Journal of Dentistry 157 (2025) 105728

Contents lists available at ScienceDirect

Journal of Dentistry

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

From inconsistent annotations to ground truth: Aggregation strategies for annotations of proximal carious lesions in dental imagery

Vanessa Klein [a,b], Martha Büttner [a], Gerd Göstemeyer [c], Sarina Rolle [a], Antonin Tichy [b,d], Falk Schwendicke [a,b,**], Noah F. Nordblom [a,*]

[a] Department of Oral Diagnostics, Digital Health and Health Services Research, Charité – Universitätsmedizin Berlin, Berlin, Germany
[b] Conservative Dentistry and Periodontology, LMU University Hospital, LMU Munich, Munich, Germany
[c] Department of Operative and Preventive Dentistry, Charité–Universitätsmedizin Berlin, Berlin, Germany
[d] Institute of Dental Medicine, First Faculty of Medicine of the Charles University, Prague, Czech Republic

ARTICLE INFO

Keywords:
Artificial Intelligence
Fuzzy labels
Ground truth labels
Label aggregation methods
Annotation strategies
Near-Infrared Light Transillumination (NILT)
Radiographs
Caries diagnosis

ABSTRACT

Objectives: Annotating carious lesions on images is challenging. For artificial intelligence (AI) applications, the aggregation of heterogeneous multi-examiner annotations into one single annotation (*e.g.* via majority voting, MV) is usually needed. We assessed different aggregation strategies for multi-examiner annotations of primary proximal carious lesions on orthoradial radiographs and Near-Infrared Light Transillumination (NILT) images.
Methods: A total of 1007 proximal surfaces from 522 extracted posterior teeth were assessed by five dentists. Histological analysis provided the gold standard. Surfaces were classified as (1) sound, (2) enamel lesion or (3) dentin lesion. Four label aggregation strategies - MV, Weighted Majority Voting (WMV), Dawid-Skene (DS), and multi-annotator competence estimation (MACE) - were applied to unimodal (radiographs, NILT) and multimodal (combined) datasets. The area under the receiver operating characteristic curve (AUROC) was the primary outcome metric.
Results: According to the gold standard, 637 (63 %) surfaces were sound, 280 (28 %) showed carious lesions limited to the enamel, and 90 (9 %) showed lesions extending into the dentin. For radiographs, aggregation using MACE outperformed MV, WMV and DS significantly across all lesion depths ($p < 0.002$). For NILT, MACE significantly outperformed MV across all lesion depths ($p < 0.001$) and DS for enamel and dentin lesions ($p \leq 0.002$). In the multimodal dataset, DS outperformed the other label aggregation strategies across all lesion depths significantly ($p < 0.05$).
Conclusions: The commonly applied MV may be suboptimal. There is a need for informed application of specific aggregation strategies, depending on the dataset characteristics.
Clinical significance: Most AI applications for dental image analysis are trained on a single annotation, usually resulting from aggregated multi-examiner annotations of each image. However, since these annotations are usually aggregated in an *in vivo* setting where no definitive ground truth is available, the choice of aggregation strategy plays a crucial role.

1. Introduction

Numerous studies have focused on the accuracy of dentists in detecting carious lesions on radiographic images, demonstrating limited accuracy and, more so, significant variability between different annotators, even after calibrating them [1–4]. The accuracy of these annotators is frequently assessed against a reference test, in many cases histological assessment ("gold standard") of extracted teeth [5].

Recently, dentistry experienced a surge of applications of artificial intelligence (AI), with one particularly prominent focus being image analysis, and one highly common task being carious lesion detection on radiographs [2,3,6]. Notably, these applications are trained on

* Corresponding authors at: Department of Oral Diagnostics, Digital Health and Health Services Research, Charité – Universitätsmedizin Berlin, Assmannshauser Straße 4-6, 14197, Berlin, Germany.
** Corresponding authors at: Department of Conservative Dentistry and Periodontology, LMU University Hospital, LMU Munich, Goethestraße 70, Munich, 80 336, Germany.
E-mail addresses: Falk.Schwendicke@med.uni-muenchen.de (F. Schwendicke), noah.nordblom@charite.de (N.F. Nordblom).

https://doi.org/10.1016/j.jdent.2025.105728
Received 7 February 2025; Received in revised form 26 March 2025; Accepted 29 March 2025
Available online 30 March 2025

thousands of retrospectively collected, de-identified images, where establishing a hard reference test (such as histology) is not possible [7]. In this case, and to address the described variability of single annotators, multiple annotators are used to establish the reference test against which the AI model is trained and tested [7]. Notably, multiple annotators frequently produce so-called fuzzy (heterogenous) annotations [8] which pose challenges in deciding how to best handle them [9]. While there are methods available to train deep learning models directly on such fuzzy data [9,10], the most common approach in the dental domain currently taken is to aggregate the heterogenous, fuzzy annotations into a single annotation (*e.g.* of a carious lesion being present on a tooth or surface, or not being present), and then train and test the AI model on the resulting, annotated images.

A range of strategies for this aggregation exist, for example majority voting (MV, using the annotation which has been provided most often), weighted majority voting (WMV, where certain factors alter the weight of a given annotation), or probabilistic approaches such as Dawid-Skene algorithm (DS) or multi-annotator competence estimation (MACE) [11]. Bayesian methods, like MACE, utilize probabilistic models and likelihood functions to explicitly account for uncertainty, which can be particularly valuable in dentistry, where creating reliable datasets from fuzzy labels benefits from modeling uncertainty as a probabilistic function.

We here explored the accuracy of different annotation aggregation strategies, namely MV, WMV, DS, and MACE, for detecting primary proximal carious lesions on two common imaging modalities, orthoradial radiographs and Near-Infrared-Light-Transillumination (NILT) imagery, against the gold standard for caries detection (histology). We further assessed the impact of lesion depth on this accuracy. Our study setup allowed us to gauge the generalizability of our findings across image types and detection task. We hypothesized that there were no significant differences in accuracy (measured via the area under the receiver operating characteristic curve, AUROC) between different aggregation strategies.

2. Materials and methods

2.1. Study samples

608 extracted teeth (Fig. 1), both sound and showing carious lesions of different depths, were collected. Exemplary depictions of these lesions are shown in Fig. 2. All teeth were obtained with informed consent under an ethics-approved protocol (Ethics Committee of the Charité - Universitätsmedizin Berlin, EA4/102/14). During the study, the teeth were stored at 4 °C in a 0.5 % chloramine-T solution, which was exchanged every four weeks. As for 86 teeth, the restorative status (extended restorations and other defects on both proximal surfaces) was impeding histological evaluation, the number of teeth used for the study was eventually reduced to 522. The number of surfaces included in this study was reduced to 1007, as some teeth could only contribute with one proximal surface due to restorations (exclusion criterion) (Fig. 1). Prior to testing, the teeth were cleaned and polished using a scaler (S204S9E2, Hu-Friedy, Chicago, USA) and polishing paste (Proxyt fein, Ivoclar Vivadent, Schaan, Liechtenstein).

To generate models on which imagery could be obtained in a simulated clinical fashion, teeth (four premolars and four molars per model) were mounted in lower jaw models constructed from wax (Typodont, Dentaurum, Ispringen, Germany). Both upper and lower jaw premolars and molars were assembled in a single model. Negative silicone molds (Z-Dupe Catalyst A and Base B, HS-Dubliermasse, Henry Schein, Melville, USA) of the wax models were made, and a cold-curing clear epoxy resin (Epo-Thin 2, Buehler, Lake Bluff, USA) was used to produce resin models. Wax was then applied to shape the gingiva (ROSA Modellierwachs, ORBIS-Dental, Offenbach, Germany). The models were stored at 4 °C in a 0.5 % chloramine-T solution until further analysis. Models were stored for up to 7 months prior to radiograph acquisition and up to 14 months prior to NILT imaging.

2.2. Imagery

The generation of imagery followed established protocols [12]. Each model generated two images per imaging modality (radiographs, NILT), representing the left and right sides. Radiographs were taken using a

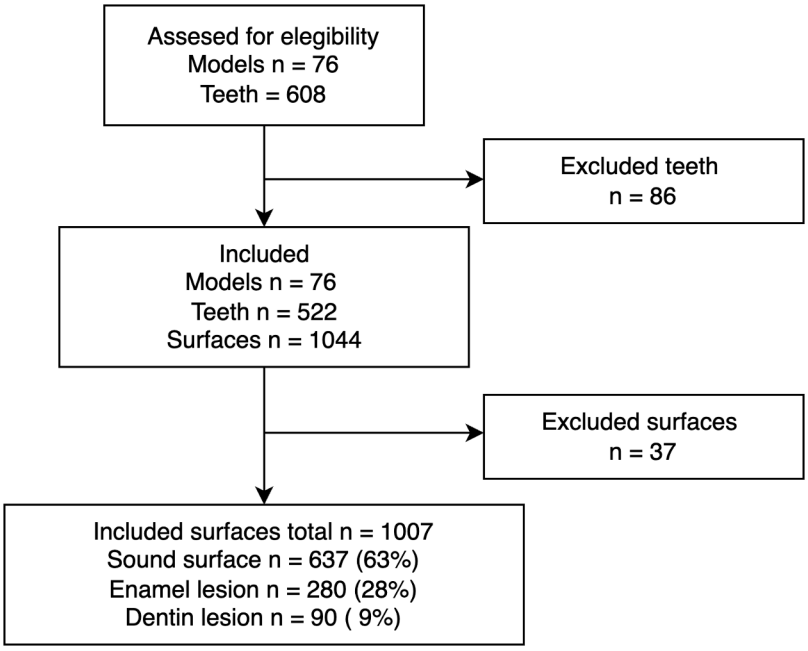

Fig. 1. Study samples. From the total assessed sample (n = 608 teeth assembled in 76 models), 522 teeth were included in the study, enabling the histological evaluation of 1044 tooth surfaces. Of these surfaces, 37 exhibited restorations or other defects and were thus unsuitable for the detection of primary carious lesions. Consequently, the number of surfaces evaluated was reduced to 1007.

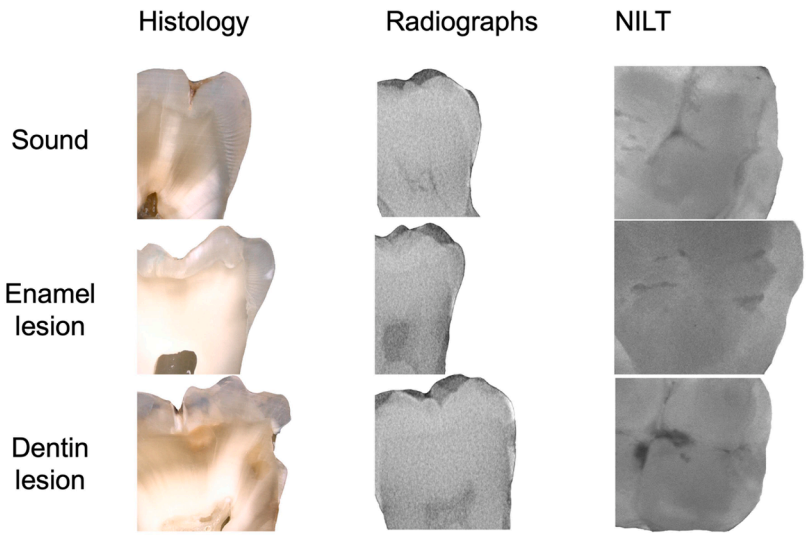

Fig. 2. Representative examples of dental surfaces with varying lesion depths, assessed through histology, radiographs, and Near-Infrared Light Transillumination (NILT).

digital X-ray sensor plate (VistaScan, DÜRR DENTAL SE, Bietigheim-Bissingen, Germany) and a radiation source (Heliodent Plus, Sirona Dental Systems GmbH, Bensheim, Germany) with an exposure time of 0.04 s, operating at 70 kV and 7 mA. To ensure consistent imaging conditions, a mounting aid (LEGO System A/S, Billund, Denmark) was used to position each model identically, as described elsewhere [13]. A 15-mm-thick plexiglass plate was placed between the radiation source and the model to simulate soft tissue scattering. NILT images were captured using DIAGNOcam (KaVo, Biberach, Germany) and the associated software (KaVo Integrated Desktop version 2.4.1.6821). The models were mounted in a dummy head (Phantomkopf P-6, Frasaco, Tettnang, Germany), with the operating light switched off as per the manufacturer's instructions. NILT images were taken vertically to the occlusal surfaces of each tooth, as described before [14].

2.3. Gold standard

Histological examination served as gold standard for assessing the presence and depth of carious lesions. Teeth were first removed from the resin models and sectioned in a mesio-distal direction using a band saw (Exakt 300, Exakt Apparatebau, Norderstedt, Germany). The section was made close to the area where the carious lesion on each tooth was clinically presumed to be the deepest. The sectioned teeth were then embedded in a cold-curing clear methyl-methacrylate-based resin (Technovit 4004, Kulzer, Hanau, Germany) and mounted on glass slides. As each tooth could present with two surfaces (mesial, distal) being carious on both image types, we first assessed the depth of the clinically more extended (deeper) lesion and then progressively ground and polished the tooth halves down with a micro grinder (EXAKT 400 CS, Norderstedt, Schleswig-Holstein) with a series of sandpapers (grit size 1200, 2400, and 4000), ensuring all relevant areas of the tooth were adequately evaluated.

To assess the depth of carious lesions and verify whether other lesions had been reached, photographs were taken every 300 micrometers using a digital light microscope (VHX-5000, Keyence, Osaka, Japan) at 20× magnification. Surfaces were classified as sound, enamel lesion, or dentin lesion. Throughout the study, the annotators had no access to the gold standard data.

2.4. Annotation process

The digital imagery was annotated by five dentists with varying levels of professional experience (2 - 21 years). While some were employed in university clinics, others worked in private dental practices. All dentists received a detailed written guideline on how to annotate carious lesions on proximal tooth surfaces (mesial, distal). Lesions within the enamel, *i.e.* to the dentinoenamel junction (DEJ) were labeled as enamel caries, while lesions beyond DEJ were labeled as dentin caries. A custom-built annotation software was used [15]. This classification approach enabled comparability between imaging modalities.

2.5. Annotation aggregation strategies

We evaluated four different annotation aggregation strategies to generate a final, aggregated label for each surface: (1) MV, (2) WMV, (3) the Dawid-Skene algorithm (DS) and (4) multi-annotator competence estimation (MACE) described below. A summary of the advantages and disadvantages of each aggregation strategy is provided in Table 1.

Majority voting (MV): The most often occurring class (sound surface, enamel lesion, dentin lesion) from all five annotators was used to generate the aggregated annotation. For each surface, the class that

Table 1
Advantages and disadvantages of label aggregation methods compared in this study.

Method	Advantages	Disadvantages
Majority Voting (MV)	Simple and easy to implement consensus prediction. Encapsulates the collective judgment of all annotators.	Does not account for annotator reliability. Can be biased if there are more unreliable annotators.
Weighted Majority Voting (WMV)	Incorporates annotator reliability into the decision process. Utilizes ground truth data to calculate weights.	Requires a subset of data with known ground truth annotations. May be affected by noisy annotations if weights are not accurate in the subset.
Dawid-Skene Algorithm (DS)	Iteratively estimates true annotations and annotator reliability. Reduces the impact of noisy or unreliable annotators.	Computationally intensive due to iterative processes. Assumes that annotator reliability can be accurately estimated.
Multi-Annotator Competence Estimation (MACE)	Models annotator reliability and spamming behavior. Effective at handling noisy annotations.	Computationally intensive due to iterative process. May require more data to accurately estimate parameters.

received the highest number of votes was selected as the final label. This method resulted in a consensus prediction based on the majority agreement of the annotators. The final aggregated annotations represented a single annotation for each surface, encapsulating the collective judgment of the five annotators.

Weighted majority voting (WMV): For weighted majority voting, the dataset was split in two parts: a small subset which contained each lesion depth 20 times and where ground truth labels are available, and the remaining dataset. For the small subset, each annotator's probability of correctly assigning each lesion depth is calculated. The model then determines the final aggregated label by weighting annotators' contributions based on their reliability, summing these probabilities across all annotators, and selecting the label with the highest total probability.

Dawid-Skene algorithm (DS): A Dawid-Skene algorithm implementation was used to estimate the true labels from fuzzy data [16,17]. In DS, different weights are assigned to each annotator based on their reliability, ensuring that more competent and reliable annotators have a greater influence on the final label. This is achieved through an iterative Expectation-Maximization (EM) process, which alternates between estimating the most likely true label and updating annotator reliability scores. As the iterations progress, annotator accuracy is dynamically adjusted, improving label aggregation and reducing the impact of noisy or unreliable annotators. Once the model converges, the final prediction is made based on the estimated labels and the refined annotator reliability scores.

Multi-annotator competence estimation (MACE): MACE is a Bayesian model designed to estimate true labels from noisy annotations while accounting for annotator reliability and spamming behavior [17,18]. MACE models initial annotator credibility and the likelihood of random guessing (spamming behavior) in the dataset. These parameters are iteratively updated based on observed data, allowing the model to adjust for annotator inconsistencies. While MACE is more effective at handling noisy annotations, it is computationally more intensive than DS due to the additional modelling of spamming tendencies.

2.6. Statistical evaluation

The level of statistical significance was set to < 0.05. Fleiss' Kappa was calculated to assess inter-annotator agreement [19]. Individual annotators were evaluated by calculating their sensitivity, specificity, F1-score and AUROC. Mann-Whitney-U tests were performed to test significant differences between imaging modalities and lesion depths on account of the absence of a normal distribution. For annotation aggregation, AUROC was the primary evaluation metric. Metrics were calculated for each imaging dataset separately (radiographs, NILT) as well as for the combined, multimodal dataset. Aggregated labels were derived using the methods described above. We further performed an analysis stratified according to lesion depth. The normality of the data was assessed visually using Q-Q plots and statistically tested with the Shapiro-Wilk test. Since most subgroups violated the assumption of normality, we utilized the Kruskal-Wallis tests, followed by Dunn's post-hoc test with Holm adjustment for multiple comparisons, to evaluate the annotation aggregation strategies.

2.7. Software

Python (version 3.12.2) along with several supporting libraries was utilized for data preparation, cleaning and analysis. *Pandas* (version 2.2.1) and *NumPy* (version 1.26.4) were used for data cleaning and numerical computations, while *Seaborn* (version 0.12.2) was employed for data visualization. *Scikit-learn* (version 1.4.1.post1) was used to calculate annotation metrics, and *SciPy* (version 1.15.0) was used to perform the Shapiro-Wilk test and Kruskal-Wallis tests. *Scikit-posthocs* (version 0.11.2) was used for post-hoc Dunn's test with adjusted p-values for multiple comparisons using the Holm method. *Statsmodels* (version 0.14.4) was used to compute Fleiss' Kappa values. *Crowd-kit* (version 1.4.1) was utilized for annotation aggregation.

3. Results

According to the histological examination, 637 (63 %) surfaces were sound, 280 (28 %) showed carious lesions limited to the enamel, and 90 (9 %) showed lesions extending into dentin (Fig. 1). True positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) for the annotators are presented in Table 2 as summary statistics and descriptive annotation metrics are presented in Table 3. Fleiss' Kappa values for the five annotators were 0.41 for radiographs and 0.42 for NILT, indicating moderate agreement.

Across all imaging modalities and lesion depths, variability in diagnostic performance was observed. For lesions detected on radiographs, median sensitivity was highest for sound surfaces (0.89, $p < 0.05$) followed by dentin lesions (0.74) and enamel lesions (0.46). Median specificity for radiographs reached 0.96 for dentin lesions, 0.91 for enamel lesions and was lowest (0.6, $p < 0.05$) for sound surfaces (Table 3). In NILT images, median sensitivity was highest for sound surfaces (0.91, $p < 0.05$), 0.6 for dentin lesions and 0.49 for enamel lesions. Median specificity was highest for dentin lesions (0.98, $p < 0.05$), 0.91 for enamel lesions and 0.58 for sound surfaces (Table 3). There were no significant differences across dataset modalities ($p > 0.05$).

Analysis of annotation aggregation strategies was stratified by dataset modality (radiographs, NILT and multimodal) and lesion depth (sound, enamel lesion, and dentin lesion; Fig. 3). Kruskal-Wallis test revealed statistically significant differences between aggregation strategies across modalities and surface classes (all $p < 0.001$). Dunn's post-hoc tests were conducted to determine pairwise differences (Supplementary Table 1). For radiographs, MACE outperformed MV, DS and WMV significantly across all surface classes (all $p \leq 0.002$). MACE yielded median AUROC values of 0.77 for sound surfaces, 0.73 for enamel lesions, and 0.88 for dentin lesions (Table 4). In the NILT dataset, MACE demonstrated superior performance compared to MV across all surface classes ($p \leq 0.001$) and higher performance than DS in enamel and dentin lesions ($p < 0.001$). MACE achieved median AUROC values of 0.76 for sound surfaces, 0.72 for enamel lesions, and 0.87 for dentin lesions (Table 4). In the multimodal dataset, DS outperformed the other label aggregation strategies across all surface classes (all $p < 0.05$). Median AUROC was 0.79 for sound surfaces, 0.74 for enamel lesions, and 0.90 for dentin lesions. Further details can be found in the

Table 2
Descriptive statistics of annotation performance in percent (%) of five different annotators across datasets and lesion depth. Statistics are provided as median (minimum, maximum). TN = true negative, FP = false positive, FN = false negative, TP = true positive.

Dataset	Lesion depth	TN	FP	FN	TP
Radiography	Sound	60.0 (14.1, 75.4)	40.0 (24.6, 85.9)	10.7 (2.0, 24.2)	89.3 (75.8, 98.0)
	Enamel	91.1 (79.1, 97.4)	8.9 (2.6, 20.9)	54.3 (37.1, 96.1)	45.7 (3.9, 62.9)
	Dentin	96.4 (95.9, 99.1)	3.6 (0.9, 4.1)	25.6 (21.1, 70.0)	74.4 (30.0, 78.9)
NILT	Sound	58.1 (10.8, 77.0)	41.9 (23.0, 89.2)	8.8 (3.8, 23.5)	91.2 (76.5, 96.2)
	Enamel	91.3 (78.3, 97.1)	8.7 (2.9, 21.7)	51.1 (30.0, 93.9)	48.9 (6.1, 70.0)
	Dentin	97.7 (96.1, 99.0)	2.3 (1.0, 3.9)	40.0 (31.1, 81.1)	60.0 (18.9, 68.9)

Table 3
Stratified annotation metrics (AUROC) for different imaging modalities (radiographs, NILT), and lesion depths (sound, enamel lesion, dentin lesion). Metrics of the five annotators are provided as median (minimum, maximum).

Dataset	Lesion depth	Sensitivity	Specificity	F1-score	AUROC
Radiographs	Sound	0.89 (0.76, 0.98)	0.60 (0.14, 0.75)	0.89 (0.76, 0.85)	0.75 (0.56, 0.77)
	Enamel	0.46 (0.04, 0.63)	0.91 (0.79, 0.97)	0.46 (0.04, 0.59)	0.69 (0.51, 0.71)
	Dentin	0.74 (0.30, 0.79)	0.96 (0.96, 0.99)	0.74 (0.30, 0.75)	0.85 (0.65, 0.88)
NILT	Sound	0.91 (0.76, 0.96)	0.58 (0.11, 0.77)	0.82 (0.78, 0.85)	0.75 (0.54, 0.77)
	Enamel	0.49 (0.06, 0.70)	0.91 (0.78, 0.97)	0.57 (0.11, 0.62)	0.70 (0.52, 0.74)
	Dentin	0.60 (0.19, 0.69)	0.98 (0.96, 0.99)	0.64 (0.29, 0.72)	0.79 (0.59, 0.83)

Supplementary Table 1. Label aggregation in multimodal datasets resulted in higher AUROC-scores compared to unimodal datasets. DS, the best-performing method in multimodal datasets, outperformed MACE, the best-performing method in unimodal datasets, in all lesion depths except enamel lesions (Supplementary Table 2). Additional label aggregation metrics such as sensitivity and specificity are presented in Supplementary Table 3 and Supplementary Table 4.

4. Discussion

High quality labels are important for both training and testing AI [20,21]. However, generating gold standard annotations, such as those derived from histology, is often not feasible given ethical and practical limitations. Consequently, multiple annotators are employed to annotate each image, assuming this "crowd intelligence" to overcome the limitations of individual examiners, yielding robust annotations. Nevertheless, this raises another challenge, because these annotations need to be aggregated into one final annotation for the training and testing of AI models. We assessed which of four different annotation aggregation strategies yielded the best performance when measured against the gold standard (histology), in two dental image modalities for one specific task, detecting carious lesions. Our work is of relevance to dental AI researchers and developers, as presently, the choice of strategy is seldom justified or reported to be made on an informed basis. Our findings indicate that DS performed best in multimodal datasets, while MACE excelled in unimodal ones. Both strategies performed better than the currently common MV or (the less common) WMV. Our results also confirm a high variability in annotations when detecting carious lesions on dental images. Notably, while the annotation variability was high in our study sample, annotation quality especially of our radiographs dataset was comparably high, with median sensitivity values of 0.46 for enamel lesions and 0.74 for dentin lesions [4,22].

Tackling high inter-annotator variability to generate high quality data is not only a task limited to dentistry. Moderate inter-annotator agreement was reported for specific tasks in the field of histopathology as well [23–25]. Here, annotations are often refined using a single (or multiple) gold standard annotator(s), and it was shown that annotation accuracy increases with experience [26,27]. The same approach is not always feasible in dentistry, as it was shown that annotation accuracy does not necessarily correlate with seniority [5]. Here, imaging-modality-specific performance seems to play a more important role [4], especially in detecting enamel lesions.

All aggregation strategies applied in this study have their own strengths and limitations. MV excels in its simplicity but can be significantly impacted by individual annotators with low sensitivity and specificity, particularly when the total number of annotators is small. This is especially true in scenarios where sensitivity is low, such as with enamel lesions [22]. WMV addresses this limitation by reducing the influence of low-performing individual raters through weighting their contributions based on confidence scores. However, this approach requires a ground truth test set to determine the confidence scores for each annotator. Since these scores can vary depending on cavity depth, multiple annotations of images with known ground truth values are necessary. In our study, we used 20 images per cavity depth (sound, enamel lesion, dentin lesion) for this purpose. The key advantage of WMV is not only its potential for improved annotation accuracy but also its ability to facilitate cross-study comparisons of annotator performance, enhancing transparency in reporting.

Probabilistic aggregation methods (DS, MACE) demonstrated superior performance in our study and do not require a ground truth dataset, as they estimate the underlying ground truth by modeling annotator behavior and the observed annotation patterns. Here, uncertainties stemming from underlying patterns such as diagnostic modality used, individual annotator performance, annotation complexity and annotator biases can be accounted for. However, while the lack of a need for ground truth allows greater flexibility, it also poses a major limitation: it may hinder the ability to transparently report annotator performance, as no explicit benchmarking against a gold standard is performed. Thus, we recommend reporting annotator accuracy/aggregated annotation accuracy on a separate *in vitro* test set where ground truth values are available, to ensure transparent reporting of annotation quality.

Notably, the current study focused on detecting carious lesions, leading to surface-based classification. A growing trend in dental image analysis using AI, however, is to employ object detection or segmentation models. The latter yield pixel-wise classifications and, consequently, pixel blobs indicating the presence of a condition (*e.g.* a carious lesion) in a certain image area. Segmentation introduces additional complexity, requiring both lesion presence and depth to be identified alongside precise pixel-level location. Similarly to classification tasks, label aggregation such as majority voting can be employed, treating each pixel as an individual classification task. However, due to the high variability in the annotations, these methods might not prove as sufficient. The segmentation expectation-maximization algorithm is one method used to aggregate crowd-sourced segmentations. Here the annotator's skill is modeled as a latent parameter representing the probability of providing a correct annotation [17]. Other methods employ neural networks to learn from diverse annotations by modeling annotator-specific biases and spatial error patterns [28]. These networks can aggregate multiple, variable segmentation labels into a consensus label map, leveraging their ability to capture complex spatial characteristics of annotator mistakes. Another promising approach involves semi-supervised learning, which reduces the reliance on labeled data. By leveraging large volumes of unlabeled data, semi-supervised methods can learn to extract clinically relevant features from dental imagery without requiring explicit labels [29]. These representations can then be fine-tuned on smaller, labeled datasets or even in-vitro datasets to achieve high performance with fewer annotations.

In addition to these methods that could be applied to retrospective, clinically yielded datasets, *in vitro* data may offer opportunities to improve supervised model training and validation. Micro-CT provides high-resolution imaging that is particularly valuable for generating ground truth labels [30]. A first open *in vitro* dataset was released in 2024 which contains both micro CT data and corresponding conventional dental radiographs [31]. Annotating accuracy significantly increased if annotators had access to the corresponding micro CTs instead of merely the radiographs [31]. Notably, training and testing AI models solely on *in vitro* datasets poses significant limitations. *In vitro* data lacks the variability and complexity of real-world clinical imagery, such as patient motion, anatomical variability, and imaging artifacts.

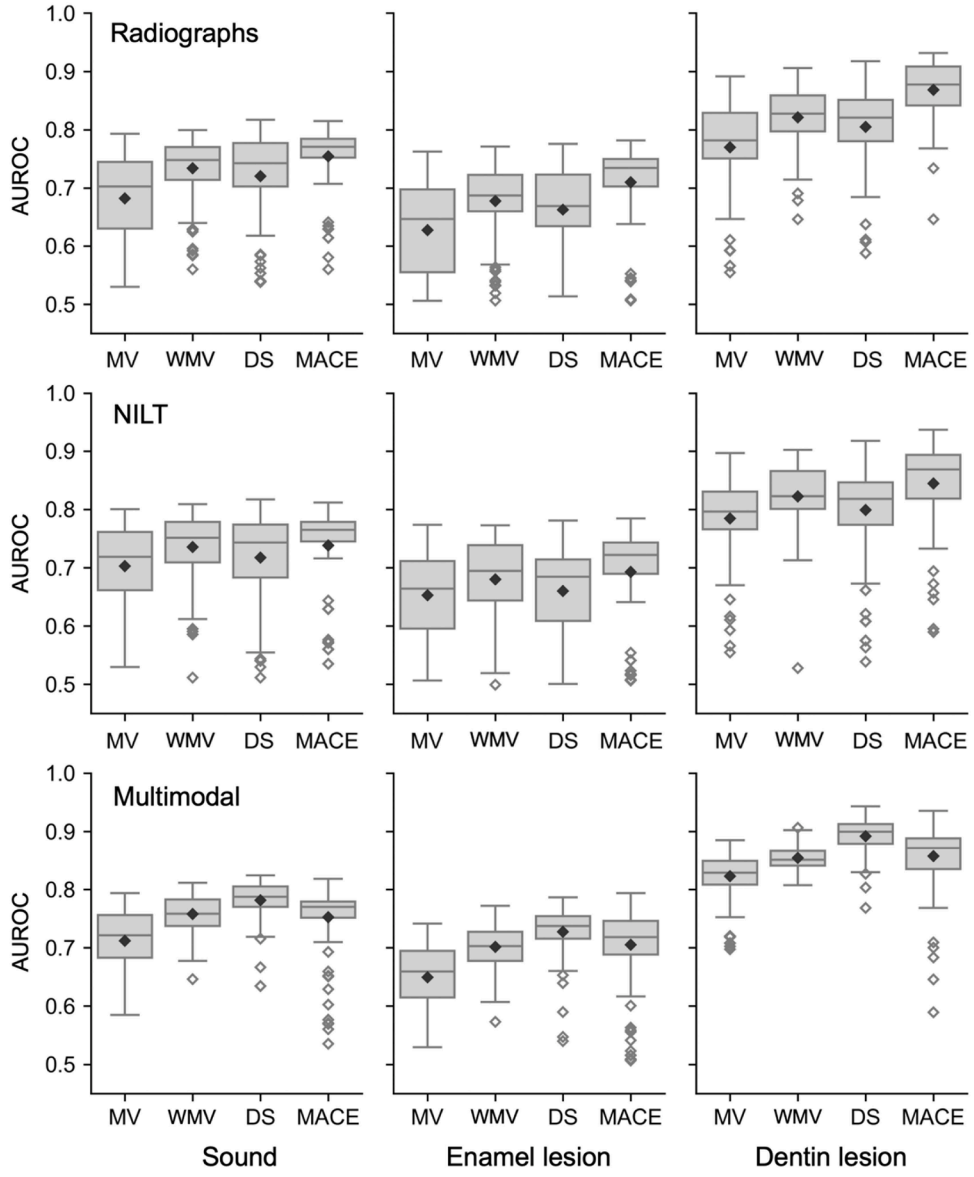

Fig. 3. Mean area under the receiver operating characteristic curve (AUROC) scores across different aggregation strategies, dataset modalities and lesion depths. Boxplots represent the distribution of AUROC values for each aggregation strategy (MV = Majority voting, WMV = Weighted majority voting, DS = Dawid-Skene Algorithm, MACE = multi-annotator competence estimation) applied to radiographs, NILT, and the multimodal dataset. Each subplot corresponds to a specific lesion depth (sound, enamel lesion, dentin lesion). The black diamonds indicate the mean AUROC for each aggregation strategy.

These datasets are also often limited in size and diversity, potentially introducing biases that undermine clinical utility. To address these limitations while benefiting from the increased sensitivity, datasets for model training could include both *in vitro* as well as *in vivo* data.

While our study provides valuable insights into the performance of annotation aggregation strategies in detecting carious lesions, several limitations must be acknowledged. These include the mentioned focus on surface-wise classification tasks and the reliance on *in vitro* data, which may limit the generalizability of our findings. Additionally, annotator variability was limited to five annotators per modality, which may not fully capture the diversity of real-world annotation scenarios. Further, bootstrapping from the original dataset allowed us to model variability across datasets; however, it may have introduced a certain bias by over-representing individual annotators or patterns present in the original data. Lastly, a class imbalance was present in our dataset, as most surfaces in our study were sound, which, however, is common in clinical imagery. This imbalance may have influenced annotator decisions and consequentially the performance of aggregation methods, particularly MV due to its reliance on majority consensus, which can lead to biases toward the dominant class and reduced sensitivity for underrepresented lesion categories. Future research should explore the applicability of our findings to segmentation tasks, evaluate the aggregation strategies on *in vivo* datasets, and incorporate additional imaging modalities and longitudinal datasets to enhance label accuracy and, consequently, the performance of AI models.

5. Conclusion

The high variability of caries annotations underscores the challenges posed by inter-annotator differences and the inherent complexity of diagnosing caries in radiographs and NILT images. Our findings demonstrate that annotation aggregation strategies such as DS and MACE outperformed MV and WMV, with DS excelling in multimodal datasets and MACE performing best in unimodal datasets. This indicated that the commonly applied simple MV may not be ideal and that the optimal aggregation strategy depends on the dataset characteristics.

Table 4
Descriptive statistics for label aggregation strategies. Median area under the receiver operating characteristic curve (AUROC) and the 95 % confidence intervals (2.5th percentile, 97.5th percentile) are reported for each aggregation strategy, stratified by dataset type and lesion depth. Bold formatting indicates the best performing strategy in each dataset (median). MV = Majority voting, WMV = Weighted majority voting, DS = Dawid-Skene Algorithm, MACE = multi-annotator competence estimation. Pairwise Dunn's post-hoc test results are presented in Supplementary Table 1.

Dataset	Lesion depth	MV	WMV	DS	MACE
Radiographs	Sound	0.70 (0.54, 0.79)	0.75 (0.59, 0.80)	0.74 (0.54, 0.80)	**0.77 (0.61, 0.80)**
	Enamel lesion	0.65 (0.51, 0.75)	0.69 (0.53, 0.76)	0.67 (0.52, 0.76)	**0.73 (0.54, 0.77)**
	Dentin lesion	0.78 (0.58, 0.87)	0.83 (0.70, 0.89)	0.82 (0.61, 0.90)	**0.88 (0.77, 0.93)**
NILT	Sound	0.72 (0.55, 0.79)	0.75 (0.59, 0.80)	0.74 (0.54, 0.81)	**0.76 (0.56, 0.80)**
	Enamel lesion	0.66 (0.51, 0.76)	0.69 (0.53, 0.77)	0.68 (0.50, 0.77)	**0.72 (0.51, 0.77)**
	Dentin lesion	0.80 (0.60, 0.87)	0.82 (0.72, 0.89)	0.82 (0.58, 0.91)	**0.87 (0.65, 0.93)**
Multimodal	Sound	0.72 (0.59, 0.78)	0.76 (0.70, 0.80)	**0.79 (0.69, 0.82)**	0.77 (0.57, 0.82)
	Enamel lesion	0.66 (0.53, 0.73)	0.70 (0.63, 0.77)	**0.74 (0.61, 0.78)**	0.72 (0.51, 0.78)
	Dentin lesion	0.83 (0.71, 0.88)	0.85 (0.81, 0.90)	**0.90 (0.83, 0.93)**	0.87 (0.69, 0.93)

Therefore, there is a need for informed application of specific aggregation strategies.

Ethics

The protocol of this study was approved by the Ethics Committee of the Charité - Universitätsmedizin Berlin (EA4/102/14).

CRediT authorship contribution statement

Vanessa Klein: Writing – review & editing, Visualization, Investigation, Data curation. **Martha Büttner:** Writing – review & editing, Supervision, Project administration, Investigation. **Gerd Göstemeyer:** Writing – review & editing, Supervision, Resources, Project administration. **Sarina Rolle:** Writing – review & editing, Resources, Data curation. **Antonin Tichy:** Writing – review & editing. **Falk Schwendicke:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Noah F. Nordblom:** Writing – original draft, Visualization, Software, Investigation, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Falk Schwendicke reports financial support was provided by German Research Foundation. Falk Schwendicke is a co-founder of dentalXrai, a startup focusing on dental radiograph analytics using artificial intelligence, but the present work was fully independent from this status. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements and funding

This work was supported by the German Research Foundation DFG KR 5457/1–1.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jdent.2025.105728.

References

[1] J. Vaarkamp, J.J. Ten Bosch, E.H. Verdonschot, E.M. Bronkhorst, The real performance of bitewing radiography and fiber-optic transillumination in approximal caries diagnosis, J. Dent. Res. 79 (2000) 1747–1751, https://doi.org/10.1177/00220345000790100301.
[2] A.G. Cantu, S. Gehrung, J. Krois, A. Chaurasia, J.G. Rossi, R. Gaudin, K. Elhennawy, F. Schwendicke, Detecting caries lesions of different radiographic extension on bitewings using deep learning, J. Dent. 100 (2020) 103425.
[3] H. Devlin, T. Williams, J. Graham, M. Ashley, The ADEPT study: a comparative study of dentists' ability to detect enamel-only proximal caries in bitewing radiographs with and without the use of AssistDent artificial intelligence software, Br. Dent. J. 231 (2021) 481–485.
[4] M. Janjic Rankovic, S. Kapor, Y. Khazaei, A. Crispin, I. Schüler, F. Krause, K. Ekstrand, S. Michou, F. Eggmann, A. Lussi, M.-C. Huysmans, K. Neuhaus, J. Kühnisch, Systematic review and meta-analysis of diagnostic studies of proximal surface caries, Clin. Oral Investig. 25 (2021) 6069–6079, https://doi.org/10.1007/s00784-021-04113-1.
[5] J. Boldt, M. Schuster, G. Krastl, M. Schmitter, J. Pfundt, A. Stellzig-Eisenhauer, F. Kunz, Developing the benchmark: establishing a gold standard for the evaluation of AI caries diagnostics, J. Clin. Med. 13 (2024) 3846, https://doi.org/10.3390/jcm13133846.
[6] L.T. Arsiwala-Scheppach, A. Chaurasia, A. Müller, J. Krois, F. Schwendicke, Machine learning in dentistry: a scoping review, J. Clin. Med. 12 (2023) 937, https://doi.org/10.3390/jcm12030937.
[7] S.E. Uribe, J. Issa, F. Sohrabniya, A. Denny, N.N. Kim, A.F. Dayo, A. Chaurasia, A. Sofi-Mahmudi, M. Büttner, F. Schwendicke, Publicly available dental image datasets for artificial intelligence, J. Dent. Res. 103 (2024) 1365–1374, https://doi.org/10.1177/00220345241272052.
[8] G. Li, J. Wang, Y. Zheng, M.J. Franklin, Crowdsourced data management: a survey, IEEE Trans. Knowl. Data Eng. 28 (2016) 2296–2319, https://doi.org/10.1109/TKDE.2016.2535242.
[9] Y. Zheng, Z. Xu, X. Wang, The fusion of deep learning and fuzzy systems: a state-of-the-art survey, IEEE Trans. Fuzzy Syst. 30 (2022) 2783–2799, https://doi.org/10.1109/TFUZZ.2021.3062899.
[10] M. Herde, D. Huseljic, B. Sick, Multi-annotator Deep Learning: A Probabilistic Framework for Classification, 2023, https://doi.org/10.48550/arXiv.2304.02539.
[11] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, M. Shokouhi, Community-based bayesian aggregation models for crowdsourcing, in: Proceedings of the 23rd International Conference on World Wide Web, ACM, Seoul Korea, 2014, pp. 155–164, https://doi.org/10.1145/2566486.2567989.
[12] K. Elhennawy, H. Askar, P.-G. Jost-Brinkmann, S. Reda, A. Al-Abdi, S. Paris, F. Schwendicke, *In vitro* performance of the DIAGNOcam for detecting proximal carious lesions adjacent to composite restorations, J. Dent. 72 (2018) 39–43, https://doi.org/10.1016/j.jdent.2018.03.002.
[13] F. Schwendicke, H. Meyer-Lueckel, M. Schulz, C.E. Dörfer, S. Paris, Radiopaque tagging masks caries lesions following incomplete excavation *in vitro*, J. Dent. Res. 93 (2014) 565–570, https://doi.org/10.1177/0022034514531291.
[14] F. Schwendicke, K. Elhennawy, S. Paris, P. Friebertshäuser, J. Krois, Deep learning for caries lesion detection in near-infrared light transillumination images: a pilot study, J. Dent. 92 (2020) 103260, https://doi.org/10.1016/j.jdent.2019.103260.
[15] T. Ekert, J. Krois, F. Schwendicke, Building a mass online annotation tool for dental radiographic imagery, (2018). https://openreview.net/forum?id=SkcgYEoiG (accessed February 7, 2025).
[16] A.P. Dawid, A.M. Skene, Maximum likelihood estimation of observer error-rates using the EM algorithm, Appl. Stat. 28 (1979) 20, https://doi.org/10.2307/2346806.
[17] D. Ustalov, N. Pavlichenko, B. Tseitlin, Learning from crowds with crowd-Kit, J. Open Source Software 9 (2024) 6227, https://doi.org/10.21105/joss.06227.
[18] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, E. Hovy, Learning Whom to Trust with MACE, in: L. Vanderwende, H. Daumé III, K. Kirchhoff (Eds.), Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 1120–1130. https://aclanthology.org/N13-1132/(accessed. January 6, 2025).
[19] J.L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educ. Psychol. Meas. 33 (1973) 613–619, https://doi.org/10.1177/001316447303300309.

[20] L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, H. Harmouch, The effects of data quality on machine learning performance, (2022). https://doi.org/10.48550/arXiv.2207.14529.
[21] M. Büttner, L. Schneider, A. Krasowski, J. Krois, B. Feldberg, F. Schwendicke, Impact of noisy labels on dental deep learning-calculus detection on bitewing radiographs, J. Clin. Med. 12 (2023) 3058, https://doi.org/10.3390/jcm12093058.
[22] F. Schwendicke, M. Tzschoppe, S. Paris, Radiographic caries detection: a systematic review and meta-analysis, J. Dent. 43 (2015) 924–933, https://doi.org/10.1016/j.jdent.2015.02.009.
[23] D.C. Paech, A.R. Weston, N. Pavlakis, A. Gill, N. Rajan, H. Barraclough, B. Fitzgerald, M. Van Kooten, A systematic review of the interobserver variability for histology in the differentiation between squamous and nonsquamous non-small cell lung cancer, J. Thorac. Oncol. 6 (2011) 55–63, https://doi.org/10.1097/JTO.0b013e3181fc0878.
[24] G. Nir, S. Hor, D. Karimi, L. Fazli, B.F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R.S. Wilson, K.A. Iczkowski, M.S. Lucia, P.C. Black, P. Abolmaesumi, S.L. Goldenberg, S.E. Salcudean, Automatic grading of prostate cancer in digitized histopathology images: learning from multiple experts, Med. Image Anal. 50 (2018) 167–180, https://doi.org/10.1016/j.media.2018.09.005.
[25] L.J.H. Smits, E. Vink-Börger, G. van Lijnschoten, I. Focke-Snieders, R.S. van der Post, J.B. Tuynman, N.C.T. van Grieken, I.D. Nagtegaal, Diagnostic variability in the histopathological assessment of advanced colorectal adenomas and early colorectal cancer in a screening population, Histopathology 80 (2022) 790–798, https://doi.org/10.1111/his.14601.
[26] O. Iizuka, F. Kanavati, K. Kato, M. Rambeau, K. Arihiro, M. Tsuneki, Deep learning models for histopathological classification of gastric and colonic epithelial tumours, Sci. Rep. 10 (2020) 1504, https://doi.org/10.1038/s41598-020-58467-9.
[27] S. Shinohara, A. Bychkov, J. Munkhdelger, K. Kuroda, H.-S. Yoon, S. Fujimura, K. Tabata, B. Furusato, D. Niino, S. Morimoto, T. Yao, T. Itoh, H. Aoyama, N. Tsuyama, Y. Mikami, T. Nagao, T. Ikeda, N. Fukushima, O. Harada, T. Kiyokawa, N. Yoshimi, S. Aishima, I. Maeda, I. Mori, K. Yamanegi, K. Tsuneyama, R. Katoh, M. Izumi, Y. Oda, J. Fukuoka, Substantial improvement of histopathological diagnosis by whole-slide image-based remote consultation, Virchows. Arch. 481 (2022) 295–305, https://doi.org/10.1007/s00428-022-03327-2.
[28] L. Zhang, R. Tanno, M.-C. Xu, C. Jin, J. Jacob, O. Cicarrelli, F. Barkhof, D. Alexander, Disentangling human error from ground truth in segmentation of medical images, Adv. Neural. Inf. Process. Syst. 33 (2020) 15750–15762.
[29] A. Taleb, C. Rohrer, B. Bergner, G. De Leon, J.A. Rodrigues, F. Schwendicke, C. Lippert, J. Krois, Self-supervised learning methods for label-efficient dental caries classification, Diagnostics 12 (2022) 1237.
[30] C. Boca, B. Truyen, L. Henin, A.G. Schulte, V. Stachniss, N. De Clerck, J. Cornelis, P. Bottenberg, Comparison of micro-CT imaging and histology for approximal caries detection, Sci. Rep. 7 (2017) 6680, https://doi.org/10.1038/s41598-017-06735-6.
[31] R.E.G. Valenzuela, P. Mettes, B.G. Loos, H. Marquering, E. Berkhout, Enhancement of early proximal caries annotations in radiographs: introducing the Diagnostic Insights for Radiographic Early-caries with micro-CT (ACTA-DIRECT) dataset, BMC Oral. Health 24 (2024) 1325, https://doi.org/10.1186/s12903-024-05076-x.