



Intraoral radiograph anatomical region classification using neural networks

Nikolaos Kyventidis¹ · Christos Angelopoulos²

Received: 15 November 2020 / Accepted: 27 January 2021 / Published online: 24 February 2021
© CARS 2021

Abstract

Purpose Dental radiography represents 13% of all radiological diagnostic imaging. Eliminating the need for manual classification of digital intraoral radiographs could be especially impactful in terms of time savings and metadata quality. However, automating the task can be challenging due to the limited variation and possible overlap of the depicted anatomy. This study attempted to use neural networks to automate the classification of anatomical regions in intraoral radiographs among 22 unique anatomical classes.

Methods Thirty-six literature-based neural network models were systematically developed and trained with full supervision and three different data augmentation strategies. Only libre software and limited computational resources were utilized. The training and validation datasets consisted of 15,254 intraoral periapical and bite-wing radiographs, previously obtained for diagnostic purposes. All models were then comparatively evaluated on a separate dataset as regards their classification performance. Top-1 accuracy, area-under-the-curve and *F1*-score were used as performance metrics. Pairwise comparisons were performed among all models with Mc Nemar's test.

Results Cochran's *Q* test indicated a statistically significant difference in classification performance across all models ($p < 0.001$). Post hoc analysis showed that while most models performed adequately on the task, advanced architectures used in deep learning such as VGG16, MobilenetV2 and InceptionResnetV2 were more robust to image distortions than those in the baseline group (MLPs, 3-block convolutional models). Advanced models exhibited classification accuracy ranging from 81 to 89%, *F1*-score between 0.71 and 0.86 and AUC of 0.86 to 0.94.

Conclusions According to our findings, automated classification of anatomical classes in digital intraoral radiographs is feasible with an expected top-1 classification accuracy of almost 90%, even for images with significant distortions or overlapping anatomy. Model architecture, data augmentation strategies, the use of pooling and normalization layers as well as model capacity were identified as the factors most contributing to classification performance.

Keywords Machine learning · Artificial intelligence · Neural networks, computer · Dentistry · Dental informatics · Diagnostic imaging

Introduction

Background

According to conservative estimates, half a billion dental diagnostic radiological examinations are performed annually, with a global average of 74 per 1000 inhabitants, representing 13% of all diagnostic radiology testing [1]. Three hundred million intraoral radiographs are produced annually in the US and the EU [2, 3]. An increasing share is digitally acquired due to clinical advantages, radiation protection considerations and financial barrier lifting. Being non-physical, digital radiographs tend to be indefinitely

✉ Nikolaos Kyventidis
nkyventidis@gmail.com

Christos Angelopoulos
angelopoulosc@gmail.com

¹ School of Dentistry, Aristotle University of Thessaloniki, 28is Oktobriou 62, 54 642, Thessaloniki, Greece

² School of Dentistry, Aristotle University of Thessaloniki, Faculty of Dentistry, University Campus, 54 124, Thessaloniki, Greece

stored, leading to accumulation in vast archives, especially in large institutions.

Digital X-ray images are accompanied by standardized metadata, ideally generated during production. However, manual recording is often necessary, leading to inaccuracies attributed to lack of staff motivation or training [4]. Deficiencies in healthcare datasets are a well-documented problem associated with labor repetition and high repair cost.

Automated algorithms could become a third party responsible for generating standardized, high-quality metadata, allowing labor reallocation to more creative tasks and uplifting metadata value above the cost of maintenance if deployed at scale.

An intraoral radiograph's anatomical region is a piece of metadata consisting of predetermined anatomical classes that correspond to standard radiographic projections, currently manually recorded by the human operator. Reliable recording is vital for diagnosis and the benefits of DICOM and is a prerequisite for the implementation of hanging protocols and the creation of standardized radiographic layouts [5, 6]. Additionally, it can be valuable for machine learning model development [7] and the searchability of radiographic archives. Automated classification exclusively from pixel data would enable its generation on the modality or database level and eliminate the need for manual preselection.

Since the extent of the presumably depicted anatomy is known, it can be expressed as a problem of image classification among predefined classes, a fundamental problem in the field of computer vision. Recent advancements, especially after the introduction of the still-relevant AlexNet architecture in the ILSVRC competition [8–10], have allowed the domination of the field by a variety of algorithms where neural networks (especially convolutional ones) achieve relatively good performance in the task of classifying natural images into different classes [11].

The subsequent release of libre software that abstracts underlying development processes enabled the rapid development of relevant applications by independent researchers. In 2017, over 300 applications related to radiological imaging had been published [12], while in the US, major organizations in radiology recently issued a roadmap for future research in the field of machine learning in relation to radiology [7].

All the previously discussed benefits could be possibly provided by employing convolutional neural network architectures to achieve automated classification of intraoral radiograph anatomical regions.

In addition, exposing radiology workers to this recently introduced field through applications designed to eliminate trivial tasks can enhance familiarization with its concepts and shortcomings, leading to wider acceptance without the potential implications of diagnostic applications.

The purpose of this study is to systematically develop literature-based neural networks architectures (models) capable of classifying the anatomical region of intraoral radiographs based exclusively on pixel data, as well as to deduce the most appropriate architectures and training strategies by cross-comparing model performance on a predefined dataset. By using libre software, a simple model development methodology, a small dataset and limited computational and financial resources, wide adoption and reproducibility are facilitated. To our knowledge, no similar studies currently exist in the literature.

Materials and methods

The STARD 2015 [13] and CLAIM [14] checklists for the standardization and enhancement of the quality of diagnostic accuracy and artificial intelligence studies were followed where applicable.

Dataset generation

The present study is a retrospective study utilizing archived intraoral radiographs obtained for diagnostic purposes. No subjects were exposed to X-rays for the purposes of this study. All subjects provided written consent.

Adult patients of any age and gender were included in chronological order without further inclusion or exclusion criteria. Non-adult patients were excluded. Included images were digital periapical or bite-wing radiographs obtained by the same modality (PSP plates, SOREDEX Scanora scanner).

The original uncompressed image data were fully anonymized and randomly shuffled by a hash-based algorithm. The dataset was evaluated for content relevance, technical and visual quality and proper classification by one evaluator with 8 years of experience, excluding images of inadequate content or poor quality.

Twenty-two unique anatomical classes were identified. Class definitions assumed that each projection clearly depicts the region of three consecutive teeth, except for the anterior classes 12–22 and 32–42 which included four. Classes LBW1, LBW2 and RBW1, RBW2 were merged due to data scarcity.

The resulting dataset was then randomly split into an 80% training subset and a 20% validation subset for model training. Proper dataset and split sizes were determined by a pilot study.

Model definition

Model definitions and training methodology were based on the dominant initial choices derived from an extensive review of the literature [11] and are described in Table 1.

All models were developed using the Keras API v2.2.4, the Anaconda distribution of Python programming language v3.6.7 and TensorFlow v1.14 as a backend, by adding convolutional architectures as feature extractors on top of a multilayer perceptron classifier and trained using fully supervised learning with backpropagation of error.

Initially, a fully connected two-layer multilayer perceptron (1024 and 512 wide) with flattened input was trained as a baseline classifier. Then, a simple convolutional network of three convolutional blocks as described in TensorFlow's documentation (convolutional layers of width 64, 128 and 128 respectively, a 3×3 filter size, and a max pooling level) was added as a baseline feature extractor.

A group of more advanced and innovative convolutional networks were then consecutively added; the deep but simple convolutional architecture VGG16 [15], the MobileNetV2 as a balanced architecture against model size and performance [16] and the Inception-ResNetV2 as a high-capacity, high-performing architecture especially suitable for small datasets [17, 18].

Additional models were generated with the insertion of batch normalization layers [19] and the use of both the flattening or the global average pooling layer [20] as bridging between the feature extractor part and the MLP classifier.

Dropout layers of 0.20–0.50 rates [21] and three different data augmentation strategies applied on-the-fly were used as regularization (described in Table 2).

The final layer of every model was a softmax-activated dense layer with a range equal to the number of classes (22), so that model output could be expressed as multiple percentages of per class prediction confidence that add up to 100%. The class with the highest confidence was the top model choice for the evaluation of top-1 accuracy.

All models were initialized with default Keras parameters and trained with categorical cross-entropy as a loss function, Adam optimizer [22], a batch size of 64, and learning rate reduction by 50% in validation top-1 accuracy plateaus. Training lasted 100 epochs with early stopping if validation top-1 accuracy or loss function stopped improving after a set number of epochs. ReLUs [23] were used as activation functions. All other hyperparameters were constant.

Input resolution was 224×224 grayscale, or RGB in models requiring three-channel input. The corresponding Keras preprocessing function was used; otherwise, images were rescaled to the 0–1 range.

Reproducibility and fair comparisons among models were facilitated by a common seed value for all pseudo-random number generators.

Class imbalance was mitigated using a weighted version of the loss function based on class weights calculated on the validation subset.

Table 1 Model definitions. MLP: multilayer perceptron, GAP: Global Average Pooling

Model	Bridging layer	Layer count	Parameter count (millions)	Data augmentation
<i>Baseline group 1, MLP with two hidden layers</i>				
0	–	8	51.9	None
1	–	8	51.9	Typical
2	–	8	51.9	Aggressive
<i>Baseline group 2, MLP with two hidden layers and Batch Normalization</i>				
3	–	10	52.1	None
4	–	10	52.1	Typical
5	–	10	52.1	Aggressive
<i>Baseline group 3, Convolutional network with 3 convolutional blocks + MLP</i>				
6	Flatten	17	178.1	None
7	Flatten	17	178.1	Typical
8	Flatten	17	178.1	Aggressive
<i>Baseline group 4, Convolutional network with 3 convolutional blocks + MLP</i>				
9	GAP	17	1.1	None
10	GAP	17	1.1	Typical
11	GAP	17	1.1	Aggressive
<i>Baseline group 5, Convolutional network with 3 convolutional blocks and Batch Normalization + MLP with Batch Normalization</i>				
12	Flatten	24	178.8	None
13	Flatten	24	178.8	Typical
14	Flatten	24	178.8	Aggressive
<i>Baseline group 6, Convolutional network with 3 convolutional blocks and Batch Normalization + MLP with Batch Normalization</i>				
15	GAP	24	1.1	None
16	GAP	24	1.1	Typical
17	GAP	24	1.1	Aggressive
<i>Advanced group 1, VGG16 architecture + MLP with Batch Normalization</i>				
18	Flatten	30	41	None
19	Flatten	30	41	Typical
20	Flatten	30	41	Aggressive
<i>Advanced group 2, VGG16 architecture + MLP with Batch Normalization</i>				
21	GAP	30	15.7	None
22	GAP	30	15.7	Typical
23	GAP	30	15.7	Aggressive
<i>Advanced group 3, MobilenetV2 architecture + MLP with Batch Normalization</i>				
24	Flatten	166	67.2	None
25	Flatten	166	67.2	Typical
26	Flatten	166	67.2	Aggressive
<i>Advanced group 4, MobilenetV2 architecture + MLP with Batch Normalization</i>				
27	GAP	166	4.1	None
28	GAP	166	4.1	Typical
29	GAP	166	4.1	Aggressive

Table 1 (continued)

Model	Bridging layer	Layer count	Parameter count (millions)	Data augmentation
<i>Advanced group 5, InceptionResnetV2 architecture + MLP with Batch Normalization</i>				
30	Flatten	791	94.3	None
31	Flatten	791	94.3	Typical
32	Flatten	791	94.3	Aggressive
<i>Advanced group 6, InceptionResnetV2 architecture + MLP with Batch Normalization</i>				
33	GAP	791	56.4	None
34	GAP	791	56.4	Typical
35	GAP	791	56.4	Aggressive

Model evaluation

Evaluation was performed on a separate test dataset, not involved in model training, consisting of 261 intraoral radiographs of patients, balanced for both sexes and all five age groups described in the NHANES [24], to reduce dataset bias. Its size allows the detection of large discrepancies in accuracy between the test and validation subsets, without being large enough to impact the training dataset.

Total training time, per sample prediction time, loss function minimization, top-1 accuracy, precision, recall, F1-score and area under the curve were obtained by Keras and Scikit-Learn.

Metrics were macro-averaged from all classes, where applicable. Since some classes were considerably under-represented, in-depth per class performance analysis was deemed out-of-scope.

Top-1 class predictions for all test images were dichotomized to either true or false predictions in a one-vs-all fashion against the ground truth, then used for statistical analysis.

A significance level of 0.05 was set for all statistical tests. Comparison of the proportions of misclassifications across all models was performed with Cochran's Q test. Pairwise comparisons across models were performed with McNemar's test. *P* values were adjusted with the Bonferroni correction [25–27].

Test calculations were performed with IBM SPSS statistical package version 25, *p* value adjustments with R statistical package version 3.6.1 and ROC curves were calculated with Scikit-Learn version 0.22.

Results

Dataset generation

Out of a total of 17,781 images, 15,254 were accepted for further processing, resulting in a training subset of 12,213 images and a validation subset of 3041 images. Due to prior randomization and anonymization, the exact number of subjects included in the study is unknown. Class weights ranged from 1 to 26.65. A full dataset description is given in Table 3.

A total of 36 models were trained and evaluated. Summaries of model training history and model performance across the test dataset are presented in Table 4. Learning curves for each model are found in "Appendix of ESM".

Comparison among all models

Cochran's *Q* test indicates a statistically significant difference across the 36 models in terms of the proportion of misclassifications on the test dataset, $\chi^2_{(35)} = 3034.949$, $p < 0.001$.

Table 2 Data augmentation strategies with output examples

	No augmentation	Typical augmentation	Aggressive augmentation
<i>Brightness</i>	100%	80–120%	80–120%
<i>Rotation (degrees)</i>	0	up to 10	up to 90
<i>Shifting</i>	0%	5%	10%
<i>Zooming</i>	100%	95–105%	90–110%
<i>Shearing</i>	0%	5%	10%



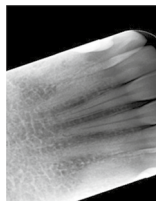




Table 3 Per anatomic class sample populations and weights. RBW: right bite-wing, LBW: left bite-wing

Class	Anatomic Region	Training samples	Validation samples	Test samples	Class weight
0	RBW	562	140	20	4.95
1	16–18	134	33	3	21.00
2	15–17	333	83	7	8.35
3	14–16	644	160	21	4.33
4	12–14	360	89	14	7.79
5	11–13	674	168	8	4.13
6	12–22	1796	448	13	1.55
7	21–23	831	207	12	3.35
8	22–24	383	95	7	7.29
9	24–26	784	195	23	3.55
10	25–27	287	71	9	9.76
11	26–28	106	26	1	26.65
12	36–38	312	78	8	8.88
13	35–37	253	63	7	11.00
14	34–36	215	53	15	13.08
15	32–34	220	55	12	12.60
16	32–42	2776	693	26	1.00
17	42–44	198	49	9	14.14
18	44–46	289	72	11	9.63
19	45–47	244	60	7	11.55
20	46–48	256	64	7	10.83
21	LBW	556	139	21	4.99

Post hoc analysis—pairwise comparisons

An overview of the levels of statistical significance for the Bonferroni-adjusted p values for model pairwise comparisons with McNemar's test is shown in Fig. 1. A complete table of p values is found in Appendix.

Failed models

MLPs without batch normalization (0, 1, 2) failed to train (top-1 accuracy 0.05–0.10, $F1$ -score 0–0.01 and AUC equal to 50). Therefore, a statistically significant difference with any other model ($p < 0.001$ for all pairs) was to be expected, except for members of the same group.

Comparisons among advanced models

No statistically significant differences were observed among advanced group models (18–35), regardless of bridging layer or data augmentation strategy ($p > 0.05$ for all pairs). Advanced models had top-1 accuracy ranging from 0.81 to 0.89, $F1$ -score between 0.71–0.86 and AUC of 0.86–0.94.

Comparison of baseline models against advanced models

Baseline models 3, 4, 6, 7, 12 and 13 showed no statistically significant differences against the advanced group, except for pairs 4–18 ($p < 0.05$), 4–22 ($p < 0.05$) and 4–31 ($p < 0.05$), indicative of comparable high performance (top-1 accuracy 0.77–0.87, $F1$ -score 0.70–0.81 and AUC 0.84–0.90).

In contrast, models 5, 10, 11 and 17 showed statistically significant differences with all models in the advanced group ($p < 0.001$ for all pairs) as a result of poor performance (top-1 accuracy 0.46–0.61, $F1$ -scores 0.33–0.48, AUC 0.67–0.75).

Models 14, 16 showed statistically significant differences with the advanced group except model 24 ($p < 0.001$ against models 18, 21, 22, 26–35, $p < 0.01$ against models 19, 23, 24 and $p < 0.05$ against model 20). Their top-1 accuracy was 0.69 and 0.68, $F1$ -scores 0.60 and 0.61, respectively, and AUC 0.81 for both.

Comparisons among baseline models

Models 11 and 17 (top-1 accuracy 0.46, $F1$ -score 0.33, AUC 0.67 and top-1 accuracy 0.47, $F1$ -score 0.36, AUC 0.68) had the lowest performance among all models (except for models 0, 1, 2) and showed a statistically significant difference

Table 4 Training history and model performance evaluation on the test dataset. AUC: area under the curve

Model	Training duration	Epochs	Per-sample prediction time	Test loss	Test top-1 accuracy	Test precision	Test recall	Test <i>F1</i> -score	Test AUC
0	0:47:35	32	0.01	15.31	0.05	0.00	0.05	0.00	0.50
1	1:12:28	32	0.01	14.51	0.10	0.00	0.05	0.01	0.50
2	1:12:17	32	0.01	14.51	0.10	0.00	0.05	0.01	0.50
3	2:29:51	100	0.01	0.67	0.80	0.76	0.75	0.74	0.87
4	3:44:45	100	0.01	0.62	0.77	0.73	0.70	0.70	0.84
5	3:42:03	100	0.01	1.07	0.61	0.50	0.50	0.47	0.74
6	1:01:57	26	0.01	0.54	0.81	0.78	0.75	0.75	0.87
7	4:47:11	100	0.01	0.32	0.87	0.81	0.81	0.81	0.90
8	4:45:29	100	0.01	0.68	0.75	0.71	0.70	0.67	0.84
9	3:05:34	100	0.01	0.69	0.72	0.59	0.59	0.58	0.79
10	4:21:05	100	0.01	1.19	0.61	0.49	0.52	0.48	0.75
11	4:09:30	97	0.01	1.48	0.46	0.36	0.37	0.33	0.67
12	4:36:47	100	0.02	0.63	0.84	0.83	0.79	0.78	0.89
13	2:38:00	52	0.02	0.54	0.81	0.79	0.80	0.74	0.90
14	4:09:10	83	0.02	0.84	0.69	0.71	0.63	0.60	0.81
15	3:19:08	100	0.01	0.73	0.75	0.72	0.70	0.66	0.84
16	3:01:18	66	0.01	0.78	0.68	0.68	0.63	0.61	0.81
17	4:25:10	100	0.01	1.47	0.47	0.49	0.39	0.36	0.68
18	2:39:04	27	0.03	0.39	0.87	0.82	0.85	0.82	0.92
19	3:15:59	31	0.03	0.40	0.84	0.82	0.80	0.78	0.90
20	7:19:47	71	0.03	0.46	0.83	0.78	0.77	0.76	0.88
21	2:40:13	29	0.03	0.34	0.86	0.79	0.79	0.78	0.89
22	5:48:16	58	0.03	0.32	0.88	0.82	0.85	0.82	0.92
23	6:23:39	65	0.03	0.47	0.83	0.78	0.75	0.75	0.87
24	1:19:44	26	0.03	0.56	0.81	0.77	0.73	0.71	0.86
25	3:17:16	40	0.03	0.42	0.84	0.81	0.82	0.79	0.90
26	4:58:44	62	0.03	0.47	0.84	0.81	0.81	0.79	0.91
27	2:14:40	48	0.02	0.48	0.89	0.83	0.86	0.82	0.93
28	3:16:23	41	0.02	0.32	0.88	0.85	0.82	0.81	0.91
29	5:11:48	66	0.02	0.47	0.86	0.81	0.80	0.79	0.90
30	4:07:29	26	0.12	0.32	0.87	0.84	0.84	0.81	0.92
31	5:54:21	31	0.11	0.30	0.89	0.85	0.87	0.84	0.93
32	7:33:58	48	0.11	0.40	0.85	0.80	0.83	0.79	0.91
33	4:45:28	29	0.12	0.42	0.85	0.78	0.79	0.77	0.89
34	5:16:05	32	0.12	0.28	0.89	0.85	0.88	0.86	0.94
35	7:47:02	48	0.13	0.37	0.86	0.77	0.77	0.76	0.88

with any other baseline model ($p < 0.001$ except pairs 11–10, 17–10 where $p < 0.01$ and 11–5, 17–5 where $p < 0.05$).

Models 5 and 10 also showed a statistically significant difference compared to all baseline models ($p < 0.001$ with models 3, 4, 6, 7, 12, 13 and $p < 0.05$ with models 8, 11, 15, 17) as well as low performance (top-1 accuracy 0.61, *F1*-score 0.47, AUC 0.74 and top-1 accuracy 0.61, *F1*-score 0.48, area AUC 0.75), excluding pairs 5–9, 5–10, 5–14, 5–16, 10–5, 10–9, 10–14, 10–16 where no statistically significant difference was observed.

Models 8, 9, 14, 15, 16 (top-1 accuracy 0.68–0.75, *F1*-score 0.58–0.67, AUC 0.79–0.84) showed a statistically significant difference in the proportion of errors against models 11, 17 ($p < 0.001$) and the highest performing baseline model 7 ($p < 0.001$ with models 9, 14, 16 and $p < 0.01$ with models 8, 15). In addition, model 8 showed a statistically significant difference with models 5 and 10 ($p < 0.05$). Models 14, 15, 16 also showed significant differences in the pairs 14–12 ($p < 0.01$), 14–13 ($p < 0.01$), 15–5 ($p < 0.05$), 15–10 ($p < 0.05$) 16–12 ($p < 0.001$), 16–13 ($p < 0.05$). This

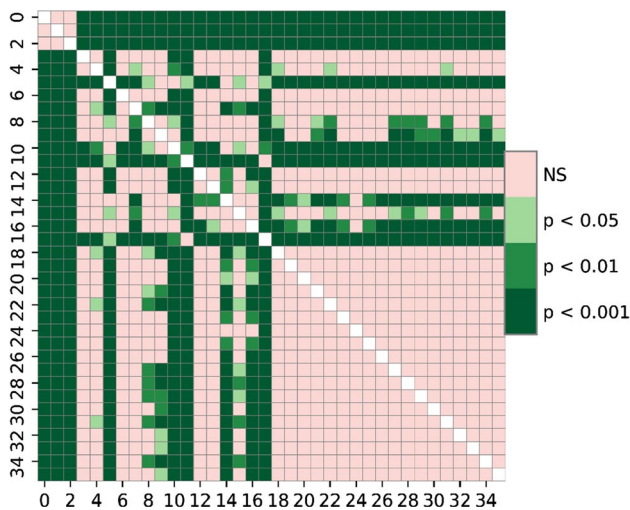


Fig. 1 Significance levels of multiple pairwise comparisons among models with McNemar's test, adjusted with the Bonferroni correction. NS: non-significant

implies that this subset of models lies in the middle in terms of performance of the baseline model group.

Baseline models 3, 4, 6, 12, 13 (top-1 accuracy 0.77–0.84, *F1*-score 0.70–0.79, AUC 0.84–0.90) showed a statistically significant difference with models 5, 10, 11, 17 ($p < 0.001$ except pair 4–10 where $p < 0.01$), as well as pairs 4–7 ($p < 0.05$), 12–14 ($p < 0.01$), 12–16 ($p < 0.001$), 13–14 ($p < 0.01$), 13–16 ($p < 0.05$), while no statistically significant difference was observed between them.

Baseline model 7 was the highest performing model of the baseline group (top-1 accuracy 0.87, *F1*-score 0.81, area AUC 90). There was a statistically significant difference with every other model in the baseline group ($p < 0.001$ with models 5, 9, 10, 11, 14, 16, 17, $p < 0.01$ with models 8, 15 and $p < 0.05$ with model 4) except for the high-performing baseline models 3, 6, 12, 13, with which no statistically significant difference was observed.

Discussion

This study investigated whether neural networks could classify the anatomical region of intraoral radiographs based solely on image data, and the influence of their architectural elements. According to our findings, it is feasible with an expected top-1 accuracy of 80–90%, when trained with small datasets.

Intraoral images usually depict very similar anatomical features, especially when they are part of a series from the same patient, where a significant amount of overlap is expected. Therefore, an understanding of the relative positioning of the depicted structures is essential for their

classification. On the contrary, deep learning classifiers have mostly been studied with images of discrete objects against different backgrounds. Bearing that in mind, testing with a multitude of architectures was deemed appropriate as some of them could fail to address this unique challenge.

Multilayer perceptrons (MLPs)

MLPs without normalization failed to train. However, the introduction of batch normalization [19] allowed training comparable to that of advanced models. This finding indicates that input and layer normalization may allow non-convolutional architectures to perform adequately in radiographic image classification tasks. However, these models performed poorly alongside data augmentation, indicating a dependency on input images with little variation.

Baseline convolutional models

Most baseline convolutional models achieved adequate performance. Batch normalization [19] improved training and performance in some architectures. A typical data augmentation strategy resulted in better training, but the introduction of an aggressive strategy led to deteriorating performance.

Advanced models

All advanced models had comparable performance and outperformed most baselines. However, due to the strict nature of the Bonferroni p value adjustment, subtle differences may not be elucidated, while different errors may occur on the test dataset for each model. Furthermore, some models trained irregularly, as it is evident by their learning curves.

Data augmentation

Data augmentation is a common technique for generating more samples from small datasets, and it is considered vital to prevent overfitting, especially in models with large capacities. It can also function as a measure of a model's resilience to improper input, as represented by the aggressive strategy.

Baseline model performance was severely downgraded with aggressive data augmentation, although some were robust or even benefited from subtle transformations. Most advanced group models were resistant to aggressive data augmentation and capable of managing degraded images.

Global average pooling

Using a global average pooling layer [20] in baseline models significantly limited their performance, a finding directly associated with the resulting marked reduction in model capacity.

On the other hand, parameter reduction seemed to favor the advanced group, where it showed no detrimental effects even when applied concurrently with aggressive data augmentation. A regularization effect could also be observed in many models' learning curves.

Clinical recommendations

Based on the above, the use of models from the advanced group trained with aggressive data augmentation and a Global Average Pooling layer [20] as a bridge between the feature extractor and the classifier parts is recommended.

Choosing an architecture should be a compromise between other parameters, such as resources and time availability. In a clinical setting where long waiting times can be a major disadvantage, a smaller and faster model such as MobileNetV2 might be more useful, while the Inception-ResnetV2-based architecture could be better suited for tasks such as database maintenance.

Limitations

Our models demonstrate all limitations inherent in most deep learning models. They were developed empirically on natural images, which prior studies have shown not to be the same as X-ray imaging feature-wise [28]; they lack theoretical explanation for their performance and have high computational costs not favoring experimentation. They also lack outcome justification and are vulnerable against specific images containing irrelevant features able to trigger a predictable output (adversarial samples), making trusted input a necessity.

Such models supposedly require large amounts of data to train. Building large data sets with medical data is a laborious undertaking with serious ethical, legal and financial considerations, while for smaller datasets containing well-structured images, solutions other than neural networks may be more efficient. However, in this study good performance was achieved with limited data.

In addition to the fore-mentioned, a significant limitation is the introduction of dataset bias in model outputs. Sources of dataset bias in this study were the use of only high-quality radiographs out of a single modality, which were diagnostic and contained mostly tooth or prosthesis imaging (rarely depicted exclusively bone structures or contained instruments). Equally important is the fact that these models are only able to replicate the classification criteria applied by the evaluator (evaluator bias). The above could significantly limit our models from performing consistently under different conditions. Training with a multitude of diverse datasets could partially resolve this issue. Currently, model

generalizability to datasets with different specifications is not guaranteed.

Another limitation is class imbalance, a direct result of uneven clinical demand and the retrospective nature of the study. Creating a perfectly balanced dataset would require either excluding a large portion of our dataset, with a possible decline in performance, or a prospective study design, exposing subjects to X-rays for study purposes and breaching the ALARA principle without clearly determined benefits. In view of the above, using a weighted loss function seems to be an appropriate compromise. However, all models exhibited their worst performance in the underrepresented classes.

Recommendations for further research

Further gains in performance are anticipated by training with a larger and more balanced dataset, by using loss functions that count in the inherent order of anatomical regions, with the transfer learning and fine-tuning technique and with the combined use of model ensembles.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11548-021-02321-4>) contains supplementary material, which is available to authorized users.

Author contributions Kyventidis Nikolaos involved in conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft, visualization. Angelopoulos Christos involved in resources, data curation, writing—review and editing, supervision, project administration.

Funding This study has received no external funding.

Availability of data and material Not publicly available.

Code availability Code will be published in the following link: <https://github.com/nkyventidis/intraoral-radiograph-classifier>.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest, financial or otherwise.

References

1. [UNSCEAR] United Nations Scientific Committee on the Effects of Atomic Radiation, Sources and effects of ionizing radiation: United Nations Scientific Committee on the Effects of Atomic Radiation (2008) UNSCEAR report to the General Assembly, with scientific annexes. United Nations, New York, p 2010
2. [FDA] Food and Drug Administration, Dental Radiography: Doses and Film Speed (2017). <https://www.fda.gov/radiation-emitting-products/nationwide-evaluation-x-ray-trendsnext/dental-radiography-doses-and-film-speed>. Accessed 25 June 2020

3. Horner K, Rushton VV, Tsiklakis K, Hirschmann P, Stelt PF, Glenny A, Velders X, Pavitt S (2004) European guidelines on radiation protection in dental radiology: the safe use of radiographs in dental practice. *Radiat Prot* 136:11–17
4. Horner K (2012) Radiation protection in dental radiology. In: Proceedings of international conference 3–7 December 2012, International Atomic Energy Agency, Bonn, Germany, 2012
5. [NEMA] National Electrical Manufacturers Association (2005) Digital Imaging and Communications in Medicine, Supplement 60: Hanging Protocols, 2005
6. [NEMA] National Electrical Manufacturers Association (2019) Digital Imaging and Communications in Medicine, PS3.17 2019 - Explanatory Information, 2019
7. C. Langlotz, B. Allen, B. Erickson, J. Kalpathy-Cramer, K. Bigelow, T. Cook, A. Flanders, M. Lungren, D. Mendelson, J. Rudie, G. Wang, K. Kandarpa (2019) A roadmap for foundational research on artificial intelligence in medical imaging: from the (2018) NIH/RSNA/ACR/The academy workshop. *Radiology* 291:781–791
8. Deng J, Dong W, Socher R, Li J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition. IEEE, pp. 248–255
9. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, vol 25. Curran Associates, Inc., pp 1097–1105
10. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg A, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. *Int J Comput Vis IJCV* 115:211–252
11. Rawat W, Wang Z (2017) Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput* 29:2352–2449
12. Litjens G, Kooi T, Bejnordi B, Setio A, Ciompi F, Ghafoorian M, Van Der Laak J, Van Ginneken B, Sánchez C (2017) A survey on deep learning in medical image analysis. *Med Image Anal* 42:60–88
13. Bossuyt P, Reitsma J, Bruns D, Gatsonis C, Glasziou P, Irwig L, Lijmer J, Moher D, Rennie D, de Vet H, Kressel H, Rifai N, Golub R, Altman D, Hooft L, Korevaar D, Cohen J (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology* 277(2015):826–832
14. Mongan J, Moy L, Kahn CE (2020) Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell* 2:e200029. <https://doi.org/10.1148/ryai.2020200029>
15. Simonyan K, Zisserman A, (2015) Very deep convolutional networks for large-scale image recognition, In: 3rd Int. Conf. Learn. Represent. ICLR 2015 San Diego CA USA May 7–9 2015 Conf. Track Proc
16. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L (2018) MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, 2018, pp 4510–4520
17. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017) Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of thirty-first AAAI conference on artificial intelligence. AAAI Press, pp 4278–4284.
18. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: 2016 IEEE conference on computer vision and pattern recognition CVPR, 2016, pp 2818–2826
19. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, 2015, pp 448–456
20. Lin M, Chen Q, Yan S (2014) Network In Network, In: 2nd Int. Conf. Learn. Represent. ICLR 2014 Banff AB Can. April 14–16 2014 Conf. Track Proc
21. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15:1929–1958
22. Kingma D, Ba J (2015) Adam: a method for stochastic optimization, In: 3rd Int. Conf. Learn. Represent. ICLR 2015 San Diego CA USA May 7–9 2015 Conf. Track Proc
23. Nair V, Hinton G (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of 27th international conference on international conference on machine learning, Omnipress, Madison, WI, USA, 2010, pp 807–814
24. [NCHS] National Center for Health Statistics (1999) National Health and Nutrition Examination Survey Data., U.S. Department of Health and Human Services, Hyattsville, MD, 1999.
25. Dietterich T (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 10:1895–1923
26. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
27. García S, Herrera F, Shawe-Taylor J (2008) An extension on “statistical comparisons of classifiers over multiple data sets” for all pairwise comparisons. *J Mach Learn Res* 9:2677–2694
28. Chow L, Paramesran R (2016) Review of medical image quality assessment. *Biomed Signal Process Control* 27:145–154

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.