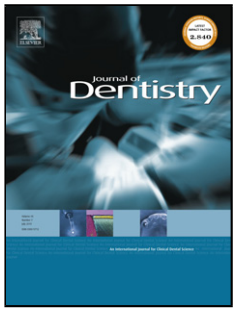
Journal Pre-proof



Convolutional neural networks for dental image diagnostics: A scoping review

Falk Schwendicke, Tatiana Golla, Martin Dreher, Joachim Krois

PII: S0300-5712(19)30228-3

DOI: https://doi.org/10.1016/j.jdent.2019.103226

Reference: JJOD 103226

To appear in: *Journal of Dentistry*

Received Date: 14 August 2019

Revised Date: 28 October 2019

Accepted Date: 1 November 2019

Please cite this article as: Schwendicke F, Golla T, Dreher M, Krois J, Convolutional neural networks for dental image diagnostics: A scoping review, *Journal of Dentistry* (2019), doi: https://doi.org/10.1016/j.jdent.2019.103226

Convolutional neural networks for dental image diagnostics: A Scoping Review

Short title: Convolutional neural networks for image diagnostics

Falk Schwendicke[1*], Tatiana Golla[1*], Martin Dreher[1], Joachim Krois[1]

[1] Department of Operative and Preventive Dentistry, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Germany

*joint first authors

Corresponding author: Prof. Dr. Falk Schwendicke MDPH, Department of Operative and Preventive Dentistry Charité – Universitätsmedizin Berlin Campus Benjamin Franklin, Aßmannshauser Str. 4-6 14197 Berlin, Tel.: (+49) 30 450 562 556; Fax: (+49) 30 450 7562 556, e-Mail: falk.schwendicke@charite.de

Abstract

Objectives: Convolutional neural networks (CNNs) are increasingly applied for medical image diagnostics. We performed a scoping review, exploring (1) use cases, (2) methodologies and (3) findings of studies applying CNN on dental image material.

Sources: Medline via PubMed, IEEE Xplore, arXiv were searched.

Study selection: Full-text articles and conference-proceedings reporting CNN application on dental imagery were included.

Data: Thirty-six studies, published 2015-2019, were included, mainly from four countries (South Korea, United States, Japan, China). Studies focussed on general dentistry (n=15 studies), cariology (n=5), endodontics (n=2), periodontology (n=3), orthodontics (n=3), dental radiology (2), forensic dentistry (n=2) and general medicine (n=4). Most often, the detection, segmentation or classification of anatomical structures, including teeth (n=9), jaw bone (n=2) and skeletal landmarks (n=4) was performed. Detection of pathologies focused on caries (n=3). The most commonly used image type were panoramic radiographs (n=11), followed by periapical radiographs (n=8), Cone-Beam CT or conventional CT (n=6). Dataset sizes varied

between 10–5,166 images (mean 1,053). Most studies used medical professionals to label the images and constitute the reference test. A large range of outcome metrics was employed, hampering comparisons across studies. A comparison of the CNN performance against an independent test group of dentists was provided by seven studies; most studies found the CNN to perform similar to dentists. Applicability or impact on treatment decision was not assessed at all.

Conclusions: CNNs are increasingly employed for dental image diagnostics in research settings. Their usefulness, safety and generalizability should be demonstrated using more rigorous, replicable and comparable methodology.

Clinical significance: CNNs may be used in diagnostic-assistance systems, thereby assisting dentists in a more comprehensive, systematic and faster evaluation and documentation of dental images. CNNs may become applicable in routine care; however, prior to that, the dental community should appraise them against the rules of evidence-based practice.

Keywords: Artificial Intelligence; CNNs, Dentistry; Diagnostics; Evidence-based Dentistry; Images

Introduction

In medicine, the application of techniques from the realm of artificial intelligence (AI) and specifically computer vision is increasingly common. Deep learning, using convolutional neural networks (CNNs) has been demonstrated to have remarkable potential to assist doctors in fields like dermatology (to detect skin cancer), opthalmology (to detect and discriminate different types of retinopathy) and radiology (for example, to detect abnormalities in chest x-rays) [1-3]. While there is evidence for CNNs having the potential to detect pathologies, but also anatomical structures on medical images with similar or even higher accuracy as medical professionals, there is also indication for potential bias in the algorithms underlying these applications, with possible risks of limited robustness and generalizability [4].

CNNs perform tasks like (1) detecting structures (deciding if an organ or, in dentistry, a tooth etc. is present on an image) or pathologies (e.g. nodes in the lung, caries lesions on teeth), (2) segmenting them (e.g. identifying the exact shape of the organ, tooth, or pathology on an image) and (3) classifying them (e.g. being able to separate the left from the right kidney or to label each tooth in a dentition, distinguishing acquired and developmental enamel white spot lesions from each other).

To do so, CNNs, which are a particular type of machine learning models, are trained with data that is provided in form of pairs of imagery data and a corresponding outcome for the image (so-called supervised learning). Depending on the task, this outcome may be a unique label such as affected or not affected, or a list of labels (e.g. teeth names or pathologies) and related areas on the image that capture the corresponding structure or pathological region. By providing the data in an iterative and repeated manner, the CNN's internal structure, more specifically its model parameters (referred to as model weights), are optimized to represent the statistical structure of the input data and its mapping to a particular label [5]. The training process is repeated until the incremental improvement in the model's predictive power allow to map an input image to a particular label (Fig. 1).

To perform such training, a reference test is needed to provide a label. This reference test can be constituted via a "gold standard" (e.g. histological assessment for carious lesions). More often, though, a label is provided by human experts who "annotate" images, e.g. depending on the task they decide if something (structure, pathology) is present, where exactly it is or what shape it has. This procedure of course defines "the truth" according to the annotator, with a single annotator being only limitedly accurate. Hence, in many cases, multiple labels are provided (multiple experts annotate the same image), and these multiple votes are synthesized to define the reference test [6].

The trained CNN is then validated, ideally against an independent dataset (which the CNN has never seen before), and performance metrics can be derived. While in medicine measures like accuracy (the % of correct detections per all labels), the area-under-the receiver-operating-curve (AUC) or the associated sensitivity (correct positive detections per all positive labels) and specificity (correct negative detections per all negative labels) are frequently employed, technical disciplines usually use other metrics, like recall (a synonym for sensitivity), precision (also known as positive predictive value), the F1-score (the harmonic mean of precision and recall) or similarity scores (indicating if labelled and predicted image segments overlap).

In dentistry and oral medicine, a vast number of images are obtained each year. In most European countries, for example, dental radiography (including panoramic, bitewing, periapical and cephalometric radiographs) is likely to account for the majority of all radiographs taken, with an estimated mean of 250-300 dental images being taken per 1000 individuals in 2010 [7]. Considering that further imagery (photographs, 3-D-surface scans, fluorescence images etc.) are employed, too, there seems great potential for CNN application. However, so far, it has not been systematically assessed which fields have employed CNNs, and which applications are in the focus. It is also not clear which methods (imagery types, reference tests, CNN architectures, outcome metrics) are common, and how CNNs currently perform when compared against human experts. Having such information could guide future applied

research for AI techniques in dentistry, but also support the strive for methodological standardization. Moreover, it could help to assess if CNNs are actually useful for clinical practice.

We aimed to perform a scoping review to compile studies on CNNs for dental image diagnostics. Scoping reviews are also executed in a systematic, replicable manner, but usually focus to identify knowledge gaps, scope the literature, clarify concepts or assess research conduct and rigor and thereby, eventually, inform research, educational and clinical policy and priorities [8, 9].

Material and Methods

For this study, we defined a scoping review as a study which "aims to map the literature on a particular topic or research area and provide an opportunity to identify key concepts, gaps in the research; and types and sources of evidence to inform practice, policymaking, and research" [10]. The conduct of this review follows the Arksey and O'Malley framework, modified by Levac [11, 12], omitting the consultation step. Reporting follows the PRISMA statement [13].

Our review questions were (1) For which applications have CNNs been applied in dentistry and oral medicine?, (2) Which methods are used in these studies to establish datasets, develop, train and test the model, and report the model performance?, and (3) What were the findings of these studies when comparing CNN performance against that of human examiners?.

Search

The systematic literature search was conducted for studies published 2000 to May 10th 2019; the search was performed by one reviewer (TG). We systematically searched three databases, namely Medline via PubMed, IEEE Xplore and arXiv:

- Medline is the most widely used medical database publishing mainly journal articles.
- IEEE (Institute of Electrical and Electronics Engineers) Xplore is a digital library for journal articles, conference proceedings, technical standards, and related materials on computer science, electrical engineering and electronics, and related fields. The database allows access to more than 4.5 million technology documents, more than 193 peer-reviewed journals and more than 1,700 global conferences (https://ieeexplore.ieee.org/Xplore/home.jsp).
- arXiv.org is an electronic archive of electronic preprints for research articles of scientific topics such as physics, mathematics, computer science, and statistics, among others. In particular in the fast-developing fields of machine learning and artificial intelligence

newest developments and findings are presented via electronic preprints. Many articles are published later in more traditional journals and are accepted for being submitted to highly renowned conferences in that particular field, such as NIPS (Neural Information Processing Systems) or ICML (International Conference on Machine Learning). Notably, arXiv is not peer reviewed, but there are moderators for each area, who review the submissions [14].

Searching these different sources accounted for the differences in publication culture across disciplines (clinical, technical sciences). No language or date restrictions were applied. A three-pronged search strategy, combining the technique of interest (CNN/deep learning etc.), the image materials (radiographs etc.) and the field of interest (dentistry, or specific topics therein) was applied. The search sequence for Medline can be found in the appendix and was adopted for the other database/repository.

One reviewer (TG) screened the titles and abstracts for potential eligibility, and a second reviewer (FS) reviewed and double-checked this step. Full-texts were then retrieved and screened in duplicate. Cross-referencing from bibliographies and further hand searches were performed.

Study selection

We included full articles as well as conference proceedings, both peer reviewed and not (peer review is not automatically and immediately performed when studies are archived publicly and thereby made available, for example in repositories, which is common in the technical sciences), of original studies reporting on the application of CNN on dental or oral medicine imagery. We did not further define image materials nor did we specify what task (detection, segmentation, classification) was to be performed or which overall objective was underlying the studies. Study setting (in vitro, clinical etc.) was also not defined, as we aimed to capture studies of different nature to be as sensitive as possible. As a result, we included diagnostic accuracy studies using a CNN against some kind of reference test, mainly on routine imagery data material (i.e. retrospective studies).

Data extraction and analysis

The following variables were extracted:

(1) General study details (title, primary author, journal or conference, date of publication)
(2) Study characteristics (field and application, image type, number of images, index and reference test, if available additional comparator tests)
(3) Outcomes and outcome metrics.

(4) Findings.

Extraction was performed by one reviewer (TG) and extractions were discussed with a second reviewer (FS) in detail. Extraction was first performed in verbatim and study fields, applications, image types, and reference tests eventually summarized categorically after establishing useful classes for synthesis. Further analysis was performed narratively, as well as by descriptive statistics and tabularization. Meta-analysis was not attempted given the heterogeneity in settings, conduct of index and reference tests, and outcome metrics (see below). Note that no formal heterogeneity test was applied, as outcomes were largely not combinable even for attempting such testing. Also, as mentioned, no further consultation with stakeholders was sought at this point.

Results

Search and study selection

We identified 323 records, around half of them from PubMed and the other half from technical databases and repositories, without any duplicative entries across databases (Fig. 2). A total of 36 studies were eventually included (Table 1); details on the studies excluded at the full-text step can be found in the appendix Table S1 (most excluded studies did not employ CNNs).

Included studies

All included studies were published between 2015 and 2019, with the number of publications increasing each year (Fig. 3). Studies were published across 14 countries, with the majority of first authors being located in South Korea (n=8 studies), the United States (n=6), Japan (n=5) or China (n=5). They were diagnostic accuracy studies, largely performed on routinely collected imagery material (rather than prospectively set-up datasets).

Study fields and applications

The studies addressed clinical problems in the following disciplines (Table 2); general dentistry (n=15 studies), cariology (n=5), endodontics (n=2), periodontology (n=3), orthodontics (n=3), dental radiology (2), forensic dentistry (n=2) and general medicine (n=4). The application found most often was the detection, segmentation or classification of anatomical structures, including teeth (n=9), jaw bone (n=2) and skeletal landmarks (n=4). Three further studies aimed to detect structures, but also restorations and pathologies. Further applications were the detection of carious lesions (n=3), biofilm classification (n=2), and the classification of endodontic treatment conditions or results (n=2). Detecting and classifying periodontal inflammation or bone loss (n=2), and detection and classification of facial features (n=1) was also performed. Two studies

investigated image quality enhancement in the field of dental radiology using CNNs. Additionally, CNNs were applied for forensic reasons as well as in general medicine (detection of osteoporosis, atherosclerotic carotid plaques and sinusitis).

Datasets and image types

Dataset sizes varied between 10 and 5,166 images (mean 1,053 images). A separate test dataset was used in ten studies; sixteen studies performed testing within the overall dataset (largely via cross-validation). In eight studies, only the total number of images and no information regarding training or testing data was provided. The most commonly used image type were panoramic radiographs (used in 11 studies), followed by periapical radiographs (n=8), Cone-Beam CT (CBCT) or conventional CT (n=6), photos (n=3), bitewing or cephalometric radiographs (both n=2) and 3D-mesh, quantitative light fluorescence or biomarker images (each n=1). One study assessed both CBCT-Scans and panoramic radiographs.

CNN architectures

Nearly half of all studies reported to have used individually constructed CNN architectures, including variations of pretrained CNNs. If assessing pretrained CNNs, AlexNet (n=7 studies) was used most often, followed by VGG16 (n=5), ResNet (n=4), U-net (n=3), VGG19 and GoogLeNet (each n=2), and V-net or DenseNet (each n=1).

Reference and comparative test

In all but four studies information about the reference test was provided. In fourteen studies, the reference test was established by medical professionals, mainly dentists (n=7 studies), dentists and radiologists (n=3), radiologists (n=2), dentists and students (n=1) and physicians (n=1). In ten further studies, the reference test was established by “experts” (n=6), “dental experts” (n=3) or “radiology experts” (n=1), with no further description of their qualification. In five studies, the reference test was established “manually” without further information. One study used micro-CT images to define the reference test, two studies used labelled images from databases without specifying who or how this labelling had been performed.

The number of human annotators who assessed each image varied between 1 (n=6 studies), 2 (n=5), 3 (n=3), 4 (n=2) and 5 (n=2). In eighteen studies no information regarding the number of annotators was provided. In 17 studies the performance of the CNN was compared to other deep learning techniques or CNN architectures as second (comparative) test, but not to human observers. In twelve studies the performance of the CNN was compared to human observers. In five of these the human observers were identical with the reference test.

Outcome metrics

In nine studies, the only reported outcome metric was accuracy. In nine further studies, further metrics were used in addition to accuracy, namely the AUC (n=4 studies), sensitivity/specificity (n=4), precision/recall (n=3), Dice similarity coefficient (n=2), F1-score, Hausdorff distance or boundary error (each n=1). Three studies chose F1-score as their only outcome metric, while further two combined it with precision/recall reporting. A range of further metrics (including mean absolute error, mean average best overlap, recall-overlap rate, mean squared error, peak signal-to-noise ratio, structure similarity index, information fidelity criterion and noise quality measure, all n=1 studies) were also employed.

Findings

The model performance varied widely between studies, tasks, and used metrics (see above). The most frequent tasks were tooth classification, with a reported mean accuracy of 0.77 to 0.98, and detection of carious lesions, with a reported mean accuracy of 0.82 to 0.89. From the twelve studies comparing CNNs against human examiners, three found the CNN to outperform humans, seven found similar performance of CNN and humans, and in one study the CNN performed worse than the humans.

Discussion

CNNs show highly promising performance when assessing imagery material in medicine. Their application in healthcare could allow more comprehensive, reliable and accurate image assessment and disease detection, thereby facilitating more effective, efficient and safer care [15]. The present study assessed, using a systematic review, the application of CNNs in dental and oral medical research, and evaluated the employed methodology and resulting rigour of the included studies. Based on our review, a number of findings emerged, which need to be discussed.

The use of CNNs in dental research has been continuously growing since the first published application in 2015, but an “explosion” (as can be seen in medicine, where 42 PubMed entries on “CNN” are available in 2015, and over 800 in only the first half of 2019) is not in sight. It is conceivable that dentistry is experiencing a lag time, following medicine, of several years. This may come to the detriment of patients in case these new technologies provide benefits. However, it may also allow to define methodological standards and introduce rigour early on, whilst other disciplines have already a larger body of evidence which suffers from risk of bias [16].

Only about half of all identified study reports were retrieved from Medline, which is usually a major source for medical and dental research. Instead, a significant proportion of studies were

published in technology-focused databases/repositories. The latter, while not being necessarily peer-reviewed, come with the advantage of swift publication and open access to the community. Researchers may prioritize such publication strategies given the pace of technological advances and the associated turnover of (technological) methods. It remains unclear, though, if the validity of studies provided in only in repositories (without peer, but possibly public review) is lower than those from established journals [17]. Generally, it is obvious that a significant proportion of the identified research is not originated in the dental, but the technical sciences, and also published differently.

The main image material used in the included studies were radiographs, specifically panoramic radiographs or (to a lesser degree) CBCT. This may be grounded in the general relevance of radiographs for dental diagnostic and decision-making. For panoramic images, it may further reflect the frequency this image material is employed in daily clinics (in Germany, for example, nearly 8 million panoramic radiographs are taken each year) [18]. Moreover, both panoramic and in particular CBCT images are complex and burdensome to assess given the vast number of structures and possible pathologies to assess or screen for. Hence, virtual assistance is of high interest for the practitioner.

The main task CNNs had so far been tested for were structure identification and segmentation; only few studies tackled the detection of pathologies, where one would expect the greater benefit given the difficulties in dental image diagnostics and the resulting limited reliability and accuracy [19, 20]. It can be assumed that so far, researchers focused on the more basic task of identifying structures (most often teeth), with this being the basis of a more complex detection system, where pathologies are then detected and assigned to the identified tooth in a subsequent step.

A range of CNN architectures have been employed. Most often, individual or at least individualized networks were used, possibly as most pretrained networks could not be leveraged on the rather small amount of data in most studies (the average sample size was around 1,000 images, while most pretrained CNNs were developed on tens or hundred thousands of images, oftentimes not in the medical arena, where sparse data are generally a problem) [21]. Notably, dental researchers adopted network architectures published in 2015 or 2016, indicating a possible lag in the adoption of latest architectures. However, it is also conceivable that very deep CNNs, published recently [22], could not be leveraged using the available datasets.

A major aspect of CNN application is how the reference test is established. In most included studies, human annotators were employed; in many cases even a single annotator had been labelling images. This may be acceptable when structures, especially those easy to identify by a human expert (like teeth) are labelled. However, when training CNNs to detect pathologies,

a higher number of experts should be used to overcome the limitations of a single expert. It should be highlighted that if this is done, though, different approaches as to how derive one single reference label are available (different majority vote schemes, for example), the impact of which is not clear [23]. The definition of a reference test should ideally follow a standard which the dental community, jointly with technical sciences, should establish.

The performance of CNNs was reported in a range of metrics, most of which come with limitations. (1) The accuracy can be highly misleading when class imbalances are present, as is often the case with pathologies [24]. Apical lesions, for example, are not found frequently; the tooth level prevalence seems to be below 10% (only every tenth or even fewer teeth are affected in most populations). In such cases, accuracies of 90% or more can be achieved only be guessing the majority class ("a tooth is not affected by an apical lesion"), without this being clinically useful. (2) The AUC is only partially informative and not useful for decision making, where over- or under-detection are not equally important. Hence, the sensitivity or specificity should be additionally reported. (3) It should be highlighted that most studies did not display metrics which are conferring information towards clinical benefit (some studies reported on the PPV=precision, while negative predictive values were rarely reported), or assessed in how far using CNNs aid detection, diagnosis and decision making. From most studies, it remains also unclear if using the CNNs is more or less accurate than the assessment by a single dentist. Notably, some studies provided such second comparative test group of human observers; often, tough, the observations from these humans were also used as labels for training the CNN. In such case, humans will always perform remarkably well given themselves being (at least part of) the reference test. There was also no information on the applicability, efficiency, or safety of CNNs when applied. Overall, our study calls for the definition of a standard set of outcomes and performance metrics to be used in CNN studies in dentistry and oral medicine.

Generally, there seems to be a high need for standardizing methodology to increase robustness, comparability and generalizability. Such standardization should start with the steps used to recruit image material and prepare the dataset. Detailed and comprehensive reporting on modelling and validation strategies should also be demanded. An open access standard test dataset to be used for each image type (panoramic images, bitewings etc.) could further ensure that CNN performance can be reliably assessed and compared.

This review comes with a number of limitations. First, given this being a scoping review, our review question was less defined than when using a conventional systematic review, and our outcomes are rather broad. Moreover, the search was performed by one reviewer, and also the extraction was only doublechecked by a second reviewer. This was decided pragmatically, as our inclusion criteria were broad and during screening, studies were retained rather than excluded if in doubt. Second, study findings were compiled narratively and according to a

systematization which we devised post-hoc; using a different strategy to synthesize data may have emerged further or different findings. Third, only a limited number of databases were searched, assuming these to cover the vast majority of publications. Overall, and despite these limitations, our study is informative as to which fields of applications are currently in the focus when applying CNNs in dentistry and oral medicine, which methodologies are used and what limitations they suffer from. Our findings should be used to guide and further develop future studies in the field.

Conclusions

The present review, with the discussed limitations, compiled and appraised studies employing CNNs on dental or oral medicine imagery. Overall, a wide range of applications, so far largely on radiographs, was found, with the current focus being detection and segmentation of anatomical, not pathological structures. We identified significant heterogeneity in the reported methods and outcomes, and it can be assumed that the applied methods partially determine the performance, but more so the external and internal validity of the study findings. There seems to be great need for standardizing the methodology in CNN studies in dentistry and oral medicine, including the definition of a standard set of outcomes and outcome metrics relevant to stakeholders. Aspects like applicability and impact on decision-making should be considered, and performance outcomes replicated on independent datasets before adopting CNNs for clinical practice.

Conflict of interest

This study was funded by the authors and their institution. There are no conflicts of interest to report.

References

[1] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542(7639) (2017) 115-118.
[2] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P.C. Nelson, J.L. Mega, D.R. Webster, Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, Jama 316(22) (2016) 2402-2410.
[3] M. Mazurowski, M. Buda, A. Saha, M. Bashir, Deep learning in radiology: an overview of the concepts and a survey of the state of the art, arXiv:1802.08717v1 (2018).
[4] G. Marcus, Deep Learning: A Critical Appraisal, arXiv.org https://arxiv.org/abs/1801.00631 (2018).
[5] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521(7553) (2015) 436-44.
[6] T. Walsh, Fuzzy gold standards: Approaches to handling an imperfect reference standard, Journal of dentistry 74 Suppl 1 (2018) S47-s49.
[7] HSE, Population Dose from Dental Radiology: 2010 2010.
[8] Z. Munn, M.D.J. Peters, C. Stern, C. Tufanaru, A. McArthur, E. Aromataris, Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach, BMC medical research methodology 18(1) (2018) 143.
[9] H.L. Colquhoun, D. Levac, K.K. O'Brien, S. Straus, A.C. Tricco, L. Perrier, M. Kastner, D. Moher, Scoping reviews: time for clarity in definition, methods, and reporting, Journal of clinical epidemiology 67(12) (2014) 1291-1294.
[10] H.M.L. Daudt, C. van Mossel, S.J. Scott, Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework, BMC medical research methodology 13(1) (2013) 48.
[11] D. Levac, H. Colquhoun, K.K. O'Brien, Scoping studies: advancing the methodology, Implementation Science 5(1) (2010) 69.
[12] H. Arksey, L. O'Malley, Scoping studies: towards a methodological framework, International Journal of Social Research Methodology 8(1) (2005) 19-32.
[13] D. Moher, L. Shamseer, M. Clarke, Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement, Syst Rev 4 (2015).
[14] M. McKinney, arXiv.org, Reference Reviews 25(7) (2011) 35-36.
[15] L. Liu, J. Xu, Y. Huan, Z. Zou, S.C. Yeh, L. Zheng, A Smart Dental Health-IoT Platform Based on Intelligent Hardware, Deep Learning and Mobile Terminal, IEEE journal of biomedical and health informatics (2019).
[16] K.B. Nielsen, M.L. Lautrup, J.K.H. Andersen, T.R. Savarimuthu, J. Grauslund, Deep Learning-Based Algorithms in Screening of Diabetic Retinopathy: A Systematic Review of Diagnostic Performance, Ophthalmology. Retina 3(4) (2019) 294-304.
[17] D.M. Herron, Is expert peer review obsolete? A model suggests that post-publication reader review may exceed the accuracy of traditional peer review, Surgical endoscopy 26(8) (2012) 2275-80.
[18] KZBV, KZBV Jahrbuch 2017, 2017.
[19] F. Schwendicke, M. Tzschoppe, S. Paris, Radiographic caries detection: A systematic review and meta-analysis, Journal of dentistry 43(8) (2015) 924-33.
[20] Z.Z. Akarslan, M. Akdevelioglu, K. Gungor, H. Erten, A comparison of the diagnostic accuracy of bitewing, periapical, unfiltered and filtered digital panoramic images for approximal caries detection in posterior teeth, Dento maxillo facial radiology 37(8) (2008) 458-63.
[21] G. Chartrand, P.M. Cheng, E. Vorontsov, M. Drozdzal, S. Turcotte, C.J. Pal, S. Kadoury, A. Tang, Deep Learning: A Primer for Radiologists, Radiographics : a review publication of the Radiological Society of North America, Inc 37(7) (2017) 2113-2131.
[22] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, Recent advances in convolutional neural networks, Pattern Recognition 77 (2018) 354-377.
[23] J. Krois, T. Ekert, L. Meinhold, T. Golla, B. Kharbot, A. Wittemeier, C. Dorfer, F. Schwendicke, Deep Learning for the Radiographic Detection of Periodontal Bone Loss, Scientific reports 9(1) (2019) 8495.

[24] J. Krois, C. Graetz, B. Holtfreter, P. Brinkmann, T. Kocher, F. Schwendicke, Evaluating Modeling and Validation Strategies for Tooth Loss, Journal of dental research (2019) 22034519864889.

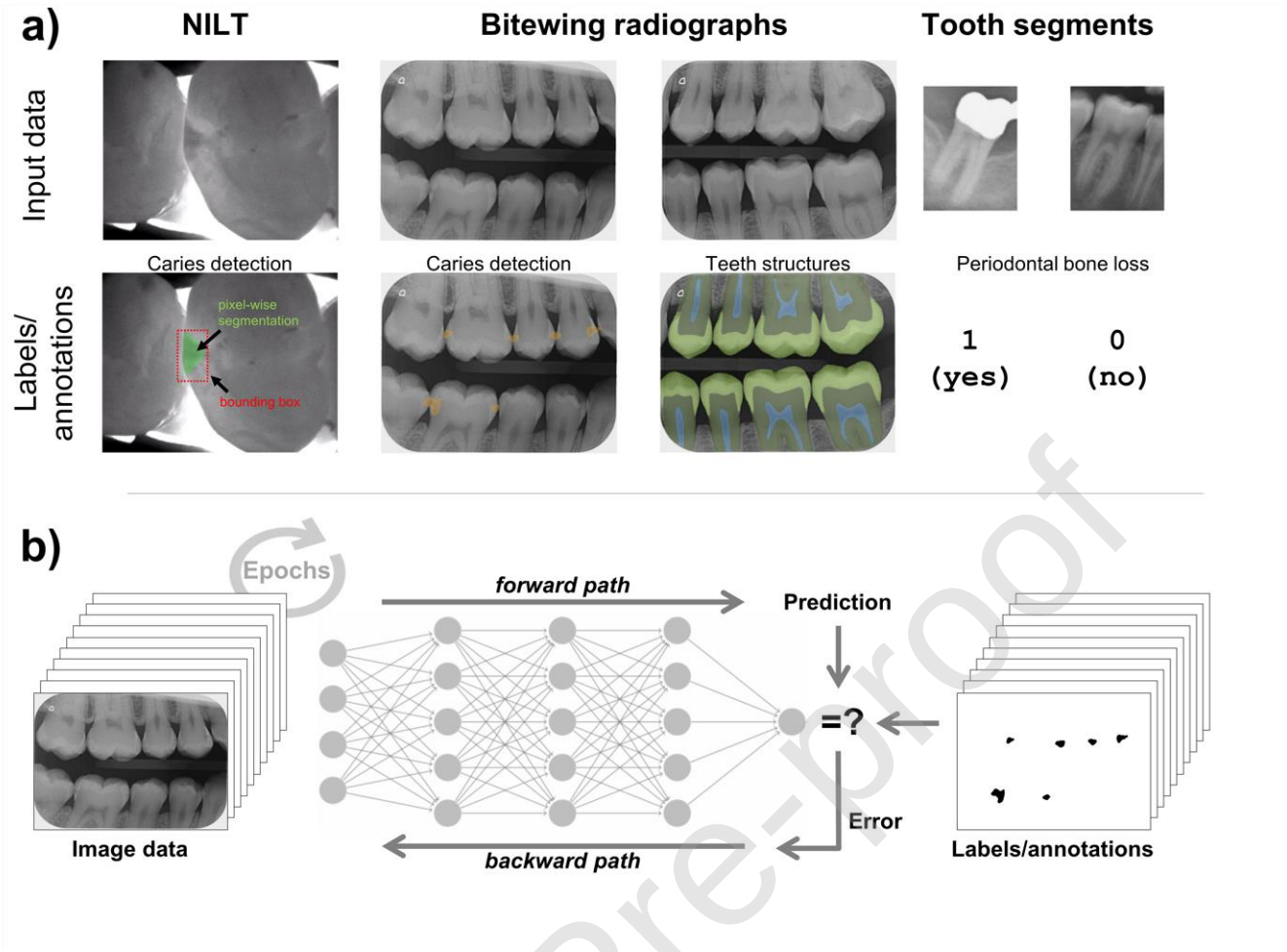


Figure 1: a) Potential labelling strategies for different dental image modalities such as (from left to right) near infrared-light transillumination (NILT), bitewing radiographs, or tooth segments extracted from panoramic radiographs. Labels and/or annotations are provided by experts in the field, in a pixels-wise fashion, in form of bounding boxes or as (binary) image class labels (e.g. 0 and 1, corresponding to a positive and negative class). b) Training of a neural network, demonstrated on the bitewing caries detection image from (a). In an iterative and repeated procedure (referred to as epochs) image data is passed through a network (from left to right, so-called forward path). The model's output, its prediction, is compared to a ground truth (e.g. the labels and annotations provided by human experts). Wrong outcomes cause the error to be propagated backwards (backward path) through the network, with small changes being applied to the model weights in response (by the so-called back propagation algorithm). Over the period of many epochs, the model weights are optimized to represent the statistical structure of the input data and its mapping to a particular label or annotations.

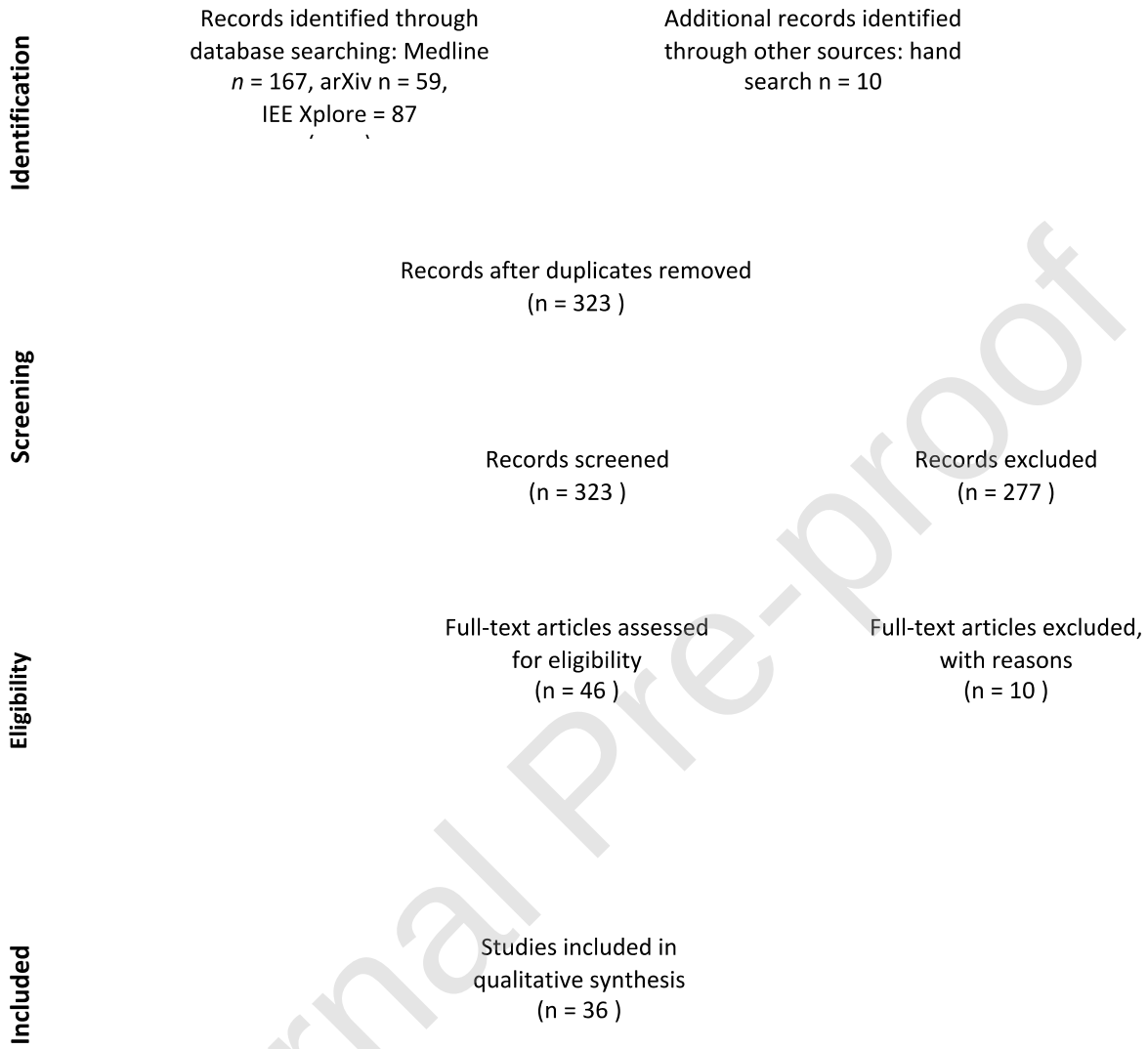


Figure 2: Flowchart of the search

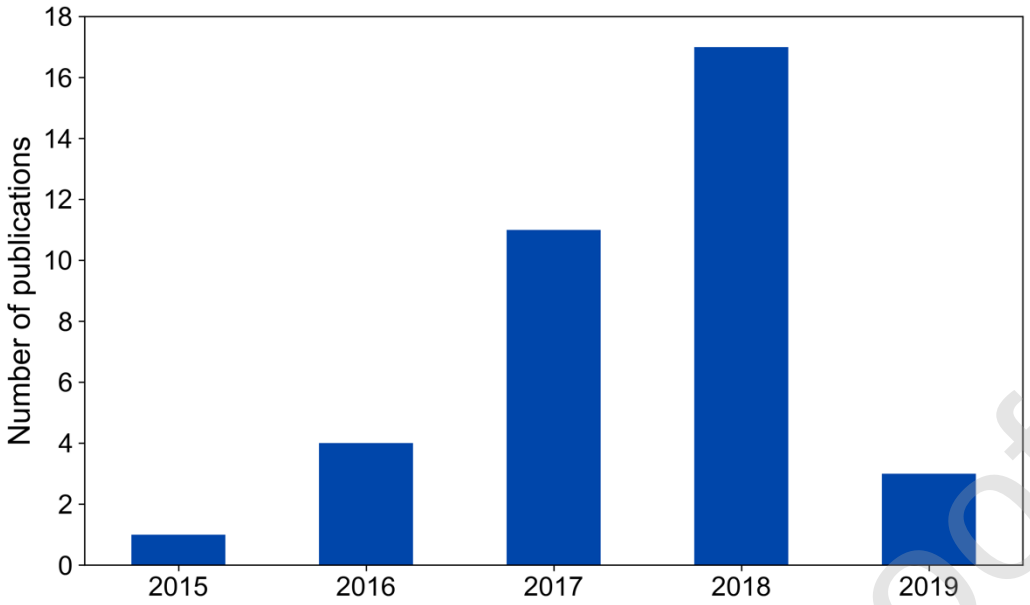


Figure 3: Number of included studies, per year of publication.

Table 1: Included studies. References can be found in the appendix.

Study	Country, Year	Field of dentistry	Topic	Image type	Images total	CNN architecture	Annotators (n)	CNN performance	Outcome metrics
Ronneberger [1]	Germany, 2015	General dentistry	Tooth segmentation	Bitewing	39	u-net	Medical doctors (2)	n.a.	Dice similarity, precision
Imangaliyev [2]	Netherlands, 2016	Cariology	Biofilm classification	QLF-Images	427	ResNet	Dental experts (n/a)	+	F1-score
Choi [3]	Korea, 2016	Cariology	Caries detection	Periapical	475	Individual CNN	Experts (n/a)	+	F1-score, precision, recall
Yu [4]	Korea, 2016	Forensic medicine	Tooth and sex classification	Panoramic	980	Individual CNN	n.a.	n.a.	accuracy
Eun [5]	Korea, 2016	General dentistry	Tooth segmentation	Periapical	600	Individual CNN	Manually (n/a)	+	Mean Average Best overlap, recall-Overlap rate
De Tobel [6]	Belgium, 2017	Forensic medicine	Age estimation	Panoramic	400	AlexNet	(n/a) (2-3)	=	AUC
Prajapati [7]	India, 2017	General dentistry	Dental pathology detection	Periapical	251	VGG16	Dentists + radiologists (n/a)	n.a.	accuracy
Miki [8]	Japan, 2017	General dentistry	Tooth detection and classification	CBCT	52	AlexNet	Manually (n/a)	n.a.	accuracy
Miki [9]	Japan, 2017	General dentistry	Tooth classification	CBCT	52	AlexNet	Manually (n/a)	n.a.	accuracy
Murata [10]	Japan, 2017	Orthodontics	Facial feature classification	Photos	704	VGG19	Dentists + students (n/a)	+	accuracy
Lee [11]	Korea, 2017	Orthodontics	Cephalometric landmark detection	Cephalometric image	300	Individual CNN	Experts (2)	-	Euclidean distance, accuracy
Oktay [12]	Turkey, 2017	General dentistry	Tooth detection and classification	Panoramic	100	AlexNet	Dentistd (1)	n.a.	accuracy
Srivastava [13]	USA, 2017	Cariology	Caries detection	Bitewing	3000	Individual CNN	Dentists (1)	+	recall, precision, F1-score
Yauney [14]	USA, 2017	Cariology	Biofilm classification	Whitelight image, Biomarker image	96	VGG16	Dental experts (n/a)	n.a.	accuracy, AUC, precision-recall

Rana [15]	USA, 2017	Periodontology	Periodontal inflammation detection	Photos	405	Individual CNN	Dental experts (1)	=	AUC, recall,, precision
Arik [16]	USA, 2017	Orthodontics	Cephalometric landmark detection	Cephalometric image	400	Individual CNN	Experts (2)	+;-	accuracy
Egger [17]	Austria, 2018	General dentistry	Mandible segmentation	CT	10	VGG-16net, FCN-32s, FCN-16s, FCN-8s	Experts (2)	n.a.	accuracy
Jader [18]	Brasil, 2018	General dentistry	Tooth segmentation	Panoramic	1500	ResNet	Prelabeled dataset (n/a)	+	accuracy, F1-Score, precision, recall, specificity
Du [19]	China, 2018	Dental radiology	Image quality enhancement	Panoramic	5166	Individual CNN	n.a.	n.a.	Mean/max. absolute errors
Yang [20]	China, 2018	Endodontics	Quality evaluation	Periapical	196	Individual CNN	Dentist (1)	=	recall, precision, F1-score
Zhang [21]	China, 2018	General dentistry	Tooth detection and classification	Periapical	1000	VGG16,	n.a.	n.a.	F1-score, precision, recall
Xu [22]	China, 2018	General dentistry	3D tooth classification	Surface point-cloud data (scan)	1200	Individual CNN	Prelabeled dataset (n/a)	n.a.	acccuracy, boundary error
Wirtz [23]	Germany, 2018	General dentistry	Tooth segmentation	Panoramic	24	u-net	Manually (n/a)	+	accuracy, DICE, F1-score, precision, recall, specificity
Hatvani [24]	Hungary, 2018	Dental radiology	Image quality enhancement	CBCT	17	u-net, Subpixel	Micro-CT (n/a)	+	Mean squared error, peak-signal-to-noise ratio , structure similarity index, information fidelity criterion, noise quality measure

Kats [25]	Israel, 2018	General medicine	Atherosclerotic carotid plaques detection	Panoramic	65	ResNet	Dentists + radiologists (2)	n.a.	accuracy, AUC, sensitivity, specificity
Hiraiwa [26]	Japan, 2018	Endodontics	Root morphology classification	CBCT, Panoramic	760	AlexNet, GoogleNet	Radiologists (n/a)	better	accuracy, sensitivity, specificity, AUC
Murata [27]	Japan, 2018	General medicine	Sinusitis detection	Panoramic	800	AlexNet	Dentists + radiologists (4)	similiar (radiologists), better (dentist)	accuracy, sensitivity, specificity, AUC
Lee [28]	Korea, 2018	Cariology	Caries detection	Periapical	3000	GoogLeNet Inception v3	Dentists (4)	n.a.	accuracy, AUC, sensitivity, specificity
Lee [29]	Korea, 2018	General medicine	Osteoporosis detection	Panoramic	1268	Individual CNN	Radiologists (n/a)	similiar	accuracy, AUC, F1-Score
Lee [30]	Korea, 2018	Periodontology	Bone loss detection	Periapical	1740	VGG19 + Individual CNN	Dentists (3)	similiar (Premolar), worse	accuracy, AUC
Zakirov [31]	Russia, 2018	General dentistry	Tooth and tooth structure detection and classification	CBCT	1791	V-Net, DenseNet	Specialists (4-5)	n.a.	accuracy, AUC
Torosdagli [32]	USA, 2018	General dentistry	Mandible segmentation and anatomical landmark detection	CBCT	300	Tiramisu (U-Net + DenseNET), Zhang's improved u-net	Experts (3)	n.a.	accuracy, sensitivity, specificity, dice similarity coefficient, Hausdorff distance
Chu [33]	USA, 2018	General medicine	Osteoporosis detection	Panoramic	108	AlexNet	Dentists (1)	n.a.	accuracy
Chen [34]	China, 2019	General dentistry	Tooth detection and classification	Periapical	1250	ResNet	Dentists (1)	similiar	recall, precision, intersection-over-union
Tuzoff [35]	Russia, 2019	General dentistry	Tooth detection and classification	Panoramic	1574	VGG16	Experts (5)	similiar	sensitivity, specificity
Joo [36]	Korea, 2019	Periodontology	Periodontal inflammation detection	Photos	1843	Individual CNN	n.a.	n.a.	accuracy

Abbreviations: AUC area-under-the-curve. CBCT Cone Beam CT. CNN convolutional neural network. QLF quantitative light-induced fluorescence. n/a not available/applicable. +/-/= significantly better, worse, or not significantly different performance.

Table 2: Application of CNNs in different fields of dentistry and oral medicine

Field	N articles	Application
General dentistry	15	Detection and classification of teeth (8), detection of oral structures including pathologies (2), segmentation of teeth (2), teeth and pathologies (1) and mandible (1), mandible segmentation and anatomical landmark detection (1)
Cariology	5	Detection of carious lesion detection (3), classification of biofilm (2)
General medicine	4	Detection of osteoporosis (2), atherosclerotic carotid plaques (1), sinusitis (1)
Orthodontics	3	Detection of cephalometric landmarks (2), classification of facial features (1)
Periodontology	3	Detection of periodontal inflammation (2), bone loss (1)
Endodontics	2	Classification of root canal fillings (1) and root morphology (1)
Dental radiology	2	Image quality enhancement (2)
Forensic medicine	2	Age estimation (1), classification of sex (1)