage into the clinical environment, however, remains slow. Despite a wealth of studies, only a few products have successfully passed regulatory hurdles and entered routine care [3]. The main barrier for this is grounded in the poor generalizability of many AI models. As models are typically trained and tested using data from one center, recorded with one technique, methodology, and represent a single population.

An AI application rarely performs similarly well if applied on data from other centers, gathered using other technical setups, representing different populations, which often differ in age, gender, socio-demographic characteristics, or oral health status. [4] Collaborative efforts (e.g., gathering data from multiple centers) may help to overcome generalizability issues and also allow smaller or less experienced research groups to participate in state-of-the-art AI research. However, such efforts are limited by privacy constraints, which lead to difficulties in exchanging particularly dental data as it is oftentimes hard to de-identify [5].

Federated Learning (FL) is a learning paradigm which enables collaborative, data-driven research between multiple centers through a privacy-by-design approach. It avoids critical exchanges of sensitive data between centers and instead relies on sharing abstract model parameters, which essentially carry the knowledge learned from this data. FL was originally aimed at parallelized training on edge devices and smartphones but has caught considerable attention in healthcare [6–9] mainly as it may assist to overcome privacy limitations and allow to train generalizable models. However, dental research on FL is still limited [10].

In the present study, we aimed to assess FL for tooth segmentation on panoramic radiographs, a specific (and exemplary) task in dental image analysis. Tooth segmentation involves labeling pixels belonging to each tooth on a panoramic, which allows to identify, classify and relate further findings (e.g., a caries or apical lesion) of an AI-based analysis to a specific tooth. It was further useful for the present study, as tooth segmentation can be relatively easily performed by humans (who label the radiographs before using them for training) and can hence be standardized across centers, reducing the effect of center-specific labeling on the outcomes of FL. We used radiographs from nine international centers and compared FL against Local Learning (LL, involving training on isolated data of each center) and Central Learning (CL, involving data pooling, e.g., under the assumption of data sharing agreements being in place). We also tested models for their generalizability across centers. Our hypothesis was that FL significantly improves the performance and generalizability in comparison with LL (i.e., when CL training is not feasible due to privacy regulations). We further investigated whether specific centers benefited particularly from FL given their specific data distribution.

2. Materials and methods

2.1. Study design

In this study, neural networks (see below) were employed to solve a multi-class tooth segmentation task on panoramic radiographs. Training was conducted with three different learning paradigms: LL, CL and FL. The resulting models were evaluated and compared in terms of performance (on their own local test dataset) and generalizability (on the combined test dataset, i.e., including data from all participating centers). The contribution of each center to FL varied considerably due to disparate data shares. In order to assess the effect of each center's contribution, FL was analyzed further using equal contributions from each center in a sensitivity analysis.

2.2. Data

The available datasets were collected by nine different centers from across the globe as part of the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) Initiative [11], namely (1) Charité – Universitätsmedizin Berlin, Berlin, Germany (Charité), (2) University of Murcia, Murcia, Spain (MU), (3) King George's Medical University, Lucknow, India (KGMU), (4) Wonkwang University College of Dentistry, Daejeon, Korea (WU), (5) Private Practice Dr. Nielsen, Rio de Janeiro, Brazil (PPN), (6) University of Kyrenia, Kyrenia, Cyprus (RBKU), (7) Shahid Beheshti University of medical sciences, Tehran, Iran (SBMU), (8) Shahed University, Tehran, Iran (SU) and (9) Private Practice Dr. Uribe, Valdivia, Chile (PP). Details on ethical approval and data protection considerations can be found in the Appendix. Each center

Table 1
The metadata provided by the nine different centers. For some centers, not all metadata was available.

ID	Country	City	Sample size (Share on overall data)	Female (%)	Age in years (SD)	Number of teeth (SD)	Device
Charité	Germany	Berlin	1881 (45.0%)	0.50	44.3 (20.0)	29 (3)	Sirona XG3D
MU	Spain	Murcia	252 (6.0%)	0.47	53.3 (14.9)	26 (5)	Vatech PAX-400C
KGMU	India	Lucknow	317 (7.6%)	0.30	45.1 (11.9)	29 (5)	Planmeca ProMax
WU	Korea	Daejeon	294 (7.0%)	0.60	46.7 (13.9)	28 (2)	Vatech PCH-2500
PPN	Brazil	Rio de Janeiro	324 (7.8%)	0.66	44.3 (19.6)	27 (5)	Kodak K9000C 3D
RBKU	Cyprus	Kyrenia	337 (8.0%)	0.52	– (–)	28 (4)	Sirona SL
SBMU	Iran	Tehran	374 (9.0%)	0.56	33.7 (18.0)	28 (5)	Planmeca Dimax 3
SU	Iran	Tehran	255 (6.1%)	0.45	44.2 (23.0)	29 (4)	n/a
PPU	Chile	Valdivia	143 (3.4%)	0.57	35.5 (18.0)	29 (3)	Sirona SL

n/a not available.

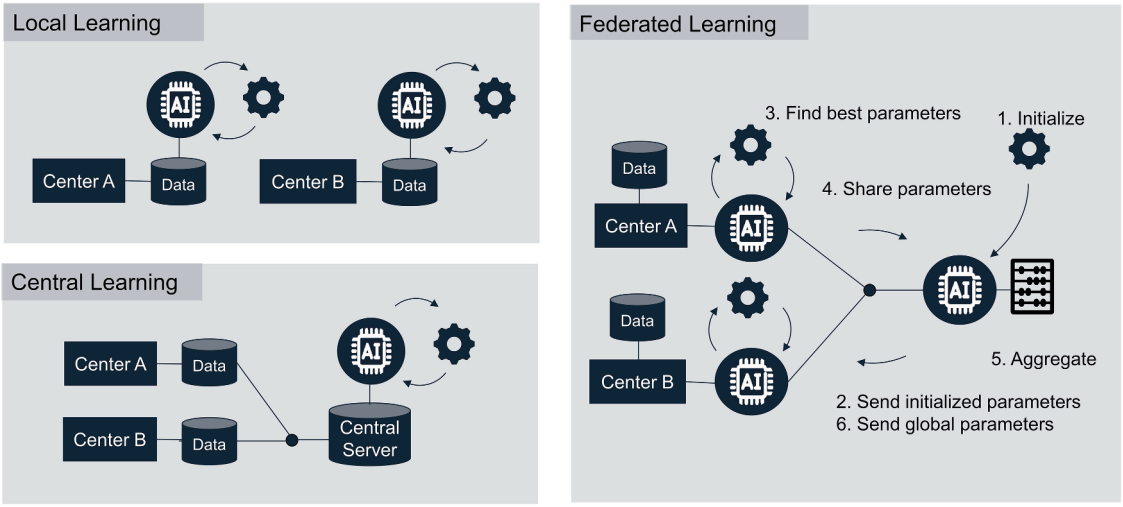


Fig. 1. Differences in Local, Central and Federated Learning. In Local Learning each center trains on its own data only. In Central Learning training data is shared and a global model is trained on the combined dataset. In Federated Learning training is done on local data and the respective computed model weights are aggregated at a server without exchanging the data itself.

provided a convenient sample from already existing radiographs from routine care. The exclusion criteria encompassed: edentulous patients, primary teeth, severe positional artifacts, inappropriate exposure parameters and metallic artifacts. The resulting datasets consisted of 143 to 1881 panoramic radiographs per center leading to 4177 images in total. These highly differing numbers of provided images represent a typical scenario in real-world applications. Aggregated metadata such as the mean number of teeth per panoramic are reported in Table 1.

For the annotation of the datasets, all teeth in each panoramic were segmented and classified according to the FDI World Dental Federation notation, which resulted in 32 classes in total (one per tooth). Each image was segmented by one expert and then verified by a second independent expert. The group of experts consisted of a range of supervised final year dental students and experienced dentists. Labeling was performed independently under standardized conditions using an in-house custom-built annotation tool, which was employed in a wide range of previous work [12–14].

This study was conducted over a 5-fold cross-validation, where data were divided into 60% training data (3 folds), 20% validation data (1 fold), and 20% test data (1 fold). For LL and FL, these data splits per center remained separated. For CL, the single center datasets were merged fold-wise to create a centralized version of the data. The assignment of images for each fold was identical over all learning paradigms.

2.3. Learning paradigms

Image analysis of radiographs is often conducted with neural networks, which are built through an arrangement of mathematical units (artificial neurons) and connections with certain values (referred to as model parameters or weights) between them. Neural networks must be trained to learn the inherent patterns of the images. During training, the model sees exemplary images and optimizes its weights until it is capable to generate the desired output with a low error. Models are typically initialized with predefined values, oftentimes stemming from training on existing datasets (e.g., the ImageNet dataset containing everyday RGB images [15]). In this study, we refer to varying logistics of the learning process (e.g., training location, utilized data) as different learning paradigms. The learning paradigm **CL** is based on data sharing, which allows to pool data of different centers in one central location, where training is conducted. Notably, CL may not always be feasible due to privacy regulations. In this case, centers have either the option to rely on **LL**, which involves training on their data only or to join a FL initiative. In **FL**, participants refrain from exchanging sensitive data and instead share abstract model parameters, which essentially carry the knowledge learned from their data. The exchange and aggregation of parameters in FL are determined by the FL protocol. The most popular FL protocol, referred to as Federated Averaging (FedAvg) was proposed by McMahan et al. [16] and works as follows: After all participants agree on a suitable machine learning approach, the FL server distributes the initialized model parameters to all FL participants to kick off the first round of FL. Participants use these parameters as a starting point to train their model on their local data for a predefined number of local epochs. Secondly, all participants send their model parameters, which carry the knowledge from their local data, back to the server. The local contributions are weighted according to the dataset sizes of the participants. The server then averages all local model parameters to form a set of global model parameters. These global parameters are then distributed to all participants for the next round of FL training. FL is eventually put to a stop when a certain stopping criterion is met, e.g., a predefined number of epochs. A high-level overview of FL and the two alternative training processes, namely Local Learning and Central Learning is visualized in Fig. 1. A more detailed illustration of the FL procedure is displayed in Appendix Fig. 1.

2.4. Training procedure

The implementation and training parameters of the three learning paradigms (FL, LL and CL) are represented in the Appendix. As described previously, different centers provided different number of images (as would likely be the case clinically) ranging from 143 images (3.4% on the overall data share) provided by PPU to 1881 (45% on the overall data share) by Charité. Further numbers are reported within the metadata in Table 1. These large differences in dataset sizes may affect the global model performance of FL heavily as the data contribution of each center directly defines its contribution to FL. Hence, in a sensitivity analysis, FL training was repeated with each participant contributing equally to the global model.

2.5. Performance metrics and statistical analysis

Model performances were primarily quantified by a tooth-based F1-score (F1-score$_{tooth}$), where true positives, false positives and false negatives were computed on a tooth-level instead of the typical pixel-level, as described in the Appendix. Secondary metrics were the pixel-wise F1-score$_{pixel}$, sensitivity and precision (positive predictive value (PPV)). All

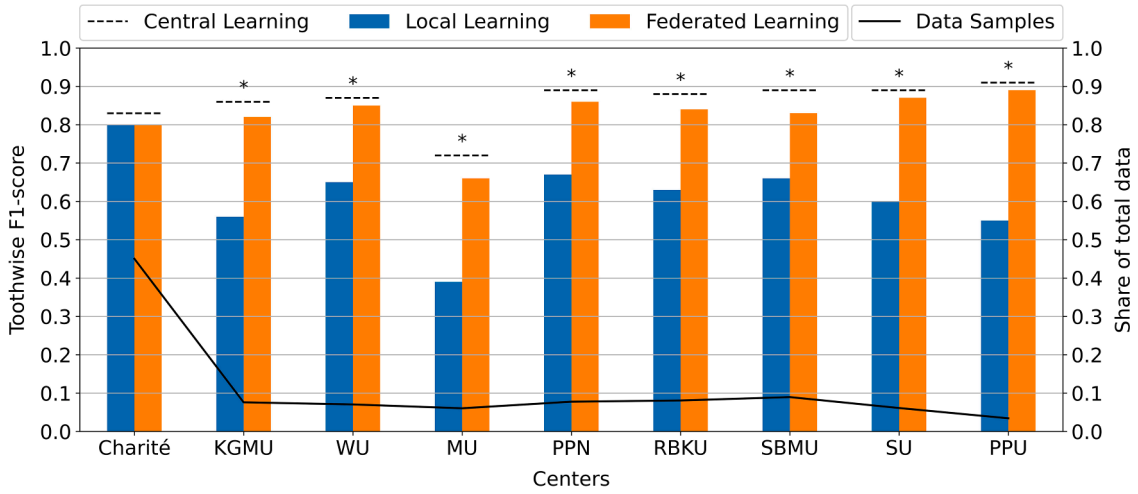


Fig. 2. Model performance (mean toothwise F1-score) of the three different learning paradigms, measured on test sets of each single participant (center) (left y-axis). Asterisk indicates statistical significance ($p<0.05$/Mann-Whitney) between Federated Learning and Local Learning. The relative dataset size from each center is indicated by the black line (right y-axis).

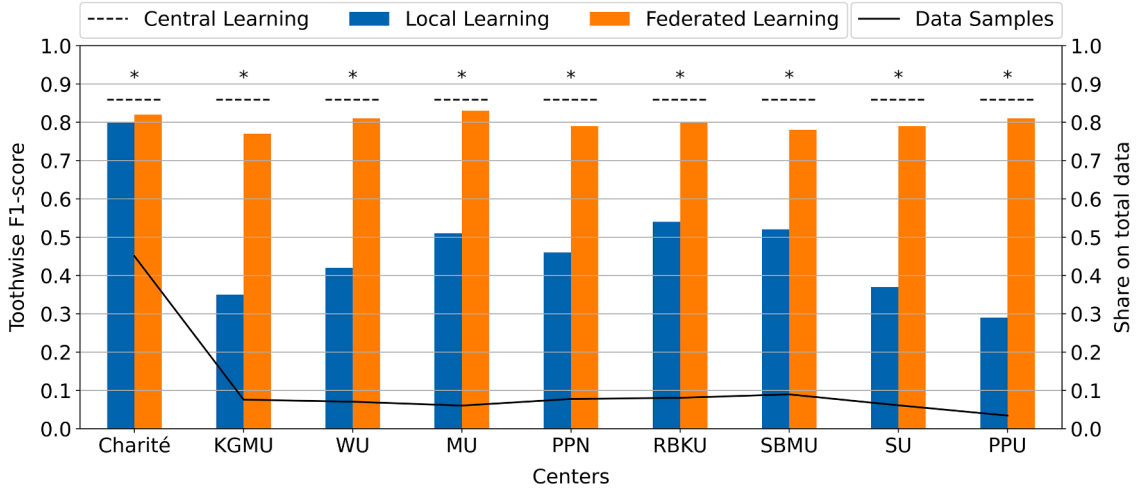


Fig. 3. Generalizability of models developed using the three different learning paradigms, measured as the mean of performances on the test sets pooled from all participants (centers). Asterisk indicates statistical significance ($p<0.05$/Mann-Whitney) between Federated Learning and Local Learning. The relative dataset size from each center is indicated by the black line (right y-axis).

models were evaluated on their own local test dataset to quantify model performance. For generalizability, testing was performed on the combined test data, i.e., including their own and test data of the other participants. We formally tested for statistically significant differences with the non-parametric Mann-Whitney-U-Test. *p*-values level below a significance level of 0.05 were considered statistically significant. The statistical analysis was performed within Python 3.9.2 and SciPy v1.6.2 (SciPy 2021).

3. Results

The model performances of the three different learning paradigms on the local test sets are reported in Fig. 2. For 8 out of 9 participants, FL outperformed LL with statistical significance. Participant PPU, for instance, reached a mean (SD) F1-score$_{tooth}$ of 0.55 (0.032) with LL, which was dramatically outperformed ($p = 0.006$) by FL with an F1-score$_{tooth}$ of 0.888 (0.025). Only for the participant Charité, which provided the largest share of the data, there was no significant difference between LL and FL ($p = 0.338$). For all participants, CL significantly outperformed both FL and LL. Details on other metrics are provided in the Appendix Tables 1-3. *p*-values of the non-parametric Mann-Whitney-U-Test are represented in Appendix Table 4.

The generalizability of the models was captured on the combined test data, which included the test data of the dedicated participant combined with the test data of all other centers (Fig. 3). For all participants, FL outperformed LL ($p<0.05$). Participant KGMU, for instance, reported an F1-score$_{tooth}$ for LL of 0.351 (0.134), which was outperformed by FL with 0.768 (0.117). However, FL (and LL) showed a generalizability gap towards CL ($p<0.001$). Exemplary error cases for the centers KGMU, RBKU and PPU in comparison with their ground truth are reported in Fig. 4. LL was particularly challenged by segmenting restorations, third molars and teeth visualized in low contrast, e.g., lower anteriors due to the overlap with the vertebrae.

As described, centers provided a different number of images for FL, which lead to different contributions for FL in the base-case analysis. Hence, in our sensitivity analysis, equal contributions of each center were employed. This significantly deteriorated the model performance of the largest data provider (Charité) from 0.803 (0.01) to 0.777 (0.007) ($p = 0.008$). For all other centers, however, there was no significant difference in model performance ($p>0.05$). Generalizability of the models showed no statistically significant difference between the base-case and the sensitivity analysis. A visualization of the comparison

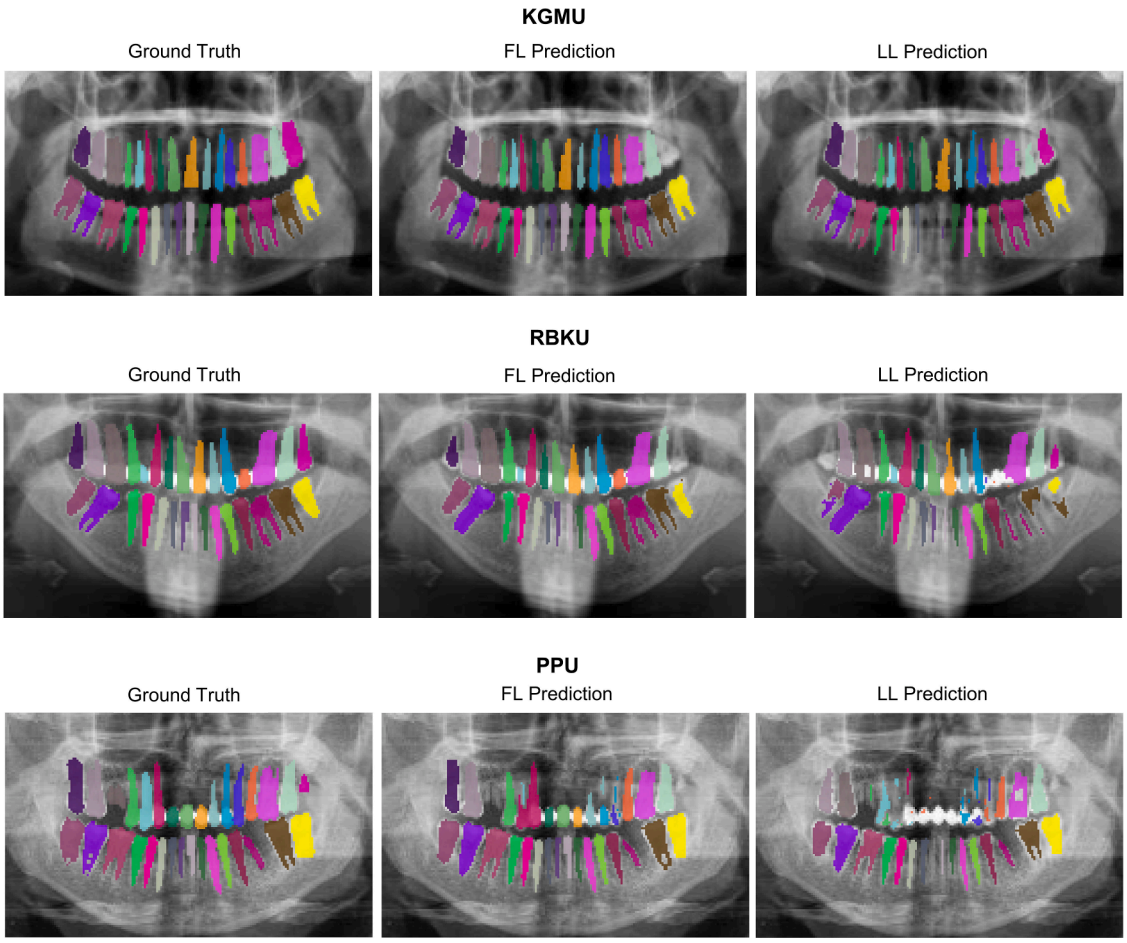
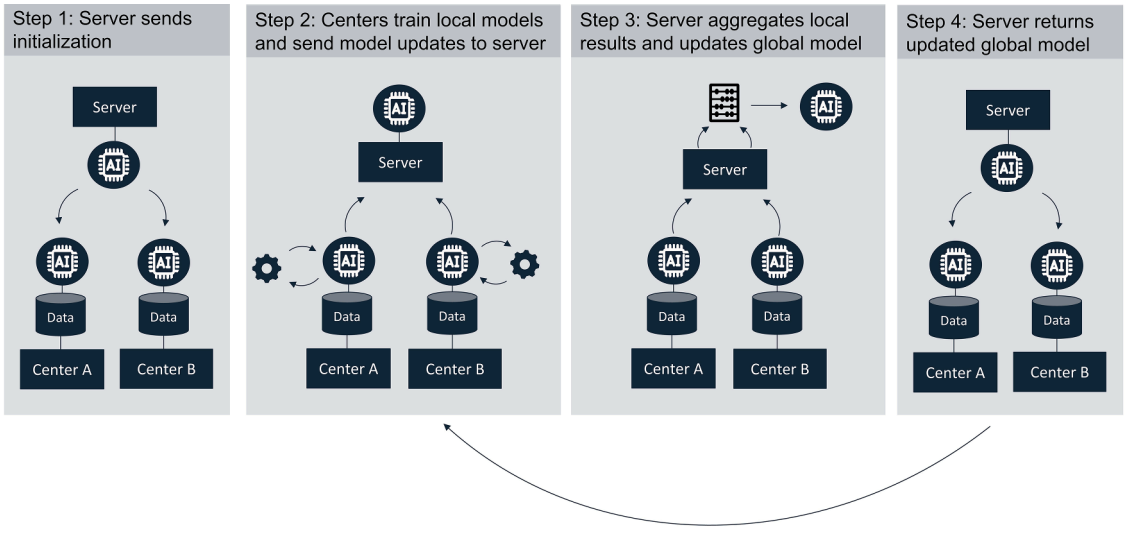
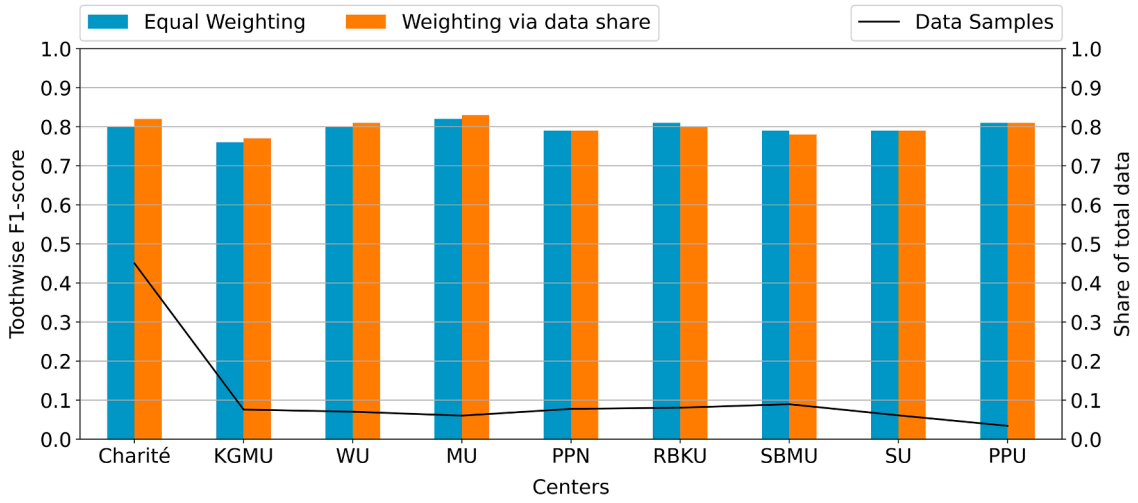


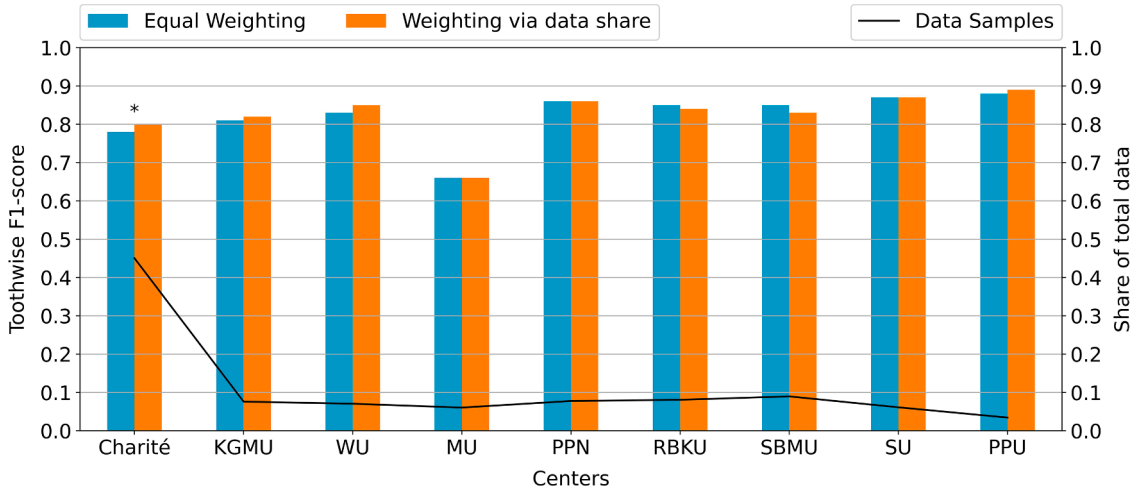
Fig. 4. Ground truth and exemplary predictions of FL and LL on the test set of KGMU, RBKU and PPU (top to bottom).



Appendix Fig. 1. Detailed step-by-step description of the FL procedure.



Appendix Fig. 2. Generalizability (mean toothwise F1-score) of models developed using two different weighting schemes within the FL procedure, measured as the mean of performances on the test sets from all participants (centers). Asterisk indicates statistical significance ($p<0.05$/Mann-Whitney).The relative dataset size from each center is indicated by the black line (right y-axis).



Appendix Fig. 3. Model performance (mean toothwise F1-score) of FL using two different weighting schemes, measured on test sets of each single participant (center) (left y-axis). Asterisk indicates statistical significance ($p<0.05$/Mann-Whitney). The relative dataset size from each center is indicated by the black line (right y-axis).

and all metrics are provided in Appendix Fig. 2-Appendix Fig. 3 and Table 5. *p*-values of the non-parametric Mann-Whitney-U-Test are reported in Appendix Table 6.

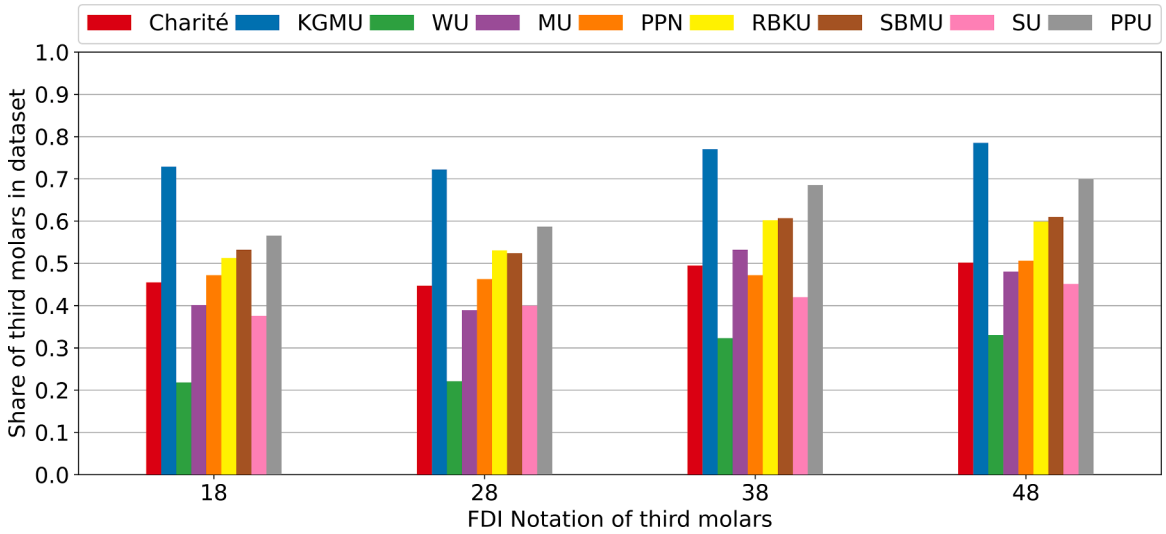
4. Discussion

Dental data is oftentimes considered as particularly sensitive given for instance its usage in forensics [5]. Data sharing of such sensitive data is challenging and administratively complex due to layered and potentially locally varying data protection regulations. Based on this, CL is often not feasible. However, to train generalizable deep learning models, data from different centers is crucial as models from LL may not perform well on data from different centers. FL may be considered as an alternative to LL, when CL is not applicable a priori. Based on the amount of sensitive dental data, FL may be particularly relevant for oral and dental research and applications. In the present study, we compared FL with LL and quantified their performance and generalizability gaps towards CL (as the ideal option) on an exemplary task, tooth segmentation on panoramic radiographs.

We found that for the majority of participants (8 of 9), models trained with FL achieved better performances on their local test sets than those trained with LL. The local datasets of the eight participants were relatively small and seemed to be insufficient to learn the inherent structure of their local data. Only the participant holding the majority of data (Charité) was able to reach similar model performances using LL compared with FL; the local dataset was relatively large and seemed to be suitable to learn the underlying representation of the local data. Moreover, FL yielded models that generalized significantly better than LL models for all participants, i.e., performed better on the pooled test set from all participants.

Notably, a performance and generalizability gap towards CL was observed for both FL and LL, setting out CL as “gold standard” if available. Our findings require more detailed discussion.

Charité, the largest data donor, did not benefit from FL when tested on data from its own institution but showed improved generalizability when compared to LL. It seems that the sample size was sufficient to learn their own inherent data representation but did not provide enough diversity to perform well on unseen data from other centers. This is



Appendix Fig. 4. Occurance of third molars numbered 18, 28, 38 and 48 in the datasets across the different centers expressed as share of samples with prevalence of the specific third molar.

Appendix Table 1
Primary and secondary metrics for CL quantified on test sets of all participants reported with mean value (standard deviation).

Data of Participant	F1-Score (Pixelwise)	F1-Score (Toothwise)	Sensitivity	PPV
Charité	0.89(0.003)	0.833(0.018)	0.836(0.007)	0.95(0.004)
KGMU	0.891(0.011)	0.858(0.013)	0.835(0.013)	0.956(0.009)
MU	0.854(0.004)	0.717(0.022)	0.788(0.01)	0.931(0.008)
PPN	0.897(0.009)	0.889(0.018)	0.846(0.013)	0.955(0.006)
PPU	0.911(0.005)	0.905(0.016)	0.86(0.006)	0.968(0.009)
RBKU	0.899(0.005)	0.881(0.013)	0.848(0.006)	0.958(0.004)
SBMU	0.904(0.007)	0.885(0.006)	0.852(0.006)	0.963(0.009)
SU	0.902(0.009)	0.894(0.013)	0.857(0.007)	0.953(0.012)
WU	0.906(0.005)	0.867(0.01)	0.856(0.007)	0.962(0.004)

notable, as also centers with large datasets may benefit from FL especially if the developed model is not meant to be used solely in that single institution. FL may increase generalizability for any given institutions' deep learning models and hence ease its translation into the clinical environment.

Different participants seem to reach varying model performances using FL. For example, MU reached a mean (SD) F1-score$_{tooth}$ of 0.663 (0.036) while PPU showed a mean (SD) F1-score$_{tooth}$ of 0.888 (0.025). The reason for this lies in the way FL was conducted. Although every participant took part in FL, the final model was significantly shaped by the participant(s) carrying the most data. This is caused by the aggregation method FedAvg, which weights updates according to the participants' dataset size. Consequently, updates from the largest participant had a stronger effect on the FL model than updates from smaller participants. Hence, participants with data close to the distribution of the largest participants (in our case, Charité) performed better in FL than those with rather unique distributions. Here, the data distribution of MU (participant, which benefitted the least from FL) seemed to differ more from the data distribution of Charité than other participants. This heterogeneous nature of the data of different participants is known as dataset shift [17] and may be caused by several factors: "Covariate data shift" may be introduced through varying radiographic devices with different settings. "Prior probability shift" may be caused by varying medical standards, social, and commercial determinants in different centers and countries (e.g., age and dental status may differ and reflect population demographics, oral health and healthcare services). In our case, MU represented an older population with a mean (SD) age of 53.3 (14.9) years compared to Charité with 44.3 (20.0) years and all other participants, with a lower number of teeth (26(5)) than Charité (29(3)) and others, which may be the reason for the lower performance of FL on MU data.

Further, in our exemplary error case analysis, we found that FL models were not able to recognize third molars, since most centers provided images with third molars missing (Appendix Fig. 4). This was to the disadvantage of centers with a high number of third molars (e.g., KGMU). The effect of such shifts may be more apparent if data is not equally distributed across centers. However, this scenario could likely be expected in real life, where different donors provide differently sized datasets with potentially unique features. Rebalancing the weighting scheme of the contributions (like we did in our sensitivity analysis) seems like a tempting option to tackle the discussed biases (e.g., increasing the relative contribution of KGMU should lead to a better generalizability of all models for segmenting third molars). However, the sensitivity analysis showed that standardizing the contribution of each center equally had no effect on the generalizability of all centers, and only came with adverse effects when testing it on Charité data. Therefore, manual weighting of contributions from different centers should be carefully undertaken and monitored appropriately.

Apart from this, the performance and generalizability gaps between FL and CL, observed over the entirety of experiments, should be highlighted, and discussed. The root cause for this difference is the previously discussed dataset shift. In an ideal FL scenario, data would be independent and identically distributed, allowing the same performances in FL and CL. However, the heterogeneity of the underlying data usually hampers model convergence in FL and leads to lower performance for FL than CL [18,19–23]. An extensive hyperparameter search

Appendix Table 2
Primary and secondary metrics for FL quantified on test sets of all participants reported with mean value (standard deviation).

Model of Participant	Data of Participant	F1-Score (Pixelwise)	F1-Score (Toothwise)	Sensitivity	PPV
Charité	Charité	0.881(0.005)	0.803(0.011)	0.826(0.008)	0.945(0.004)
	KGMU	0.871(0.007)	0.806(0.017)	0.83(0.011)	0.916(0.011)
	MU	0.83(0.006)	0.612(0.035)	0.756(0.008)	0.92(0.005)
	PPN	0.887(0.013)	0.867(0.021)	0.845(0.013)	0.933(0.013)
	PPU	0.906(0.009)	0.899(0.022)	0.866(0.015)	0.95(0.004)
	RBKU	0.889(0.007)	0.859(0.02)	0.845(0.01)	0.938(0.005)
	SBMU	0.892(0.01)	0.855(0.013)	0.847(0.014)	0.943(0.006)
	SU	0.894(0.009)	0.878(0.018)	0.86(0.015)	0.93(0.01)
	WU	0.893(0.007)	0.83(0.018)	0.852(0.011)	0.938(0.009)
KGMU	Charité	0.864(0.002)	0.694(0.014)	0.789(0.005)	0.956(0.005)
	KGMU	0.879(0.01)	0.818(0.029)	0.82(0.018)	0.948(0.006)
	MU	0.814(0.005)	0.471(0.05)	0.724(0.008)	0.93(0.003)
	PPN	0.879(0.01)	0.835(0.028)	0.819(0.013)	0.95(0.007)
	PPU	0.892(0.009)	0.839(0.025)	0.831(0.013)	0.963(0.008)
	RBKU	0.878(0.004)	0.797(0.021)	0.813(0.009)	0.954(0.005)
	SBMU	0.883(0.011)	0.815(0.025)	0.818(0.016)	0.96(0.005)
	SU	0.889(0.012)	0.851(0.025)	0.837(0.017)	0.949(0.015)
	WU	0.889(0.003)	0.792(0.026)	0.828(0.01)	0.961(0.007)
MU	Charité	0.884(0.003)	0.81(0.011)	0.834(0.006)	0.941(0.006)
	KGMU	0.87(0.012)	0.796(0.029)	0.835(0.019)	0.909(0.005)
	MU	0.843(0.005)	0.663(0.036)	0.776(0.011)	0.923(0.007)
	PPN	0.888(0.011)	0.865(0.019)	0.854(0.015)	0.925(0.007)
	PPU	0.905(0.01)	0.895(0.017)	0.872(0.012)	0.939(0.009)
	RBKU	0.893(0.003)	0.859(0.015)	0.856(0.004)	0.933(0.005)
	SBMU	0.894(0.006)	0.851(0.015)	0.855(0.011)	0.937(0.004)
	SU	0.894(0.011)	0.87(0.023)	0.867(0.015)	0.922(0.011)
	WU	0.897(0.004)	0.841(0.012)	0.863(0.006)	0.935(0.007)
PPN	Charité	0.873(0.005)	0.75(0.02)	0.806(0.009)	0.951(0.003)
	KGMU	0.869(0.012)	0.796(0.029)	0.817(0.019)	0.928(0.011)
	MU	0.82(0.011)	0.55(0.05)	0.738(0.016)	0.922(0.005)
	PPN	0.889(0.013)	0.861(0.025)	0.837(0.016)	0.946(0.01)
	PPU	0.902(0.009)	0.867(0.022)	0.85(0.014)	0.96(0.006)
	RBKU	0.884(0.008)	0.823(0.018)	0.828(0.013)	0.948(0.003)
	SBMU	0.888(0.009)	0.825(0.026)	0.831(0.014)	0.954(0.006)
	SU	0.892(0.014)	0.857(0.026)	0.849(0.019)	0.94(0.013)
	WU	0.89(0.006)	0.811(0.021)	0.839(0.012)	0.947(0.006)
PPU	Charité	0.875(0.006)	0.769(0.016)	0.810(0.010)	0.951(0.004)
	KGMU	0.867(0.014)	0.803(0.028)	0.818(0.018)	0.922(0.014)
	MU	0.827(0.008)	0.576(0.030)	0.747(0.011)	0.926(0.005)
	PPN	0.884(0.012)	0.861(0.022)	0.835(0.015)	0.939(0.010)
	PPU	0.901(0.009)	0.888(0.025)	0.852(0.013)	0.957(0.007)
	RBKU	0.887(0.009)	0.846(0.022)	0.835(0.012)	0.946(0.006)
	SBMU	0.892(0.007)	0.852(0.013)	0.839(0.012)	0.952(0.001)
	SU	0.893(0.009)	0.867(0.021)	0.851(0.014)	0.939(0.011)
	WU	0.894(0.009)	0.824(0.023)	0.846(0.015)	0.949(0.009)
RBKU	Charité	0.873(0.008)	0.750(0.032)	0.806(0.014)	0.953(0.004)
	KGMU	0.868(0.008)	0.801(0.019)	0.815(0.013)	0.928(0.008)
	MU	0.819(0.012)	0.546(0.026)	0.735(0.012)	0.926(0.013)
	PPN	0.883(0.007)	0.851(0.013)	0.831(0.008)	0.942(0.011)
	PPU	0.901(0.013)	0.880(0.032)	0.850(0.017)	0.959(0.014)
	RBKU	0.888(0.009)	0.84(0.028)	0.833(0.016)	0.951(0.005)
	SBMU	0.890(0.007)	0.835(0.022)	0.834(0.013)	0.953(0.006)
	SU	0.891(0.014)	0.857(0.032)	0.847(0.019)	0.940(0.013)
	WU	0.892(0.008)	0.816(0.031)	0.839(0.016)	0.952(0.006)
SBMU	Charité	0.869(0.007)	0.731(0.02)	0.802(0.011)	0.949(0.006)
	KGMU	0.867(0.006)	0.785(0.016)	0.814(0.013)	0.927(0.008)
	MU	0.814(0.004)	0.51(0.027)	0.728(0.006)	0.922(0.006)
	PPN	0.88(0.007)	0.837(0.013)	0.826(0.009)	0.94(0.008)
	PPU	0.897(0.014)	0.861(0.029)	0.845(0.019)	0.955(0.009)
	RBKU	0.881(0.008)	0.821(0.024)	0.826(0.011)	0.945(0.007)
	SBMU	0.889(0.012)	0.834(0.018)	0.834(0.017)	0.953(0.008)
	SU	0.886(0.015)	0.852(0.033)	0.844(0.021)	0.932(0.013)
	WU	0.888(0.009)	0.805(0.02)	0.837(0.015)	0.946(0.008)
SU	Charité	0.87(0.004)	0.733(0.032)	0.800(0.01)	0.955(0.006)
	KGMU	0.873(0.009)	0.803(0.034)	0.816(0.02)	0.939(0.008)
	MU	0.821(0.009)	0.521(0.059)	0.735(0.016)	0.929(0.009)
	PPN	0.881(0.012)	0.849(0.032)	0.824(0.016)	0.946(0.01)
	PPU	0.9(0.01)	0.866(0.024)	0.844(0.016)	0.964(0.009)
	RBKU	0.884(0.006)	0.826(0.027)	0.826(0.012)	0.952(0.007)
	SBMU	0.889(0.01)	0.835(0.026)	0.83(0.017)	0.958(0.006)
	SU	0.894(0.011)	0.869(0.019)	0.846(0.018)	0.947(0.013)
	WU	0.89(0.005)	0.805(0.025)	0.833(0.013)	0.955(0.006)

(*continued on next page*)

Appendix Table 2 (*continued*)

Model of Participant	Data of Participant	F1-Score (Pixelwise)	F1-Score (Toothwise)	Sensitivity	PPV
WU	Charité	0.876(0.002)	0.765(0.017)	0.811(0.005)	0.952(0.006)
	KGMU	0.875(0.011)	0.81(0.024)	0.824(0.019)	0.933(0.011)
	MU	0.828(0.008)	0.568(0.04)	0.746(0.012)	0.93(0.006)
	PPN	0.889(0.01)	0.861(0.023)	0.841(0.015)	0.942(0.005)
	PPU	0.903(0.007)	0.882(0.02)	0.856(0.01)	0.955(0.008)
	RBKU	0.889(0.003)	0.846(0.014)	0.837(0.007)	0.948(0.007)
	SBMU	0.89(0.01)	0.832(0.03)	0.834(0.017)	0.954(0.005)
	SU	0.893(0.01)	0.869(0.017)	0.85(0.014)	0.94(0.012)
	WU	0.903(0.003)	0.853(0.008)	0.854(0.006)	0.959(0.008)

Appendix Table 3
Primary and secondary metrics for LL quantified on test sets of all participants reported with mean value (standard deviation).

Model of Participant	Data of Participant	F1-Score (Pixelwise)	F1-Score (Toothwise)	Sensitivity	PPV
Charité	Charité	0.881(0.004)	0.8(0.024)	0.823(0.007)	0.947(0.002)
	KGMU	0.85(0.012)	0.756(0.023)	0.803(0.016)	0.902(0.007)
	MU	0.817(0.003)	0.59(0.024)	0.737(0.007)	0.916(0.007)
	PPN	0.88(0.014)	0.843(0.028)	0.833(0.018)	0.932(0.009)
	PPU	0.902(0.008)	0.892(0.025)	0.862(0.011)	0.947(0.006)
	RBKU	0.885(0.006)	0.848(0.024)	0.84(0.012)	0.936(0.002)
	SBMU	0.887(0.003)	0.83(0.014)	0.84(0.006)	0.939(0.003)
	SU	0.887(0.009)	0.861(0.015)	0.852(0.009)	0.926(0.011)
	WU	0.879(0.01)	0.792(0.024)	0.829(0.014)	0.935(0.006)
KGMU	Charité	0.708(0.017)	0.203(0.046)	0.583(0.021)	0.899(0.019)
	KGMU	0.818(0.009)	0.564(0.041)	0.729(0.014)	0.933(0.014)
	MU	0.666(0.023)	0.103(0.023)	0.537(0.028)	0.879(0.026)
	PPN	0.754(0.008)	0.412(0.043)	0.648(0.014)	0.904(0.019)
	PPU	0.757(0.022)	0.37(0.027)	0.65(0.024)	0.905(0.021)
	RBKU	0.715(0.01)	0.308(0.037)	0.597(0.011)	0.89(0.018)
	SBMU	0.739(0.025)	0.356(0.065)	0.625(0.033)	0.905(0.017)
	SU	0.756(0.047)	0.441(0.087)	0.654(0.056)	0.896(0.035)
	WU	0.777(0.008)	0.402(0.032)	0.674(0.011)	0.918(0.018)
MU	Charité	0.784(0.006)	0.447(0.017)	0.699(0.01)	0.893(0.008)
	KGMU	0.769(0.021)	0.454(0.047)	0.701(0.034)	0.851(0.011)
	MU	0.775(0.01)	0.389(0.033)	0.688(0.016)	0.886(0.01)
	PPN	0.803(0.021)	0.568(0.054)	0.741(0.03)	0.877(0.014)
	PPU	0.828(0.011)	0.621(0.022)	0.772(0.02)	0.894(0.012)
	RBKU	0.796(0.011)	0.537(0.031)	0.727(0.017)	0.878(0.01)
	SBMU	0.799(0.013)	0.503(0.05)	0.732(0.02)	0.879(0.014)
	SU	0.799(0.018)	0.555(0.026)	0.741(0.022)	0.867(0.016)
	WU	0.818(0.007)	0.534(0.031)	0.754(0.016)	0.894(0.009)
PPN	Charité	0.762(0.009)	0.365(0.024)	0.649(0.01)	0.923(0.012)
	KGMU	0.741(0.021)	0.468(0.031)	0.633(0.023)	0.894(0.02)
	MU	0.676(0.017)	0.193(0.033)	0.547(0.021)	0.884(0.013)
	PPN	0.836(0.012)	0.667(0.038)	0.752(0.013)	0.942(0.011)
	PPU	0.817(0.016)	0.546(0.019)	0.718(0.02)	0.949(0.012)
	RBKU	0.752(0.011)	0.419(0.017)	0.634(0.014)	0.923(0.011)
	SBMU	0.751(0.007)	0.421(0.026)	0.635(0.011)	0.92(0.009)
	SU	0.8(0.023)	0.564(0.038)	0.706(0.03)	0.925(0.017)
	WU	0.782(0.013)	0.495(0.029)	0.684(0.012)	0.913(0.018)
PPU	Charité	0.693(0.011)	0.206(0.027)	0.566(0.014)	0.895(0.007)
	KGMU	0.612(0.036)	0.222(0.042)	0.494(0.036)	0.805(0.03)
	MU	0.608(0.019)	0.085(0.011)	0.477(0.017)	0.835(0.026)
	PPN	0.7(0.032)	0.317(0.064)	0.576(0.039)	0.893(0.013)
	PPU	0.819(0.018)	0.55(0.032)	0.726(0.018)	0.939(0.018)
	RBKU	0.729(0.012)	0.344(0.044)	0.614(0.016)	0.898(0.007)
	SBMU	0.705(0.023)	0.292(0.033)	0.584(0.025)	0.887(0.015)
	SU	0.705(0.022)	0.319(0.033)	0.584(0.027)	0.89(0.012)
	WU	0.691(0.009)	0.268(0.039)	0.569(0.013)	0.878(0.007)
RBKU	Charité	0.793(0.016)	0.452(0.051)	0.693(0.023)	0.926(0.004)
	KGMU	0.761(0.025)	0.505(0.067)	0.667(0.037)	0.886(0.012)
	MU	0.719(0.024)	0.26(0.062)	0.606(0.034)	0.886(0.011)
	PPN	0.803(0.018)	0.599(0.039)	0.717(0.028)	0.915(0.009)
	PPU	0.851(0.018)	0.67(0.049)	0.774(0.025)	0.946(0.01)
	RBKU	0.83(0.01)	0.632(0.027)	0.746(0.016)	0.936(0.003)
	SBMU	0.81(0.018)	0.585(0.056)	0.723(0.027)	0.923(0.007)
	SU	0.805(0.027)	0.611(0.06)	0.723(0.037)	0.908(0.017)
	WU	0.805(0.016)	0.555(0.046)	0.718(0.023)	0.917(0.009)
SBMU	Charité	0.789(0.007)	0.421(0.041)	0.687(0.012)	0.925(0.011)
	KGMU	0.769(0.031)	0.533(0.06)	0.673(0.038)	0.898(0.024)

(*continued on next page*)

Appendix Table 3 (*continued*)

Model of Participant	Data of Participant	F1-Score (Pixelwise)	F1-Score (Toothwise)	Sensitivity	PPV
	MU	0.706(0.015)	0.214(0.035)	0.589(0.022)	0.884(0.012)
	PPN	0.792(0.019)	0.561(0.056)	0.698(0.03)	0.916(0.009)
	PPU	0.841(0.02)	0.632(0.036)	0.761(0.021)	0.939(0.024)
	RBKU	0.803(0.012)	0.546(0.042)	0.708(0.018)	0.927(0.01)
	SBMU	0.847(0.015)	0.658(0.054)	0.766(0.023)	0.946(0.009)
	SU	0.827(0.022)	0.649(0.037)	0.749(0.022)	0.922(0.025)
	WU	0.775(0.022)	0.473(0.052)	0.676(0.025)	0.907(0.022)
SU	Charité	0.737(0.018)	0.257(0.025)	0.619(0.021)	0.913(0.008)
	KGMU	0.719(0.029)	0.391(0.03)	0.608(0.032)	0.879(0.024)
	MU	0.634(0.016)	0.101(0.021)	0.505(0.019)	0.853(0.014)
	PPN	0.763(0.018)	0.402(0.013)	0.653(0.019)	0.918(0.016)
	PPU	0.795(0.031)	0.465(0.059)	0.692(0.04)	0.935(0.015)
	RBKU	0.723(0.023)	0.334(0.034)	0.601(0.029)	0.906(0.011)
	SBMU	0.759(0.017)	0.4(0.017)	0.651(0.019)	0.91(0.014)
	SU	0.819(0.018)	0.604(0.048)	0.737(0.024)	0.923(0.01)
	WU	0.749(0.008)	0.371(0.032)	0.638(0.011)	0.907(0.014)
WU	Charité	0.729(0.01)	0.287(0.009)	0.614(0.015)	0.897(0.006)
	KGMU	0.747(0.018)	0.458(0.055)	0.651(0.027)	0.877(0.004)
	MU	0.662(0.019)	0.166(0.03)	0.542(0.028)	0.85(0.017)
	PPN	0.766(0.024)	0.49(0.058)	0.671(0.032)	0.893(0.017)
	PPU	0.795(0.019)	0.475(0.033)	0.695(0.026)	0.929(0.013)
	RBKU	0.735(0.019)	0.366(0.022)	0.622(0.028)	0.898(0.012)
	SBMU	0.731(0.035)	0.349(0.046)	0.615(0.045)	0.903(0.014)
	SU	0.774(0.023)	0.51(0.046)	0.682(0.03)	0.895(0.025)
	WU	0.852(0.006)	0.647(0.026)	0.774(0.007)	0.947(0.006)

Appendix Table 4

p-values of the non-parametric Mann-Whitney-U-Test, which formally tested for statistically significant differences of the model performance and generalizability of different learning paradigms measured as F1-score. *p*-values level below a significance level of 0.05 were considered as statistically significant.

	Model Performance		Model Generalizability	
Participant	FL vs LL	FL vs CL	FL vs LL	FL vs CL
Charité	0.338	0.018	0.035	0.001
MU	0.006	0.030	<0.001	0.001
KGMU	0.006	0.011	<0.001	<0.001
WU	0.006	0.030	<0.001	<0.001
PPU	0.006	0.105	<0.001	<0.001
PPN	0.006	0.072	<0.001	<0.001
SU	0.006	0.030	<0.001	<0.001
RBKU	0.006	0.018	<0.001	<0.001
SBMU	0.006	0.006	<0.001	<0.001

for FL may circumvent this matter to some degree, while this was not in our focus here. Employing other FL regimens not building on the assumption of identically distributed data may also facilitate to bridge the gap between FL and CL. Notably, and most relevantly, CL will not always be available in real life given data protection concerns, and FL may be the only valid alternative over LL to achieve acceptable performance and generalizability.

This study comes with a range of strengths and limitations. First, this study represents the largest cross-center study on deep learning in dentistry. Our collaborative efforts enable new research possibilities in terms of cross-center heterogeneities and biases, which is highly important to achieve models that generalize well on unseen data. Second, it is the first systematic application of FL in dentistry, which is relevant for all dental AI researchers given the high data protection barriers for pooling dental data. Third, and as a limiting factor, this study was based on a simulation of FL instead of true implementation. The latter one may be hampered by technical difficulties, as each center requires a technical expert on site, and may significantly limit the possibilities of FL in real life. Fourth, we only explored FL for one specific task, tooth segmentation, on one specific image material, panoramic radiographs. Moreover, FL may be relevant for non-image or multimodal data pools and should be explored in more depth for such applications. Further, data labeling was conducted by one expert and controlled by a second expert. This seemed justified for our task, which was rather simple to conduct. For labeling of pathologies, a larger number of experts should label each image, or a hard ground truth should be employed instead. Notably, different labeling schemes in different centers may affect both learning (particularly LL) and testing, which is why standardized labeling should be attempted [24]. Finally, we have not conducted an extensive hyperparameter search for either of the learning paradigms as it would introduce untenable computational costs and most likely will not change the outcome of this study as all learning paradigms would benefit from hyperparameter tuning.

5. Conclusion

FL boosted the model performance and generalizability on our exemplary deep learning task in nearly all involved centers in comparison to LL. FL is a more suitable alternative to LL, when CL is not practicable due to privacy regulations. Further research should be conducted to reduce the performance gap between FL and CL.

CRediT

All authors revised the paper and gave their final approval and agreed to be accountable for all aspects of the work.

Funding

Uribe was supported by European Union's Horizon 2020 grant agreement 857287 for the Baltic Biomaterials Centre of Excellence, Headquarters at Riga Technical University, Riga, Latvia.

Appendix Table 5
Primary and secondary metrics for FL with equal weighting of contributions quantified on test sets of all participants reported with mean value (standard deviation).

Model of Participant	Data of Participant	F1-Score (Pixelwise)	F1-Score (Toothwise)	Sensitivity	PPV
Charité	Charité	0.877(0.003)	0.777(0.008)	0.818(0.005)	0.945(0.004)
	KGMU	0.866(0.004)	0.796(0.009)	0.824(0.008)	0.914(0.004)
	MU	0.822(0.005)	0.581(0.029)	0.746(0.009)	0.916(0.007)
	PPN	0.882(0.013)	0.851(0.015)	0.837(0.016)	0.933(0.012)
	PPU	0.9(0.01)	0.875(0.021)	0.856(0.013)	0.948(0.009)
	RBKU	0.886(0.005)	0.839(0.019)	0.84(0.01)	0.938(0.003)
	SBMU	0.888(0.004)	0.833(0.016)	0.839(0.003)	0.944(0.006)
	SU	0.891(0.009)	0.865(0.017)	0.855(0.01)	0.93(0.009)
	WU	0.889(0.005)	0.819(0.014)	0.844(0.009)	0.939(0.004)
KGMU	Charité	0.865(0.002)	0.694(0.023)	0.791(0.005)	0.955(0.004)
	KGMU	0.877(0.007)	0.809(0.023)	0.818(0.013)	0.945(0.005)
	MU	0.814(0.008)	0.492(0.043)	0.725(0.013)	0.928(0.003)
	PPN	0.878(0.015)	0.827(0.027)	0.817(0.02)	0.95(0.01)
	PPU	0.892(0.009)	0.835(0.026)	0.832(0.014)	0.962(0.009)
	RBKU	0.877(0.007)	0.788(0.027)	0.812(0.012)	0.952(0.005)
	SBMU	0.886(0.008)	0.813(0.027)	0.823(0.012)	0.959(0.004)
	SU	0.887(0.012)	0.841(0.021)	0.835(0.015)	0.946(0.014)
	WU	0.885(0.004)	0.773(0.027)	0.821(0.009)	0.959(0.009)
MU	Charité	0.882(0.004)	0.8(0.013)	0.829(0.008)	0.942(0.005)
	KGMU	0.873(0.012)	0.806(0.025)	0.835(0.02)	0.914(0.007)
	MU	0.84(0.008)	0.665(0.048)	0.77(0.015)	0.925(0.003)
	PPN	0.888(0.009)	0.862(0.016)	0.85(0.013)	0.93(0.009)
	PPU	0.903(0.009)	0.883(0.021)	0.866(0.011)	0.944(0.009)
	RBKU	0.89(0.005)	0.853(0.014)	0.847(0.01)	0.937(0.004)
	SBMU	0.893(0.007)	0.847(0.026)	0.85(0.014)	0.941(0.004)
	SU	0.891(0.013)	0.865(0.025)	0.864(0.017)	0.921(0.011)
	WU	0.893(0.007)	0.825(0.012)	0.85(0.012)	0.94(0.011)
PPN	Charité	0.87(0.005)	0.734(0.029)	0.8(0.009)	0.954(0.004)
	KGMU	0.869(0.005)	0.798(0.022)	0.814(0.009)	0.932(0.004)
	MU	0.816(0.007)	0.539(0.036)	0.731(0.013)	0.924(0.006)
	PPN	0.886(0.014)	0.856(0.025)	0.831(0.019)	0.949(0.008)
	PPU	0.899(0.009)	0.852(0.023)	0.843(0.012)	0.962(0.01)
	RBKU	0.882(0.006)	0.815(0.029)	0.822(0.012)	0.952(0.007)
	SBMU	0.886(0.006)	0.821(0.026)	0.826(0.011)	0.956(0.004)
	SU	0.891(0.006)	0.854(0.013)	0.843(0.008)	0.944(0.012)
	WU	0.889(0.006)	0.797(0.028)	0.831(0.012)	0.956(0.008)
PPU	Charité	0.875(0.004)	0.764(0.014)	0.81(0.007)	0.953(0.004)
	KGMU	0.871(0.007)	0.808(0.017)	0.822(0.011)	0.927(0.005)
	MU	0.821(0.011)	0.57(0.038)	0.741(0.015)	0.92(0.009)
	PPN	0.887(0.01)	0.86(0.018)	0.835(0.013)	0.946(0.008)
	PPU	0.902(0.012)	0.879(0.019)	0.852(0.014)	0.958(0.01)
	RBKU	0.889(0.006)	0.846(0.015)	0.836(0.009)	0.951(0.005)
	SBMU	0.896(0.003)	0.851(0.011)	0.842(0.004)	0.956(0.003)
	SU	0.894(0.01)	0.874(0.018)	0.852(0.011)	0.941(0.01)
	WU	0.894(0.005)	0.824(0.016)	0.841(0.008)	0.954(0.005)
RBKU	Charité	0.877(0.002)	0.763(0.021)	0.811(0.005)	0.955(0.004)
	KGMU	0.876(0.008)	0.815(0.022)	0.825(0.014)	0.934(0.002)
	MU	0.829(0.007)	0.572(0.055)	0.746(0.015)	0.932(0.006)
	PPN	0.889(0.015)	0.861(0.026)	0.837(0.021)	0.948(0.007)
	PPU	0.905(0.006)	0.881(0.015)	0.854(0.006)	0.961(0.008)
	RBKU	0.891(0.004)	0.85(0.007)	0.837(0.006)	0.951(0.006)
	SBMU	0.894(0.008)	0.845(0.028)	0.838(0.012)	0.957(0.004)
	SU	0.894(0.008)	0.871(0.018)	0.852(0.009)	0.941(0.011)
	WU	0.894(0.004)	0.826(0.019)	0.842(0.007)	0.953(0.008)
SBMU	Charité	0.871(0.005)	0.734(0.025)	0.803(0.009)	0.952(0.004)
	KGMU	0.872(0.006)	0.808(0.014)	0.822(0.009)	0.928(0.006)
	MU	0.817(0.011)	0.521(0.042)	0.734(0.016)	0.922(0.004)
	PPN	0.885(0.011)	0.853(0.025)	0.832(0.015)	0.945(0.007)
	PPU	0.901(0.009)	0.867(0.016)	0.85(0.013)	0.959(0.006)
	RBKU	0.885(0.004)	0.829(0.018)	0.83(0.008)	0.948(0.006)
	SBMU	0.895(0.003)	0.848(0.009)	0.841(0.005)	0.956(0.006)
	SU	0.892(0.009)	0.873(0.015)	0.851(0.009)	0.937(0.012)
	WU	0.889(0.005)	0.803(0.015)	0.836(0.009)	0.948(0.008)
SU	Charité	0.872(0.006)	0.735(0.029)	0.803(0.01)	0.955(0.002)
	KGMU	0.873(0.011)	0.807(0.028)	0.82(0.016)	0.933(0.007)
	MU	0.82(0.01)	0.525(0.035)	0.736(0.015)	0.927(0.006)
	PPN	0.886(0.013)	0.853(0.029)	0.83(0.017)	0.95(0.008)
	PPU	0.899(0.008)	0.866(0.018)	0.845(0.007)	0.962(0.009)
	RBKU	0.886(0.006)	0.825(0.026)	0.827(0.011)	0.953(0.004)
	SBMU	0.891(0.008)	0.84(0.019)	0.833(0.011)	0.959(0.006)
	SU	0.894(0.007)	0.868(0.01)	0.848(0.01)	0.946(0.008)
	WU	0.892(0.006)	0.81(0.033)	0.836(0.009)	0.956(0.003)

(*continued on next page*)

Appendix Table 5 (*continued*)

Model of Participant	Data of Participant	F1-Score (Pixelwise)	F1-Score (Toothwise)	Sensitivity	PPV
WU	Charité	0.874(0.006)	0.753(0.024)	0.808(0.011)	0.952(0.004)
	KGMU	0.874(0.008)	0.808(0.022)	0.821(0.016)	0.934(0.01)
	MU	0.827(0.005)	0.562(0.015)	0.744(0.009)	0.931(0.003)
	PPN	0.885(0.01)	0.852(0.016)	0.834(0.012)	0.944(0.012)
	PPU	0.899(0.011)	0.866(0.035)	0.85(0.018)	0.954(0.006)
	RBKU	0.886(0.009)	0.831(0.033)	0.831(0.018)	0.95(0.004)
	SBMU	0.888(0.01)	0.826(0.023)	0.83(0.016)	0.954(0.006)
	SU	0.891(0.01)	0.861(0.016)	0.847(0.014)	0.939(0.011)
	WU	0.898(0.008)	0.835(0.03)	0.844(0.015)	0.959(0.003)

Appendix Table 6
p-values of the non-parametric Mann-Whitney-U-Test, which formally tested for statistically significant differences of the model performance and generalizability of FL with equal contributions and with contributions weighted by data share. *p*-values level below a significance level of 0.05 were considered as statistically significant.

Participant	Model Performance Equal vs Weighted	Model Generalizability Equal vs Weighted
Charité	0.006	0.083
MU	0.5	0.473
KGMU	0.265	0.534
WU	0.265	0.362
PPU	0.265	0.987
PPN	0.338	0.493
SU	0.5	0.809
RBKU	0.417	0.325
SBMU	0.072	0.255

CRediT authorship contribution statement

Lisa Schneider: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **Roman Rischke:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Joachim Krois:** Conceptualization, Methodology, Resources, Data curation. **Aleksander Krasowski:** Resources, Writing – review & editing. **Martha Büttner:** Resources, Writing – review & editing. **Hossein Mohammad-Rahimi:** Resources, Data curation. **Akhilanand Chaurasia:** Resources, Data curation. **Nielsen S Pereira:** Resources, Data curation. **Jae-Hong Lee:** Resources, Data curation. **Sergio E. Uribe:** Resources, Data curation. **Shahriar Shahab:** Resources, Data curation. **Revan Birke Koca-Ünsal:** Resources, Data curation. **Gürkan Ünsal:** Resources, Data curation. **Yolanda Martinez-Beneyto:** Resources, Data curation. **Janet Brinz:** Resources, Data curation. **Olga Tryfonos:** Resources, Data curation. **Falk Schwendicke:** Conceptualization, Methodology, Resources, Data curation, Writing – review & editing, Supervision.

Declaration of Competing Interest

FS and JK are co-founders of the dentalXrai Ltd., a startup focused on deep learning for radiograph analysis. dentalXrai Ltd. did not have any role in conceiving, conducting or reporting this study.

Appendix

Data sources, ethics and data protection

The following centers participate; ethical approval was covered by the main center in Berlin and, in accordance with local regulations others centers sought local approval, too: (1) (Charité), (2) (MU), (3) (KGMU), (4) (WU), (PPN), (6) (RBKU), (7) (SBMU), (8) (SU) and (9) (PPU). For all partners, data sharing (transfer) agreements in line with European General Data Protection Regulation (GDPR) and Berliner Landeskrankenhausgesetz (LHG, Berlin Hospital Law) were in place. As per these agreements, data donors were to ensure that they fulfilled local legal requirements for processing personal health data and to provide personal data only within the grounds of GDPR (e.g., on informed consent basis or after sufficient de-identification according to local regulation) and/or LHG (allowing the processing of data for scientific purposes).

Training

For all three learning paradigms (FL, LL and CL), UNet++ [25] with a ResNet-34 backbone provided by Iakubovskii [26] was used as model architecture given to its high performance for dental radiograph analysis [14]. The applied loss function was a combination of Focal and Dice loss. The Dice loss is based on the most used metric for evaluating segmentations, namely the Sørensen–Dice coefficient [27]. The Focal loss was developed as an extension of the binary cross-entropy loss, which tackles the issue of class imbalance by reducing the impact of easy examples (here background segmentation) to yield focus on harder examples (tooth segmentation) [28]. Training was performed with the Adam optimizer [29] with a learning rate of 0.0003. For image augmentation, the following data augmentation methods from MONAI v0.9 were applied: Random Gaussian Noise, Random Shift in Intensity, Random Gaussian Smoothing and Random Spatial Crop. All architecture parameters were initialized with pre-trained parameters on ImageNet [15] and optimization was performed over all layers of the architecture. Training was seeded identically. LL and CL were both performed for 300 epochs, while FL was performed for 500 global epochs, each including one local training epoch on the participants' site. Finally, after the last global epoch, local models were fine-tuned to their local data for four epochs.

We deliberately decided to forego early stopping, as it may employ inconsistencies in the models across centers due to varying stopping points of centers in the training process. We selected a higher number of epochs for FL as, by nature, it may take longer to converge. We further employed FedProx [18] in FL, which adds a regularization term to the loss that penalizes large deviations from the global FL model on the participants' site and improves the convergence of FL. No extensive hyperparameter search was conducted, as we aimed at model comparisons instead of maximizing model performances. LL, CL, FL and the sensitivity analysis of FL were performed on multiple NVIDIA A100 40GB GPUs and were all implemented with NVFlare v2.0 [30] and PyTorch v1.12.

Performance metrics and statistical analysis

Model performances were primarily quantified by a tooth-based F1-score (F1-score$_{tooth}$), where true positives, false positives and false negatives were computed on a tooth-level instead of the typical pixel-level. For this, the agreement of label and prediction was assessed by dividing the area of overlap by the area of union (Intersection over Union). An agreement of 0.8 or higher resulted in a true positive tooth count, while lower agreements led to a false positive count. A true negative count was a missing tooth, correctly recognized and therefore, not segmented by

the model. If the model missed a tooth completely, it was counted as false negative. All true positives, false positives and false negatives were summed up over all channels of the segmentation label before computing the F1-score. This computation results in unbiased F1-scores [31].

References

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

[2] H. Mohammad-Rahimi, S.R. Motamedian, M.H. Rohban, J. Krois, S.E. Uribe, E. Mahmoudinia, R. Rokhshad, M. Nadimi, F. Schwendicke, Deep learning for caries detection: a systematic review, J. Dent. 122 (2022), 104115, https://doi.org/10.1016/j.jdent.2022.104115.

[3] Y. Chen, K. Stanley, W. Att, Artificial intelligence in dentistry: current applications and future perspectives, Quintessence Int. 51 (3) (2020) 248–257.

[4] J. Krois, A.Garcia Cantu, A. Chaurasia, R. Patil, P.K. Chaudhari, R. Gaudin, S. Gehrung, F. Schwendicke, Generalizability of deep learning models for dental image analysis, Sci. Rep. 11 (1) (2021) 1–7.

[5] H. James, et al., Thai tsunami victim identification overview to date, J. Forensic Odontostomatol. 23 (1) (2005) 1–18.

[6] N. Rieke, J. Hancox, F. Milletari, H. Roth, S. Albarqouni, S. Bakas, M. Galtier, B. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. Summers, A. Trask, D. Xu, M. Baust, M.J. Cardoso, The future of digital health with federated learning, Npj Digit. Med. 3 (1) (2020), https://doi.org/10.1038/s41746-020-00323-1.

[7] W. Li, F. Milletarì, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M.J. Cardoso, et al., Privacy-preserving federated brain tumour segmentation. International Workshop on Machine Learning in Medical Imaging, Springer, 2019, pp. 133–141.

[8] D. Yang, Z. Xu, W. Li, A. Myronenko, H.R. Roth, S. Harmon, S. Xu, B. Turkbey, E. Turkbey, X. Wang, W. Zhu, G. Carrafiello, F. Patella, M. Cariati, H. Obinata, H. Mori, K. Tamura, P. An, B.J. Wood, D. Xu, Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan, Med. Image Anal. 70 (2021), 101992.

[9] I. Dayan, H.R. Roth, A. Zhong, et al., Federated learning for predicting clinical outcomes in patients with COVID-19, Nat. Med. 27 (10) (2021) 1735–1743.

[10] R. Rischke, L. Schneider, K. Müller, W. Samek, F. Schwendicke, J. Krois, Federated learning in dentistry: chances and challenges, J. Dent. Res. (2022), 00220345221108953.

[11] WHO/ITU. 2022. Focus Group on “Artificial Intelligence for Health”. ITU. [Accessed 2022 Oct 27]. https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx.

[12] J. Ma, L. Schneider, S. Lapuschkin, R. Achtibat, M. Duchrau, J. Krois, F. Schwendicke, W. Samek, Towards trustworthy ai in dentistry, J. Dent. Res. 101 (11) (2022) 1263–1268.

[13] C. Rohrer, J. Krois, J. Patel, H. Meyer-Lueckel, J.A. Rodrigues, F. Schwendicke, Segmentation of dental restorations on panoramic radiographs using deep learning, Diagnostics 12 (6) (2022) 1316.

[14] 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.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

[16] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273–1282.

[17] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence, Dataset Shift in Machine Learning, The MIT Press, 2009.

[18] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, Proc. Mach. Learn. Syst. 2 (2020) 429–450.

[19] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, V. Chandra, ArXiv Preprint ArXiv: 1806.00582, 2018.

[20] T.-M.H. Hsu, H. Qi, M. Brown, ArXiv Preprint ArXiv: 1909.06335, 2019.

[21] F. Sattler, K.-R. Müller, W. Samek, Clustered federated learning: model-agnostic distributed multitask optimization under privacy constraints, IEEE Trans. Neural Netw. Learn. Syst. 32 (8) (2021) 3710–3722.

[22] F. Sattler, T. Korjakow, R. Rischke, W. Samek, FEDAUX: leveraging unlabeled auxiliary data in federated learning, IEEE Trans. Neural Netw. Learn Syst. (2021).

[23] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R.G.L. D'Oliveira, H. Eichner, S.E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P.B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S.U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F.X. Yu, H. Yu, S. Zhao, Advances and open problems in federated learning, Found. Trends® Mach. Learn. 14 (1–2) (2021) 1–210, https://doi.org/10.1561/2200000083.

[24] Ş. Vădineanu, D.M. Pelt, O. Dzyubachyk, K.J. Batenburg, An analysis of the impact of annotation errors on the accuracy of deep learning for cell segmentation, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2022, pp. 1251–1267.

[25] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2019) 1856–1867.

[26] P. Iakubovskii, Segmentation Models Pytorch, 2019. https://github.com/qubvel/segmentation_models.pytorch.

[27] T.A. Sorensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, Biol. Skar. 5 (1948) 1–34.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[29] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2017. http://arxiv.org/abs/1412.6980 (Accessed 15 February 2023).

[30] H.R. Roth, Y. Cheng, Y. Wen, I. Yang, Z. Xu, Y.-T. Hsieh, K. Kersten, A. Harouni, C. Zhao, K. Lu, et al., ArXiv Preprint ArXiv: 2210.13291, 2022.

[31] L. Schneider, P. Dave, L. Arsiwala-Scheppach, F. Schwendicke, J. Krois, Exploring bias in F-score computation methods of multi-class segmentation models, in: 2021 The 5th International Conference on Video and Image Processing, 2021, pp. 76–84.