Journal of
Clinical Medicine

MDPI

Article

Developing the Benchmark: Establishing a Gold Standard for the Evaluation of AI Caries Diagnostics

Julian Boldt [1,†], Matthias Schuster [1,*,†], Gabriel Krastl [2], Marc Schmitter [1], Jonas Pfundt [1], Angelika Stellzig-Eisenhauer [3] and Felix Kunz [3]

[1] Department of Prosthetic Dentistry, University Hospital Würzburg, 97070 Würzburg, Germany; schmitter_m@ukw.de (M.S.); pfundt.jonas@stud-mail.uni-wuerzburg.de (J.P.)
[2] Center of Dental Traumatology, Department of Conservative Dentistry and Periodontology, University Hospital Würzburg, 97070 Würzburg, Germany; krastl_g@ukw.de
[3] Department of Orthodontics, University Hospital Würzburg, 97070 Würzburg, Germany; stellzig_a@ukw.de (A.S.-E.); kunz_f@ukw.de (F.K.)
* Correspondence: schuster_m3@ukw.de
† These authors contributed equally to this work.

Abstract: Background/Objectives: The aim of this study was to establish a histology-based gold standard for the evaluation of artificial intelligence (AI)-based caries detection systems on proximal surfaces in bitewing images. **Methods**: Extracted human teeth were used to simulate intraoral situations, including caries-free teeth, teeth with artificially created defects and teeth with natural proximal caries. All 153 simulations were radiographed from seven angles, resulting in 1071 in vitro bitewing images. Histological examination of the carious lesion depth was performed twice by an expert. A total of thirty examiners analyzed all the radiographs for caries. **Results**: We generated in vitro bitewing images to evaluate the performance of AI-based carious lesion detection against a histological gold standard. All examiners achieved a sensitivity of 0.565, a Matthews correlation coefficient (MCC) of 0.578 and an area under the curve (AUC) of 76.1. The histology receiver operating characteristic (ROC) curve significantly outperformed the examiners' ROC curve ($p < 0.001$). All examiners distinguished induced defects from true caries in 54.6% of cases and correctly classified 99.8% of all teeth. Expert caries classification of the histological images showed a high level of agreement (intraclass correlation coefficient (ICC) = 0.993). Examiner performance varied with caries depth ($p \leq 0.008$), except between E2 and E1 lesions ($p = 1$), while central beam eccentricity, gender, occupation and experience had no significant influence (all $p \geq 0.411$). **Conclusions**: This study successfully established an unbiased dataset to evaluate AI-based caries detection on bitewing surfaces and compare it to human judgement, providing a standardized assessment for fair comparison between AI technologies and helping dental professionals to select reliable diagnostic tools.

Keywords: dental caries; artificial intelligence; diagnostics; bitewing; radiography; benchmarking

check for updates

Citation: Boldt, J.; Schuster, M.; Krastl, G.; Schmitter, M.; Pfundt, J.; Stellzig-Eisenhauer, A.; Kunz, F. Developing the Benchmark: Establishing a Gold Standard for the Evaluation of AI Caries Diagnostics. *J. Clin. Med.* **2024**, *13*, 3846. https://doi.org/10.3390/jcm13133846

Academic Editors: Michał Strzelecki, Adam Piórkowski, Rafał Obuchowicz and Karolina Nurzynska

Received: 6 June 2024
Revised: 26 June 2024
Accepted: 28 June 2024
Published: 29 June 2024

1. Introduction

With the exponential growth in computational power across virtually all semiconductor-based devices, artificial intelligence (AI) is finding its way into medical sciences, driven by the desire to increase diagnostic accuracy, improve treatment outcomes and optimize workflow efficiency [1–3]. The increasing prevalence of articles on this subject in literature is evidence of this [4]. From identifying anatomical or pathological structures to assisting with logistical challenges, AI promises to save time and reduce costs [5–7].

In human medical imaging, AI applications show promising potential in several areas, particularly in oncology [8]. A major advantage of these AI applications is that their training is based on verified histopathological findings, thus relying on a reliable reference.

In dentistry AI, algorithms have already been developed for automated analysis of radiographs for caries diagnosis [9–12]. Image recognition in regard to caries detection has been approached using a variety of techniques [13]. However, the traditional comprehensive analysis of X-rays by the dentist is time-consuming and limited by the possibility of human error, which AI promises to largely eliminate [14–17].

In 2022, Mohammad-Rahimi et al. conducted a systematic review to evaluate the accuracy of automated caries detection systems and showed that the majority of the models included were able to deliver results with clinically acceptable performance parameters, although the quality of studies is often currently low [18]. In particular, in a systematic review and meta-analysis, Ammar and Kühnisch reported acceptable diagnostic accuracy of AI models for caries detection and classification on bitewing radiographs [19]. These radiographs are the most reliable and widely used clinical imaging method for caries diagnosis [20,21]. Despite some promising results, it has also been criticized that AI-based caries diagnostic studies often neither include an appropriate definition of caries nor provide information on the type of carious lesion detected and have limitations in regard to size and heterogeneity of the reported datasets [22–24].

The advancement of AI applications for caries detection in bitewing images relies primarily on the use of deep learning networks, primarily convolutional neural networks [25]. This iterative process begins with the compilation of large datasets of annotated bitewing radiographs, in which dental professionals delineate regions of interest corresponding to caries, healthy tooth structure and other anatomical structures [4]. These annotated images are then divided into distinct training and test sets. Using machine learning algorithms, AI-driven methods analyze the training dataset, identifying intricate patterns and extrapolating the desired results [4]. The integrity of the trained model is then evaluated against the separate test dataset, assessing its ability to analyze novel, unseen data. The accuracy of the model is quantified by comparing the predictions derived from the test dataset with the actual annotations. This dichotomy between training and test datasets is crucial to ensure that the AI model goes beyond simply memorizing specific instances from the training dataset, and instead acquires a robust understanding of the general patterns and features that are essential for accurate caries detection.

However, a fundamental limitation arises in the whole training process, which lies in the annotation of radiographs by dentists, representing the AI training gold standard. According to the Standards for Reporting Diagnostic Accuracy Studies (STARD), a gold standard is defined as an error-free reference standard that represents the best available method for determining the presence or absence of the target condition [26]. Although dentists are trained in clinical diagnosis, their sensitivity and specificity for detecting carious lesions on radiographs is somewhat limited [27–30], in particular for subtle or early stages of lesions. In addition, various factors, such as experience, knowledge, technical skills and time pressure, may influence diagnostic accuracy [31]. While it is undeniable that deep learning can identify features indicative of caries, the underlying methodology has potentially serious practical implications as the predictions only reflect sensitivity and specificity within the training and test data. This concern is exacerbated by the existence of commercial automated dental radiograph analysis software solutions, most of which lack transparency regarding the scientific basis of their AI models.

The aim of this study was, therefore, to develop reliable in vitro simulations of bitewing radiographs based on the histological gold standard to provide a basis for evaluating the performance of AI-based software currently offered by commercial vendors for the automated analysis of caries in bitewing radiographs. In addition, a reference dataset of caries diagnoses from in vitro bitewing radiographs by different examiners was created to serve as a benchmark for predicting whether AI applications can provide a diagnostic advantage to dental examiners.

2. Materials and Methods

2.1. Ethical Aspects

This study was approved by the Ethics Committee of the Medical Faculty of the University of Würzburg (15/15, 9 February 2015) and was carried out in compliance with the

Declaration of Helsinki. All teeth used were extracted for existing clinical indications, with ethical approval, voluntarily and without coercion, and were anonymized. Information provided to patients still allowed for patient withdrawal but excluded the possibility of targeted destruction of donated teeth.

2.2. *Trial Profile*

The trial profile is depicted in Figure 1.

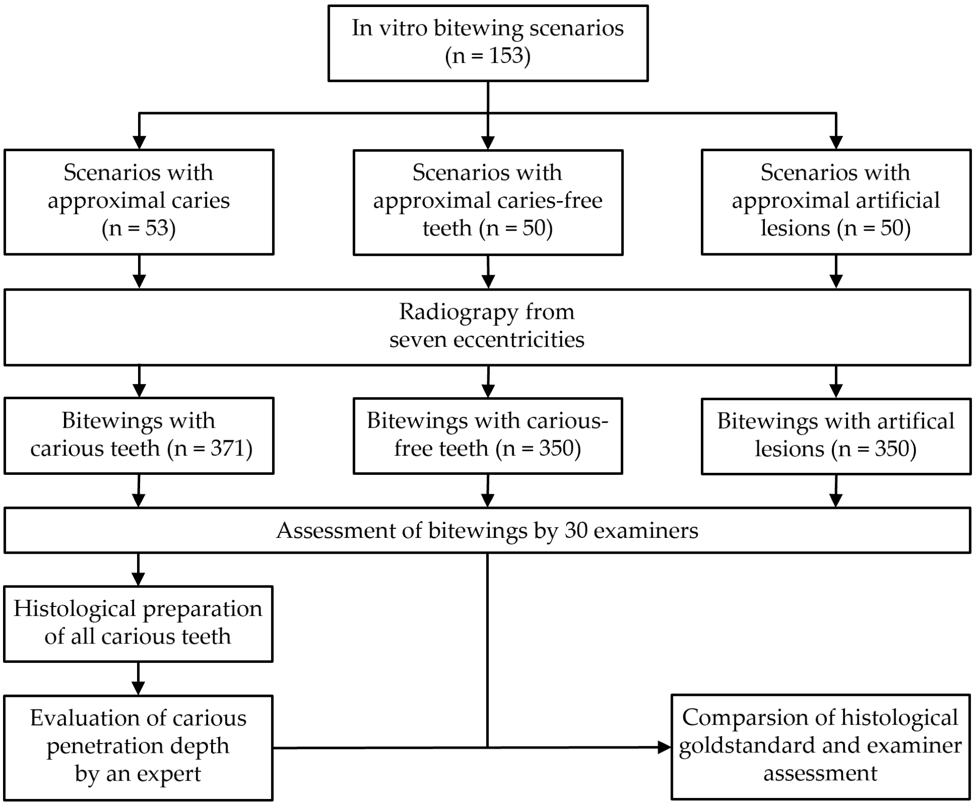
In vitro bitewing scenarios (n = 153)
Scenarios with approximal caries (n = 53)
Scenarios with approximal caries-free teeth (n = 50)
Scenarios with approximal artificial lesions (n = 50)
Radiograpy from seven eccentricities
Bitewings with carious teeth (n = 371)
Bitewings with carious-free teeth (n = 350)
Bitewings with artifical lesions (n = 350)
Assessment of bitewings by 30 examiners
Histological preparation of all carious teeth
Evaluation of carious penetration depth by an expert
Comparsion of histological goldstandard and examiner assessment

Figure 1. Trial profile.

2.3. *Tooth Selection*

This study used 179 extracted permanent human teeth that were preserved in a 1% tosylchloramide–sodium solution immediately after extraction. All teeth were obtained from various dental clinics and hospitals, ensuring a diverse representation of carious and caries-free conditions. Inclusion criteria were visually and radiographically normal and properly formed permanent teeth with restorative measures that did not significantly interfere with or prevent radiographic caries diagnosis of proximal surfaces. Exclusion criteria comprised completely decayed teeth or root remains, and teeth whose clinical appearance matched hereditary anomalies. All teeth were examined for possible carious lesions by visual inspection using a 2.5× close-up magnification loupe (GTX 2 telescope loupe system; Carl Zeiss Vision GmbH, Aalen, Germany) and tactile examination using a dental probe (EXS3A; Henry Schein Dental Deutschland GmbH, Langen, Germany). A digital single-lens reflex camera (Olympus E-400; Olympus Europa SE & Co. KG, Hamburg, Germany) with 50 mm macro lens (four thirds standard) was used to photograph each tooth from five directions (occlusal, vestibular, oral, mesial, distal). In addition, each tooth was radiographed in the vestibulo-oral and mesiodistal directions (Sirona Heliodent DS; Dentsply Sirona Deutschland GmbH, Bensheim, Germany) (Figure 2). Based on the visual, tactile and radiographic findings, two dentists classified all teeth as carious or caries-free.

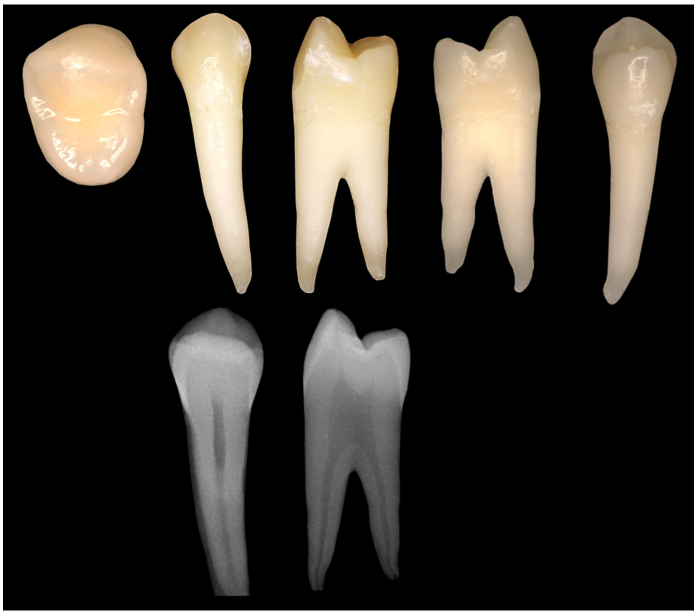

Figure 2. Photographic and radiological documentation of all teeth.

2.4. Preparation of Artificial Defects

A total of 50 caries-free teeth were used to test the ability to discriminate between carious lesions and artificial defects. The artificial defects were created on the proximal surfaces using 1 mm, 2 mm, 3 mm and 4 mm spherical diamond burs (Gebr. Brasseler GmbH & Co. KG, Lemgo, Germany). During the preparation process, the burs were inserted into the teeth, creating artificial defects half the size of the drill's diameter.

The selection of diameters ranging from 1 mm to 4 mm was based on findings of Stroud et al. on the mean enamel thickness of permanent posterior teeth [32]. This allowed for clinically accurate lesion simulations.

2.5. Bitewing Design

An occlusal holder (Split-Fixator; Scheu-Dental GmbH, Iserlohn, Germany) was fitted with Plexiglas blocks attached at the top and bottom by means of a milled groove. The teeth were embedded in Periphery Wax (Sigma Dental, Handewitt, Germany) and mounted in an anatomically and physiologically accurate configuration to standardize their position for radiographic imaging of the bitewings (Figure 3).

Despite clinical best efforts to use the parallel technique, obtaining superposition-free images of the region of interest in bitewing radiographs remains challenging. Factors such as relative positioning of the teeth, superimpositions, the curvature of the dental arch, the orientation and spatial distortion of the film during exposure and the alignment of the X-ray tube all contribute to the superposition of dental tissue in the interproximal region [33,34]. To mimic clinically relevant situations and improve data quality, the study included not only orthoradial images, but also mesial and distal eccentric images at varying angles. For this purpose, the model was fixed in a rotating vice with a graduated scale that allowed precise angular adjustments in 2-degree increments.

Each examination series yielded a total of seven radiographs, all taken with the same X-ray unit (Sirona Heliodent DS; Dentsply Sirona Deutschland GmbH, Bensheim, Germany; 60 kV, 7 mA, 0,06 ms). These included a 0° orthogonal image and 4°, 6° and 8° mesial and distal eccentric images (totaling 7 images per series) (Figure 4).

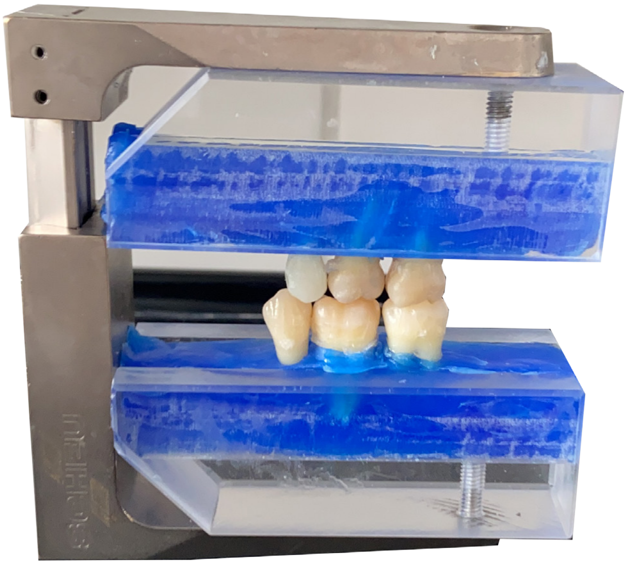

Figure 3. Occlusion holder with fixed teeth simulating clinical bitewing scenarios.

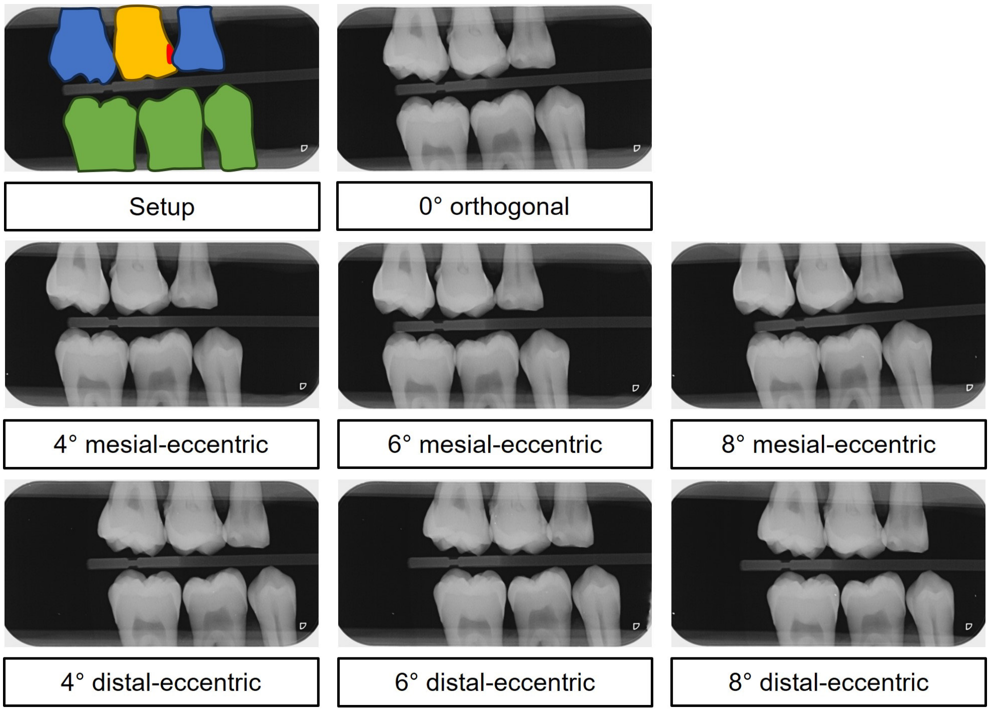

Figure 4. Digital in vitro bitewing images. Top: color-coded setup—yellow: examination tooth, red: carious lesion, blue: adjacent tooth, green: antagonistic tooth. Below: The mesial-eccentric series shows increased superimposition as the ray path becomes increasingly eccentric in the proximal region of teeth 46 and 47. Conversely, the distal-eccentric series shows increased superimposition as the ray path becomes increasingly eccentric in the interproximal region of teeth 15 and 16.

2.6. Preparation of Histological Samples

The sample preparation steps are shown in Figure 5.

After radiography, all carious teeth were subjected to an adapted standardized histological examination procedure (Figure 5) [35]. This was an elaborate process, beginning with a six-day ascending dehydration series with increasing concentrations of ethanol, followed by a six-day resin infiltration (Technovit 7200 VLC; Kulzer GmbH & Co. KG, Wehrheim, Germany) to effectively preserve carious lesions for subsequent processing (Table 1).

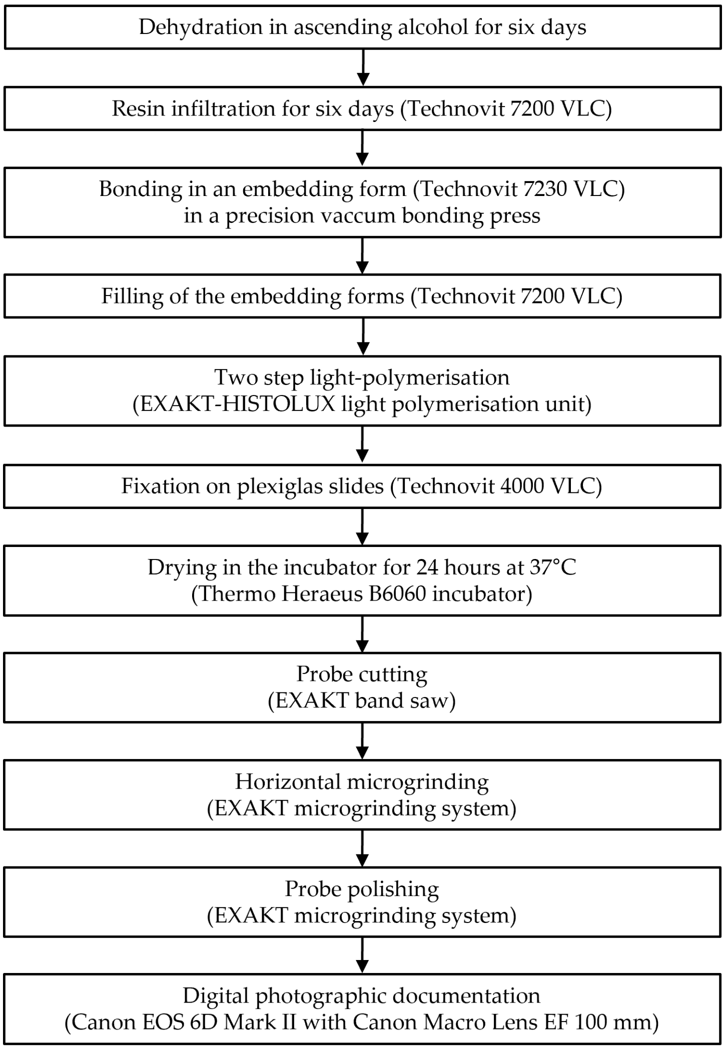
Dehydration in ascending alcohol for six days
Resin infiltration for six days (Technovit 7200 VLC)
Bonding in an embedding form (Technovit 7230 VLC) in a precision vaccum bonding press
Filling of the embedding forms (Technovit 7200 VLC)
Two step light-polymerisation (EXAKT-HISTOLUX light polymerisation unit)
Fixation on plexiglas slides (Technovit 4000 VLC)
Drying in the incubator for 24 hours at 37°C (Thermo Heraeus B6060 incubator)
Probe cutting (EXAKT band saw)
Horizontal microgrinding (EXAKT microgrinding system)
Probe polishing (EXAKT microgrinding system)
Digital photographic documentation (Canon EOS 6D Mark II with Canon Macro Lens EF 100 mm)

Figure 5. Sample preparation steps.

Table 1. Schematic overview of tooth dehydration and resin infiltration.

Time	Chemical Solution	Volume Ratio	Storage
Day 1	EtOH 70%/purif. H_2O	100	50 mL plastic flacons
Day 2	EtOH 80%/purif. H_2O	100	50 mL plastic flacons
Day 3	EtOH 90%/purif. H_2O	100	50 mL plastic flacons
Day 4	EtOH 96%/purif. H_2O	100	50 mL plastic flacons
Day 5 and 6	EtOH 99%	100	50 mL plastic flacons
Day 7	EtOH 99%/Technovit 7200 VLC	50/50	50 mL plastic flacons (in darkness)
Day 8	Technovit 7200 VLC	100	Snap-on lid tablet glass (in darkness)
Day 9	Technovit 7200 VLC	100	Snap-on lid tablet glass (in darkness)
Day 10	Technovit 7200 VLC	100	Snap-on lid tablet glass (in darkness)
Day 11	Technovit 7200 VLC	100	Snap-on lid tablet glass (in darkness)
Day 12	Technovit 7200 VLC	100	Snap-on lid tablet glass (in darkness)

purif. = purified, EtOH = ethanol.

The (carious) teeth were then bonded (Technovit 7230 VLC; Kulzer GmbH & Co. KG, Wehrheim, Germany), vestibular side down, to an embedding form (Kulzer GmbH & Co. KG, Wehrheim, Germany) using a disposable spatula and cured with UV light for 10 min in a precision vacuum bonding press (EXAKT Apparatebau GmbH & Co. KG, Norderstedt, Germany). The forms were filled with embedding resin (Technovit 7200 VLC; Kulzer GmbH & Co. KG, Wehrheim, Germany) using a disposable pipette. Pre-polymerization was performed in an EXAKT-HISTOLUX light polymerization unit (Exakt Apparatebau GmbH & Co. KG, Nordstedt) with two UV lamps for two hours, followed by the actual polymerization with eight UV lamps for a further eight hours. The polymerized blocks were fixed to Plexiglas slides (Walter-Messner GmbH, Oststeinbek, Germany) using mixed Technovit 4000 (Kulzer GmbH & Co. KG, Wehrheim, Germany) and cured with UV light for 10 min in the precision vacuum bonding press (EXAKT Apparatebau GmbH & Co. KG, Norderstedt, Germany). Before further processing, the samples were dried in an incubator (Thermo Heraeus B6060 incubator; Heraeus Holding GmbH, Hanau, Germany) for 24 h at 37 °C.

The (carious) teeth were sectioned directly in front of the lesion using a saw with a diamond-coated band 100 µm wide (EXAKT Apparatebau GmbH & Co. KG, Norderstedt, Germany) under constant water cooling. Due to inherent vibrations and the cutting width of the saw blade, a loss of tooth substance of approximately 300 µm per cut (slice) was assumed. During the cutting process, the block was fixed to the machine by a vacuum pump at 680 mbar and pulled through the saw blade by a constant force of 400 g (40 N). The hard-cut method was used to divide the carious teeth before the lesion reached its maximum extent.

This was followed by a meticulous, progressive approach to the carious defect using the wet grinding technique with the EXAKT horizontal microgrinding system and a 400 g press weight (EXAKT Apparatebau GmbH & Co. KG, Norderstedt, Germany) (Figure 6). The microgrinding unit was calibrated by grinding a microscope slide with 1200 grit Al_2O_3 sandpaper for two minutes. A difference in the slide of no more than 5 µm at four different measuring points was considered acceptable. The final step was polishing with the EXAKT horizontal microgrinding system using 2400 and 4000 grit Al_2O_3 sandpaper, with each incremental step documented by digital photographic records with a digital single-lens reflex camera (Canon EOS 6D Mark II; Canon Deutschland GmbH, Krefeld, Germany) and a macro lens (Canon Macro Lens EF 100 mm, Canon Deutschland GmbH, Krefeld, Germany) to illustrate the maximum extent of the lesion in the mesiodistal direction. The removal of tooth material between grinding was determined by measuring the thickness with a micrometer screw (EXAKT Apparatebau GmbH & Co. KG, Norderstedt, Germany).

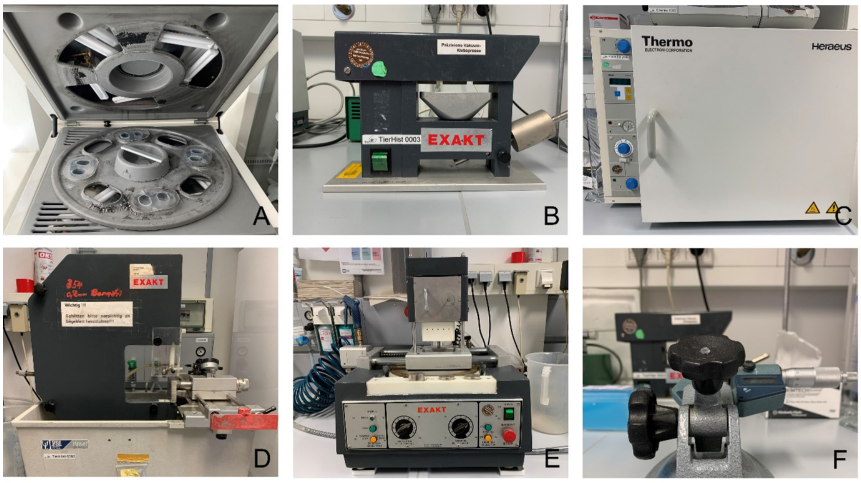

Figure 6. Equipment for the preparation of histological specimens. (**A**): EXAKT-HISTOLUX light polymerization unit, (**B**) EXAKT precision vacuum bonding press, (**C**): Thermo Heraeus B6060 Incubator, (**D**): EXAKT band saw, (**E**): EXAKT horizontal microgrinding system, (**F**): EXAKT micrometer screw.

2.7. Lesion Classification of the Histological Samples

All histological specimens, with the maximum extent of the carious lesion in mesio-distal direction, were digitally photographed and displayed on a diagnostic monitor (Nio Color 2 MP LED; Barco, Kortrijk, Brussels, Belgium) with no time limit (Figure 7). A review was performed twice at three-month intervals by an expert with extensive professional and scientific experience, following the common radiographic classification scheme (Table 2).

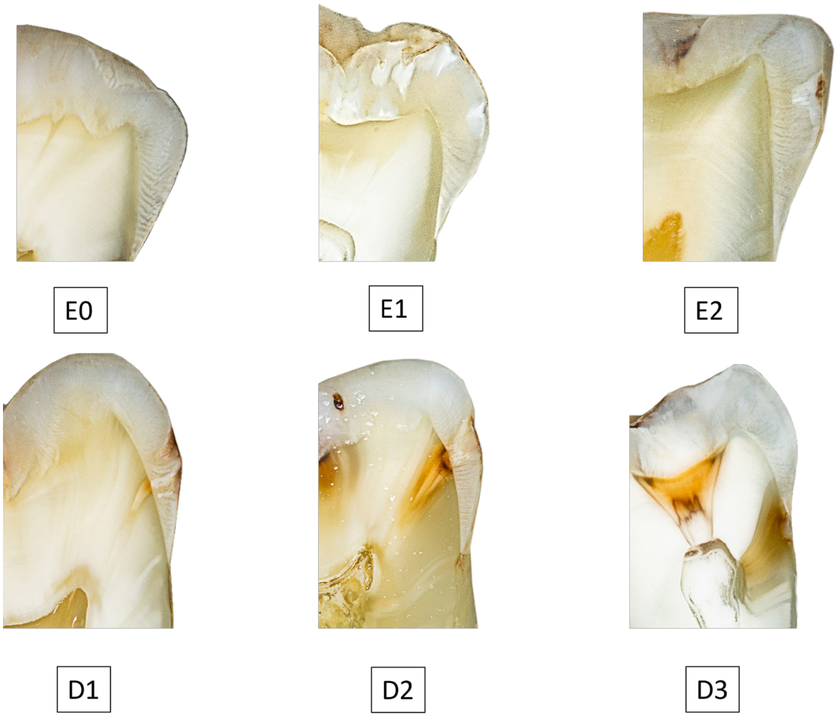

Figure 7. Histological specimen with different proximal carious lesion depths. E0 = Caries-free, E1 = Caries limited to the outer half of the enamel, E2 = Caries extending to the inner half of the enamel, D1 = Caries in the outer third of dentin, D2 = Caries in the middle third of dentin, D3 = Caries in the dentinal third close to the pulp or up to the pulp.

Table 2. Caries classification scheme.

Classification of Caries	Carious Lesion Extension
E1	Caries limited to the outer half of the enamel
E2	Caries extending to the inner half of the enamel
D1	Caries in the outer third of dentin
D2	Caries in the middle third of dentin
D3	Caries in the dentinal third close to the pulp or up to the pulp

The characteristics of the histological analysis are summarized in Table 3.

Table 3. Number of histologically confirmed carious lesions and their categorization according to the caries classification scheme.

Caries Classification	Proximal Surfaces	Percentage
E1	15	22.1%
E2	8	11.8%
D1	8	11.8%
D2	18	26.4%
D3	19	27.9%
	68	100%

2.8. Radiographic Caries Diagnostic by Dental Examiners

To benchmark dental examiners when analyzing in vitro bitewing images, 10 clinicians, 10 private practitioners and 10 students were asked to evaluate these radiographs.

Clinicians were defined as dentists providing care in a hospital setting, whereas private practitioners were defined as dentists working independently outside an institutional setting, usually in their own private practice. As a baseline, all participants were informed that all teeth would be examined for the presence or absence of proximal caries. Each participant evaluated a random selection of 35 to 36 bitewing images on a dental diagnostic monitor (Nio Color 2 MP LED; Barco, Kortrijk, Brussels, Belgium) without a time limit. All examiners were categorized according to gender, occupation and professional experience to assess respective influence on the quality of caries findings in bitewing radiographs.

2.9. Statistical Analysis and Performance Metrics

Statistical analyses were performed using R (version 4.3.2). Quality of carious lesion classification was determined by assessing the intrarater reliability using the intraclass correlation coefficient (ICC). The performance of the combined examiners was assessed using several metrics, including sensitivity, specificity, accuracy, positive and negative predictive values (PPV/NPV), area "under the curve" (AUC), F1 score and Matthews correlation coefficient (MCC).

The F1 score, a harmonic mean of precision and sensitivity, is a commonly used metric for binary classifier evaluation and ranges from 0 to 1, with higher values indicating superior classifier performance. It is defined as $2 \times \frac{(\text{PPV} \times \text{sensitivity})}{(\text{PPV} + \text{sensitivity})}$.

The Matthews Correlation Coefficient (MCC) is another key parameter for evaluating predictions against actual values and provides a reliable assessment of performance. The MCC is defined as $\frac{(\text{TN} \times \text{RP}) - (\text{FN} \times \text{FP})}{\sqrt{(\text{TP}+\text{FP}) \times (\text{T P}+\text{FN}) \times (\text{TN}+\text{FP}) \times (\text{TN}+\text{FN})}}$. An MCC value of 1 indicates a perfect prediction, while -1 indicates a complete disagreement between prediction and observation, and 0 indicates a random prediction. By including true negatives, false positives, false negatives and true positives, the MCC provides a comprehensive assessment of the predictive accuracy of the system or examiner under investigation.

De Long's test was used to compare the receiver operating characteristic (ROC) curves of histology and examiners. In addition, MCC scores were tested for differences in correlation using Bonferroni correction to compare performance across varying eccentricities of the central X-ray beam, the different carious lesion depths, gender, occupation and experience. The ability of the examiners to discriminate between artificially induced defects and true caries was investigated by comparing correct and incorrect predictions.

2.10. Sample Sice Planning

Our sample size planning was based on the number of bitewing radiographs required for accurate and reliable AI-assisted caries detection. We reviewed recent studies in this area and found that the number of bitewing radiographs used ranged from 45 to 252, with an average of 114 [9,36–40]. Due to the wide variation in the number of bitewing radiographs used in the literature, we used significantly more radiographs for testing in our study, with a total of 371 bitewing radiographs of 53 carious teeth. It can, therefore, be concluded that our sample size provides a robust dataset for evaluation.

3. Results

3.1. Examiner Characteristics

The metrics for all examiners are shown in Table 4. Private practitioners, clinicians and students were equally represented with ten examiners each. The private practitioners were almost equally divided between six examiners with less than five years' experience and four examiners with five or more years' experience. However, there was some imbalance between the two groups, with four male and six female private practitioners. All ten clinicians were evenly split between those with less than five years' experience and those with five or more years' experience, as was the gender split with five males and five females. There were eight female students compared to two males. Eight of the eleven examiners with less than five years' experience were male, followed by three female examiners in this group. In the group of nine examiners with five or more years' experience, there were three male and six female examiners. Of the thirty examiners, thirteen were male and seven were female.

Table 4. Examiner characteristics.

	Occupation			Experience		Gender	
	Private Practitioners	**Clinicians**	**Students**	**<5 Years**	**≥5 Years**	**Male**	**Female**
Occupation							
Private practitioners	10	-	-	6	4	6	4
Clinicians	-	10	-	5	5	5	5
Students	-	-	10	-	-	2	8

3.2. Reliability of Histological Lesion Classification

Intrarater reliability was very high throughout both assessment rounds (ICC: 0.993; 95%-CI [0.990; 0.995]). In two cases where the expert's categorization of carious lesions differed between the two rounds of examination, a second expert was consulted to determine the final lesion class.

3.3. Examiners Performance Metrics

All examiners reached a combined accuracy of 0.799, a sensitivity of 0.565, a specificity of 0.956, a PPV of 0.896, a NPV of 0.765, an AUC of 76.1, a F1 score of 0.693 and a MCC of 0.578 (Table 5).

Table 5. Combined examiners' performance metrics for caries detection.

Parameter	
Accuracy	0.799
Sensitivity	0.565
Specificity	0.956
PPV	0.896
NPV	0.765
F1 score	0.693
MCC	0.578
AUC	76.1

Note. AUC = area under the curve, MCC = Matthews correlation coefficient, NPV = negative predictive value, PPV = positive predictive value.

3.4. AUC

All examiners achieved a combined AUC of 76.1, whereas histology, serving as the gold standard method in caries diagnostic research, was assigned an AUC of 100 (Figure 8). Statistical analysis using De Long's test to compare the two ROC curves revealed a significantly higher performance for histology compared to the examiners' assessments (p < 0.001).

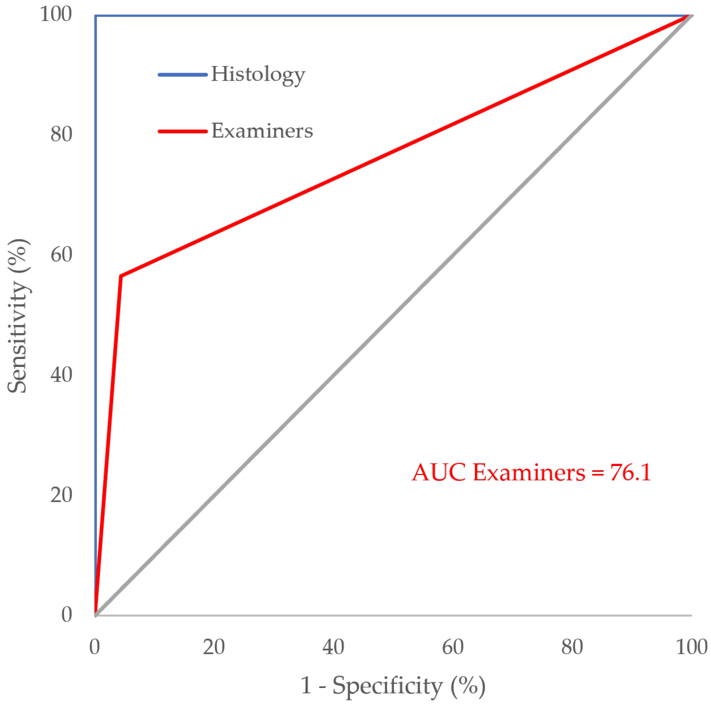

Figure 8. AUC of examiners and histology.

3.5. MCC by Lesion Class

The MCC showed variation according to the penetration depth of the carious lesions, with the best performance observed for D3 lesions (0.814), whereas E2 lesions showed the least favorable result (0.236) (Figure 9). The aggregated MCC for all lesion categories was 0.587.

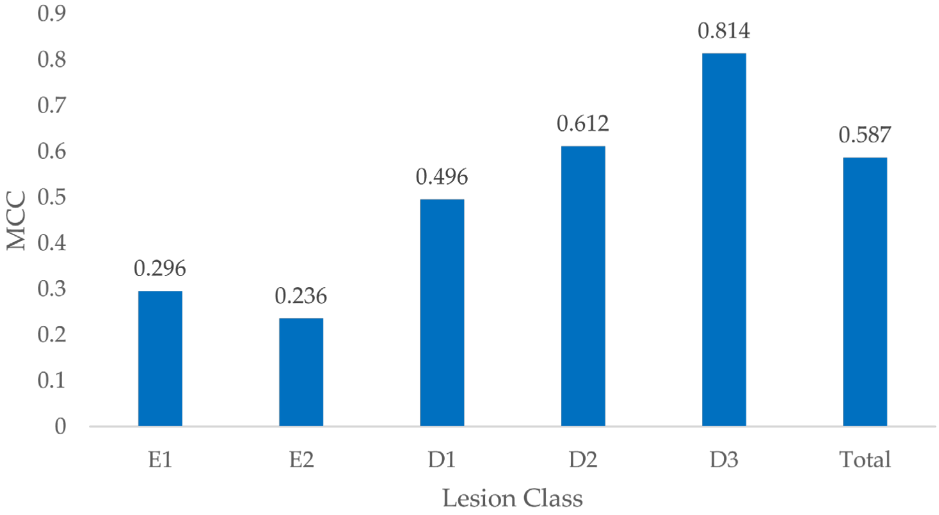

Figure 9. MCC by lesion class.

Testing for differences in MCC between different caries classifications revealed significant differences between all lesion classes ($p < 0.008$) except between E1 and E2 lesions ($p = 1$) (Table 6).

Table 6. Adjusted *p*-values for MCC comparison between lesion classes.

	Lesion Classification				
	E1	**E2**	**D1**	**D2**	**D3**
Lesion classification					
E1	-	1	<0.001 *	<0.001 *	<0.001 *
E2	1	-	<0.001 *	<0.001 *	<0.001 *
D1	<0.001 *	<0.001 *	-	0.008 *	<0.001 *
D2	<0.001 *	<0.001 *	0.008 *	-	<0.001 *
D3	<0.001 *	<0.001 *	<0.001 *	<0.001 *	-

Note. * Indicates significance (adjusted $p < 0.05$).

3.6. Gender Specific MCC

The MCC of male examiners was higher at 0.605 compared to the MCC of female examiners at 0.575 (Figure 10). However, testing for differences in MCC showed no significant effect of gender ($p = 0.44$).

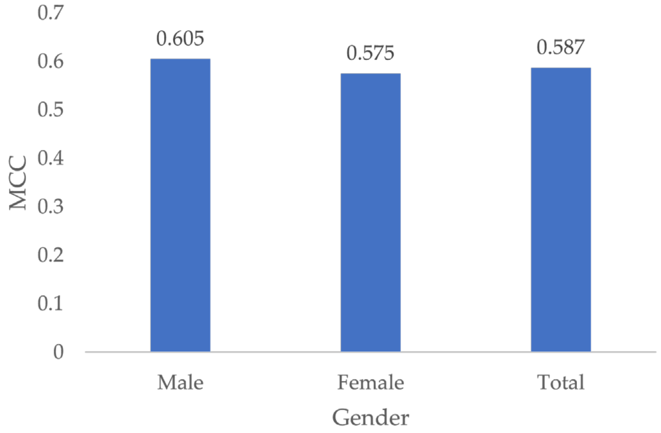

Figure 10. MCC by gender.

3.7. MCC by Occupation

Private practitioners had the highest MCC (0.595), followed by students (0.593) and clinical practitioners (0.571) (Figure 11).

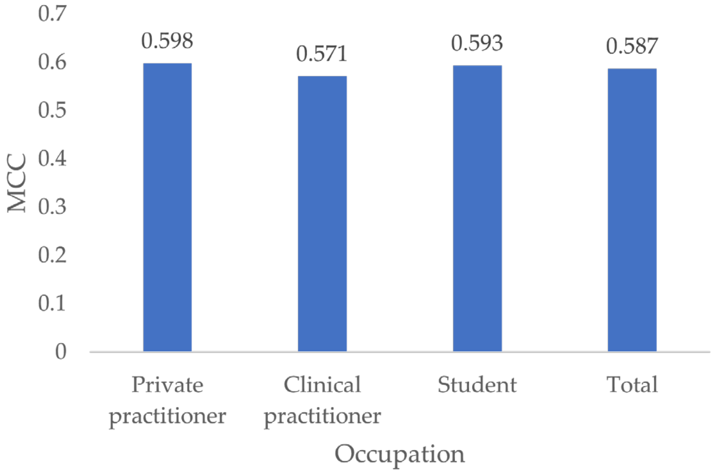

Figure 11. MCC by occupation.

Testing for differences in MCC showed no significant differences between all occupations ($p \geq 0.556$).

3.8. MCC by Experience

Dentists with less than 5 years of experience showed the best MCC (0.611), followed by students (0.593) and dentists with 5 or more years of experience (0.551) (Figure 12).

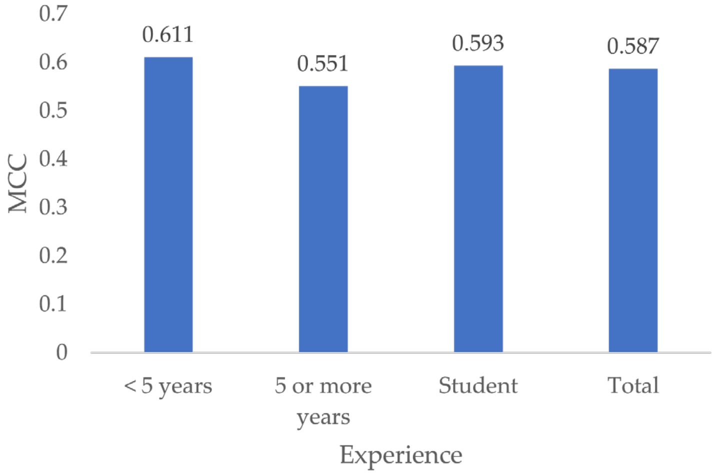

Figure 12. MCC by experience.

No significant differences were found by testing for differences in MCC according to experience ($p = 1$).

3.9. Influence of Eccentricity on MCC

Different eccentricity angles resulted in different MCC values (Figure 13).

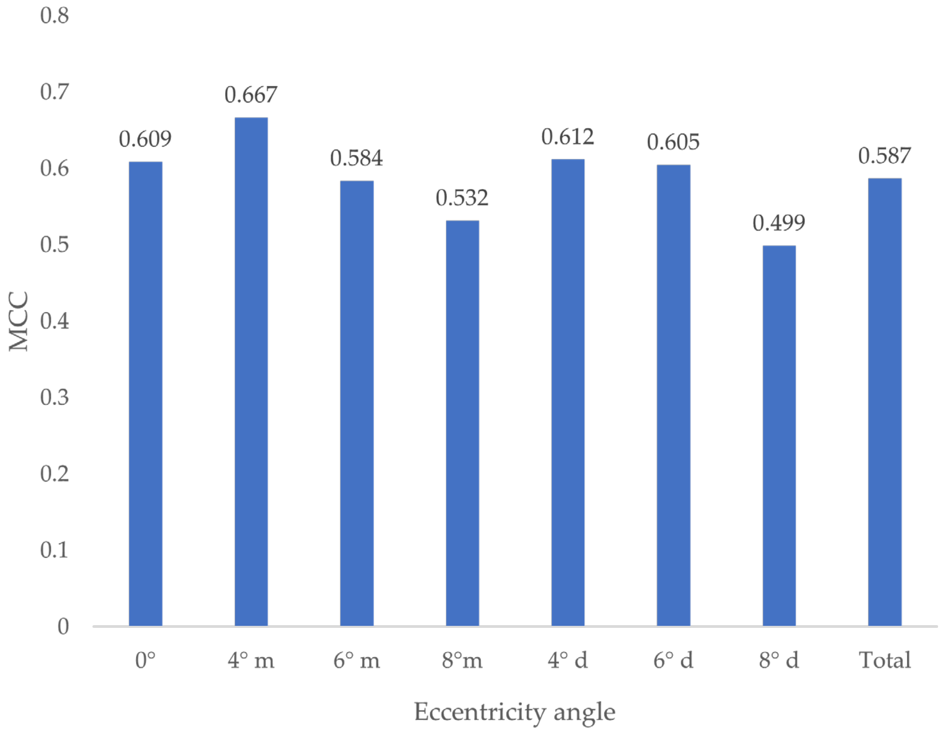

Figure 13. MCC by eccentricity angle. m = mesial-eccentric, d = distal-eccentric.

No statistically significant difference between the groups could be found ($p \geq 0.411$).

3.10. Differentiation between Carious Lesions and Artifically Induced Lesions

Out of a total of 350 artificial defects presented, 159 defects (45.4%) were identified as carious lesions by all examiners and 191 defects (54.6%) were identified as atypical for caries (Figure 14).

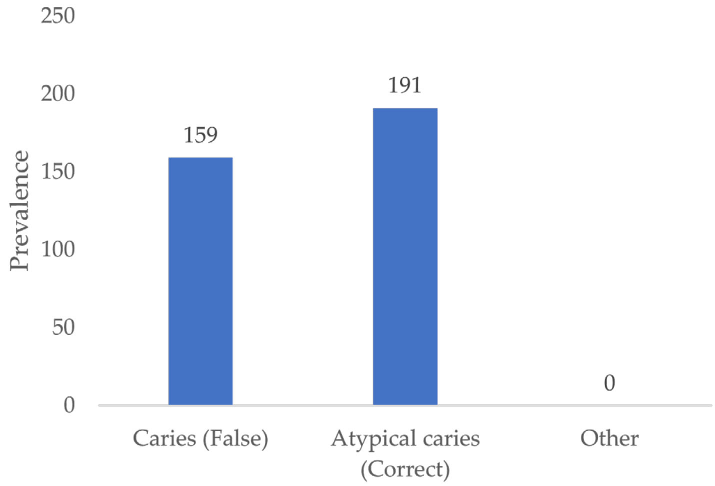

Figure 14. Assessment of artificially induced lesions.

3.11. Tooth Classification

The results indicate that 99.8% of the examiners correctly positioned the teeth depicted in the bitewing simulations according to the World Dental Federation (FDI) tooth numbering system (Figure 15).

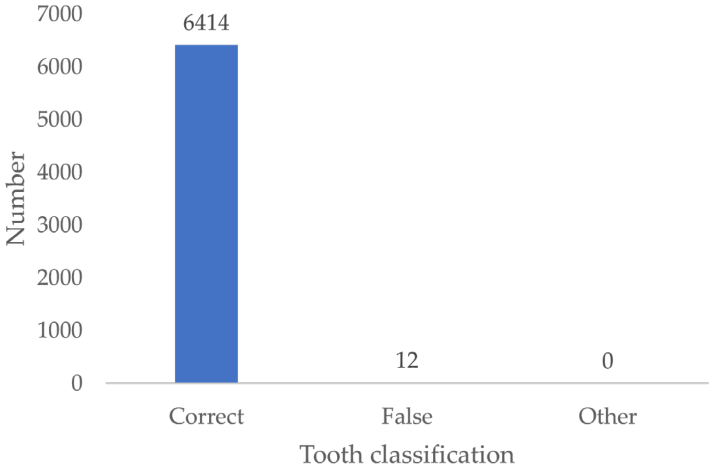

Figure 15. Tooth classification according to the FDI scheme.

4. Discussion

The European Medical Device Regulation (MDR) classifies medical imaging software as a medical device and, therefore, imposes several requirements on manufacturers to ensure safety and quality. Among other things, manufacturers are required to conduct a comprehensive clinical evaluation of their medical devices. As AI-based imaging software for caries diagnosis has been approved as a medical device, the underpinnings deserve scrutiny. The aim of this study was, therefore, to create a pool of histology-based radiographs to provide a scientifically sound testbed for such software. We are currently unaware of the existence of such a dataset.

In the context of fuzzy gold standards, several mitigation strategies have been proposed. One approach aims to supplement existing datasets with additional data from external sources [41]. By incorporating different perspectives, especially in cases where the gold standard may be imperfect, this strategy aims to improve the robustness of AI models and mitigate bias. The use of multiple diagnostic tests is also encouraged, as this can increase the transparency and reliability of diagnostic results [41]. Despite these efforts, the almost complete elimination of bias in AI-based dental caries diagnostics will, at least for an extended period, remain an elusive goal.

In general, in vitro studies provide a robust method for validating new caries diagnostic methods because they can refer to a reliable gold standard by means of histological analysis. Also, literature states that histological examination shall serve as the basis for a gold standard for the evaluation of new caries diagnostic methods [42]. Therefore, the ideal, albeit theoretical, method for evaluating diagnostic accuracy would be to first assess the diagnoses in vivo and then re-examine the same surfaces in vitro after tooth extraction using the histological gold standard [43]. However, logistical constraints and ethical considerations associated with invasive procedures, particularly the need for extraction, make this approach infeasible. Furthermore, it has been argued that differences between in vivo and in vitro results may cast doubt on the generalizability of in vitro data [43]. Nevertheless, previous studies have confirmed that no significant difference in the diagnostic accuracy of proximal carious lesions on digital radiographs can be demonstrated between in vivo and in vitro settings [44,45].

To further ensure the applicability of our results to the clinical situation, we attempted to create clinical simulations of the orofacial region on bitewing radiographs that are as realistic as possible. Nevertheless, given the complexity of the human body, accurate reproduction of anatomical structures remains difficult. To account for potential uncertainties, only findings within the coronal region were considered. This approach was intended to reduce possible distortions caused by the setup, particularly the fixation material. A limitation concerns the in vitro radiographs that did not consider external factors that could have influenced the accuracy of the radiographic diagnosis, such as the influence of metal artefacts, patient movement or incorrect positioning of the film holder on the analysis results. For reasons of standardization, all bitewing radiographs were taken on a single X-ray unit, to account for unintended variations.

For the purpose of disinfection and protection against dehydration, all extracted teeth were immersed in 1% tosylchloramide. Previous studies have shown that tosylchloramide has no discernible effect on tooth hard tissue [46–49]. A possible influence of tosylchloramide storage on the infiltration behavior of Technovit cannot be completely excluded, however it seems unlikely in view of the high success rate of histological preparations. All teeth were obtained from a variety of sources, including dental, oral and maxillofacial surgery practices and clinics. This diverse selection supports the assumption of a representative assortment of teeth across different population groups.

As already mentioned, histological examination serves as the most widely used gold standard for the validation of new caries diagnostic methods [42]. Its substantial diagnostic quality and value have been highlighted in many publications [50,51]. A major criticism of histological examinations is the frequent bisection of teeth through an arbitrary centerline [52]. This carries the risk of irreversibly destroying the presumed maximum extent of the carious lesion, thereby obscuring the true maximum depth. To overcome this, the incision was positioned anterior to the carious lesion, and the wet grinding technique was used to approach the maximum extent of the lesion. This approach ensured that the deepest carious extension was accurately identified with a high degree of confidence. The use of final multi-stage polishing ensured a consistent surface quality for subsequent expert analysis.

In our study, all 30 examiners showed a combined accuracy, sensitivity, specificity and AUC of 0.799, 0.565, 0.956 and 76.1, respectively, for the detection of carious lesions on bitewing radiographs. The literature shows a wide range of results. Kay and Knill-Jones observed a dentist sensitivity of 0.26 for the detection of dentin caries on in vitro bitewing radiographs [53]. Devlin et al. showed a sensitivity of 44% for enamel-limited lesions on bitewing radiographs among 23 examiners [54]. Mileman and van der Welle reported an

AUC of 0.88 with a sensitivity of 0.54 and specificity of 0.97 for dentin caries on bitewing radiographs. Similarly, Peers et al. demonstrated a comparable sensitivity of 0.59 for the detection of dentin caries on bitewing radiographs [55]. It, therefore, can be assumed that the results of our study are consistent with the literature, as we also could demonstrate that carious lesion depth had a significant effect on the MCC of all examiners between all lesion classes, except between enamel-limited E1 and E2 lesions. We support the assumption that in vitro radiographs provide diagnostic quality parameters similar to studies using in vivo radiographs.

Our results also showed that, contrary to expectations, the eccentricity of the central X-ray beam up to 8°, whether mesial or distal, did not appear to have a significant effect on the examiner's judgement of the presence or absence of caries. The lack of significant impact from minor eccentricities humbly suggests that clinicians may not need to be overly concerned about small variations in radiographic positioning when assessing for caries. Like our results, the study by Deprá et al. investigated the influence of the central opening angle on the diagnosis of secondary caries and also concluded that it had no influence [56]. On the other hand, Chadwick et al. investigated the influence of different central irradiation angles on visualization of proximal cavities in bitewing radiographs and found that lesions are typically diagnosed, often resulting in overtreatment [57]. However, as both comparative studies do not provide information on the size of the eccentricity examined, we are, to the best of our knowledge, the first study to provide results with tangible values.

In the present study, no significant effect of examiner experience could be demonstrated. The results, thus, contradict the findings of Geibel et al., which have shown that experienced examiners detect proximal lesions up to four times more frequently than less experienced examiners [58]. A plausible explanation for this difference could be that dental students and practicing dentists with less than five years of professional experience in our study took more time to analyze in vitro bitewing images than their colleagues with five or more years of clinical experience, as the time factor has been demonstrated to influence diagnostic accuracy significantly [31].

It was found that just over half (54.6%) of the artificial lesions were judged by the examiners to be atypical for caries, effectively distinguishing them from true carious lesions. This observation highlights the ability of human examiners to differentiate iatrogenic defects, such as those resulting from invasive treatments resulting from treatments of the adjacent teeth, from true caries cases, primarily through the assessment of lesion morphology. To the best of our knowledge, this study represents the first attempt to establish a framework for evaluating AI algorithms in this regard and to compare their performance with human judgement.

The empirical evaluation of binary classification tasks, such as the distinction between caries and healthy tooth structure, is subject of discussion. It must be noted that accuracy, as a metric, comes with the significant limitation of sensitivity to unbalanced datasets, potentially limiting the validity of the results. As the Fifth German Oral Health Study has already confirmed, caries prevalence is decreasing in all age groups, increasing the imbalance between carious and non-carious teeth on radiographs. Therefore, the suitability of accuracy to determine diagnostic quality must strongly be questioned [59]. Furthermore, Dinga et al. recommend completely omitting accuracy as sole criterion for evaluating clinical models, as it fails to take into account clinically relevant information [60]. Nevertheless, accuracy is still somewhat stubbornly used as the main parameter for performance evaluation in the literature. For the sake of comparability, we have included this metric, but explicitly point out its shortcomings. Positive predictive value (PPV), sensitivity, specificity and the F1 score, which is the harmonic mean of precision and recall, are commonly used parameters to evaluate binary classifiers [61]. However, these metrics assume that the "positive" class (in this case a detection of caries) is of primary interest, while true negatives are omitted in their calculation. Consequently, PPV, sensitivity and F1 scores are unaffected by variations in the number of true negatives, whether their value is extremely high or low. To overcome this limitation, we made use of Matthews correlation coefficient (MCC). MCC gives high values only when the predictions of all categories (true positives, true negatives, false positives and false negatives) show good performance, also considering the proportions of the positive and negative classes. As a result, the MCC is a statistically robust measure, even in the presence of unbalanced datasets.

5. Conclusions

The aim of this study was to establish a histology-based gold standard for the unbiased evaluation of AI-based caries detection systems on proximal surfaces in bitewing radiographs. Through meticulous in vitro simulations and histological analyses, we created a robust dataset to evaluate the performance of AI algorithms in caries detection and compare it to human judgement. Although AI promises to improve diagnostic accuracy and workflow efficiency, its effectiveness depends primarily on the quality of the training data and validation processes. Future research should be designed to accurately reflect the true performance of AI models using histological analysis as a benchmark. In doing so, we have laid the foundation for evaluating the real-world performance of AI systems, thereby advancing evidence-based dentistry. Ongoing advances in AI technology and regulatory frameworks require continuous refinement and validation of diagnostic tools to ensure patient safety and clinical effectiveness. The creation of a standardized database of reference histological specimens and associated radiographs could serve as a benchmark for the development and validation of new AI-based caries detection systems. This database would allow different AI systems to be compared and their performance tested against an established gold standard, helping to identify and develop the most accurate models. However, generating a histology-based dataset is time consuming and requires resources and equipment. Therefore, a simple histology-based implementation will not be readily available in the future. In addition, it remains to be seen whether newer intraoral caries detection techniques will provide higher sensitivity, which could serve as a solid basis for training dental AI systems. In conclusion, our study is an important step towards the creation of standardized evaluation protocols for AI-based caries detection, thereby promoting transparency, reliability and confidence in dental diagnostics.

Author Contributions: M.S. (Matthias Schuster), J.B., J.P. and F.K. contributed to the conception, design, data acquisition, statistical analyses and interpretation, drafting and critical revision of the manuscript; G.K., M.S. (Marc Schmitter) and A.S.-E. contributed to conception, data interpretation, drafting and critical revision of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The authors received no financial support for the research, authorship and/or publication of this article. This work was entirely self-funded by Matthias Schuster and Julian Boldt.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the University of Würzburg (protocol code 15/15, 9 February 2015).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to ethical restrictions.

Conflicts of Interest: The authors declare no potential conflicts of interest concerning the research, authorship and/or publication of this article.

Abbreviations

AI	Artificial intelligence
AUC	Area under the curve
EtOH	Ethanol
ICC	Intraclass correlation coefficient
NPV	Negative predictive value
MCC	Matthews correlation coefficient
PPV	Positive predictive value
ROC	Receiver operating characteristics
STARD	Standards for Reporting of Diagnostic Accuracy Studies
UV	Ultraviolet

References

1. Schwendicke, F.; Samek, W.; Krois, J. Artificial Intelligence in Dentistry: Chances and Challenges. *J. Dent. Res.* **2020**, *99*, 769–774. [CrossRef] [PubMed]
2. Ahmed, N.; Abbasi, M.S.; Zuberi, F.; Qamar, W.; Halim, M.S.B.; Maqsood, A.; Alam, M.K. Artificial Intelligence Techniques: Analysis, Application, and Outcome in Dentistry-A Systematic Review. *Biomed Res. Int.* **2021**, *2021*, 9751564. [CrossRef]
3. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. [CrossRef]
4. Khanagar, S.B.; Al-Ehaideb, A.; Maganur, P.C.; Vishwanathaiah, S.; Patil, S.; Baeshen, H.A.; Sarode, S.C.; Bhandi, S. Developments, application, and performance of artificial intelligence in dentistry—A systematic review. *J. Dent. Sci.* **2021**, *16*, 508–522. [CrossRef]
5. Schwendicke, F.; Rossi, J.G.; Göstemeyer, G.; Elhennawy, K.; Cantu, A.G.; Gaudin, R.; Chaurasia, A.; Gehrung, S.; Krois, J. Cost-effectiveness of Artificial Intelligence for Proximal Caries Detection. *J. Dent. Res.* **2021**, *100*, 369–376. [CrossRef]
6. Schwendicke, F.; Paris, S.; Stolpe, M. Detection and treatment of proximal caries lesions: Milieu-specific cost-effectiveness analysis. *J. Dent.* **2015**, *43*, 647–655. [CrossRef]
7. Alam, M.K.; Alftaikhah, S.A.A.; Issrani, R.; Ronsivalle, V.; Lo Giudice, A.; Cicciù, M.; Minervini, G. Applications of artificial intelligence in the utilisation of imaging modalities in dentistry: A systematic review and meta-analysis of in-vitro studies. *Heliyon* **2024**, *10*, e24221. [CrossRef]
8. Obuchowicz, R.; Strzelecki, M.; Piórkowski, A. Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing—A Review. *Cancers* **2024**, *16*, 1870. [CrossRef]
9. Bayraktar, Y.; Ayan, E. Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs. *Clin. Oral Investig.* **2022**, *26*, 623–632. [CrossRef] [PubMed]
10. Zhu, H.; Cao, Z.; Lian, L.; Ye, G.; Gao, H.; Wu, J. CariesNet: A deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image. *Neural Comput. Appl.* **2022**, *35*, 1–9. [CrossRef]
11. Park, E.Y.; Cho, H.; Kang, S.; Jeong, S.; Kim, E.-K. Caries detection with tooth surface segmentation on intraoral photographic images using deep learning. *BMC Oral Health* **2022**, *22*, 573. [CrossRef] [PubMed]
12. Chen, X.; Guo, J.; Ye, J.; Zhang, M.; Liang, Y. Detection of Proximal Caries Lesions on Bitewing Radiographs Using Deep Learning Method. *Caries Res.* **2022**, *56*, 455–463. [CrossRef] [PubMed]
13. Obuchowicz, R.; Nurzynska, K.; Obuchowicz, B.; Urbanik, A.; Piórkowski, A. Caries detection enhancement using texture feature maps of intraoral radiographs. *Oral Radiol.* **2020**, *36*, 275–287. [CrossRef] [PubMed]
14. Anil, S.; Porwal, P.; Porwal, A. Transforming Dental Caries Diagnosis Through Artificial Intelligence-Based Techniques. *Cureus* **2023**, *15*, e41694. [CrossRef] [PubMed]
15. Roosanty, A.; Widyaningrum, R.; Diba, S.F. Artificial intelligence based on Convolutional Neural Network for detecting dental caries on bitewing and periapical radiographs. *J. Radiol. Dentomaksilofasial Indones. (JRDI)* **2022**, *6*, 89–94. [CrossRef]
16. Kunz, F.; Stellzig-Eisenhauer, A.; Zeman, F.; Boldt, J. Artificial intelligence in orthodontics: Evaluation of a fully automated cephalometric analysis using a customized convolutional neural network. *J. Orofac. Orthop.* **2020**, *81*, 52–68. [CrossRef] [PubMed]
17. Kunz, F.; Stellzig-Eisenhauer, A.; Boldt, J. Applications of Artificial Intelligence in Orthodontics—An Overview and Perspective Based on the Current State of the Art. *Appl. Sci.* **2023**, *13*, 3850. [CrossRef]
18. Mohammad-Rahimi, H.; Motamedian, S.R.; Rohban, M.H.; Krois, J.; Uribe, S.E.; Mahmoudinia, E.; Rokhshad, R.; Nadimi, M.; Schwendicke, F. Deep learning for caries detection: A systematic review. *J. Dent.* **2022**, *122*, 104115. [CrossRef] [PubMed]
19. Ammar, N.; Kühnisch, J. Diagnostic performance of artificial intelligence-aided caries detection on bitewing radiographs: A systematic review and meta-analysis. *Jpn. Dent. Sci. Rev.* **2024**, *60*, 128–136. [CrossRef]
20. Gomez, J. Detection and diagnosis of the early caries lesion. *BMC Oral Health* **2015**, *15* (Suppl. 1), S3. [CrossRef]
21. Grieco, P.; Jivraj, A.; Da Silva, J.; Kuwajima, Y.; Ishida, Y.; Ogawa, K.; Ohyama, H.; Ishikawa-Nagai, S. Importance of bitewing radiographs for the early detection of interproximal carious lesions and the impact on healthcare expenditure in Japan. *Ann. Transl. Med.* **2021**, *10*, 2. [CrossRef] [PubMed]
22. Prados-Privado, M.; García Villalón, J.; Martínez-Martínez, C.H.; Ivorra, C.; Prados-Frutos, J.C. Dental Caries Diagnosis and Detection Using Neural Networks: A Systematic Review. *J. Clin. Med.* **2020**, *9*, 3579. [CrossRef]

23. Khanagar, S.B.; Alfouzan, K.; Awawdeh, M.; Alkadi, L.; Albalawi, F.; Alfadley, A. Application and Performance of Artificial Intelligence Technology in Detection, Diagnosis and Prediction of Dental Caries (DC)-A Systematic Review. *Diagnostics* **2022**, *12*, 1083. [CrossRef] [PubMed]
24. Albano, D.; Galiano, V.; Basile, M.; Di Luca, F.; Gitto, S.; Messina, C.; Cagetti, M.G.; Del Fabbro, M.; Tartaglia, G.M.; Sconfienza, L.M. Artificial intelligence for radiographic imaging detection of caries lesions: A systematic review. *BMC Oral Health* **2024**, *24*, 274. [CrossRef] [PubMed]
25. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef]
26. Bossuyt, P.M.; Reitsma, J.B.; Bruns, D.E.; Gatsonis, C.A.; Glasziou, P.P.; Irwig, L.; Lijmer, J.G.; Moher, D.; Rennie, D.; Vet, H.C.W.d.; et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *Radiology* **2015**, *277*, 826–832. [CrossRef] [PubMed]
27. Schwendicke, F.; Göstemeyer, G. Conventional Bitewing Radiographs. In *Detection and Assessment of Dental Caries: A Clinical Guide*, 1st ed.; Ferreira Zandona, A., Longbottom, C., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 109–117. [CrossRef]
28. Devlin, H.; Williams, T.; Graham, J.; Ashley, M. 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.* **2021**, *231*, 481–485. [CrossRef]
29. Vaarkamp, J.; Ten Bosch, J.J.; Verdonschot, E.H.; Bronkhorst, E.M. The Real Performance of Bitewing Radiography and Fiber-Optic Transillumination in Approximal Caries Diagnosis. *J. Dent. Res.* **2000**, *79*, 1747–1751. [CrossRef] [PubMed]
30. Bader, J.D.; Shugars, D.A.; Bonito, A.J. Systematic reviews of selected dental caries diagnostic and management methods. *J. Dent. Educ.* **2001**, *65*, 960–968. [CrossRef]
31. Hegde, S.; Gao, J.; Vasa, R.; Cox, S. Factors affecting interpretation of dental radiographs. *Dentomaxillofacial Radiol.* **2022**, *52*, 20220279. [CrossRef]
32. Stroud, J.L.; English, J.; Buschang, P.H. Enamel thickness of the posterior dentition: Its implications for nonextraction treatment. *Angle Orthod.* **1998**, *68*, 141–146. [PubMed]
33. Kamburoglu, K.; Kolsuz, E.; Murat, S.; Yüksel, S.; Ozen, T. Proximal caries detection accuracy using intraoral bitewing radiography, extraoral bitewing radiography and panoramic radiography. *Dentomaxillofacl Radiol.* **2012**, *41*, 450–459. [CrossRef]
34. Wenzel, A. Bitewing and digital bitewing radiography for detection of caries lesions. *J. Dent. Res.* **2004**, *83*, 72–75. [CrossRef]
35. Donath, K. *Die Trenn-Dünnschliff-Technik zur Herstellung Histologischer Präparate von nicht schneidbaren Geweben und Materialien: Apparate- und Methodenbeschreibung*; EXAKT-Kulzer-Druckschriften: Norderstedt, Germany, 1988.
36. Cantu, A.G.; Gehrung, S.; Krois, J.; Chaurasia, A.; Rossi, J.G.; Gaudin, R.; Elhennawy, K.; Schwendicke, F. Detecting caries lesions of different radiographic extension on bitewings using deep learning. *J. Dent.* **2020**, *100*, 103425. [CrossRef]
37. Bayrakdar, I.S.; Orhan, K.; Akarsu, S.; Çelik, Ö.; Atasoy, S.; Pekince, A.; Yasa, Y.; Bilgir, E.; Sağlam, H.; Aslan, A.F.; et al. Deep-learning approach for caries detection and segmentation on dental bitewing radiographs. *Oral Radiol.* **2022**, *38*, 468–479. [CrossRef]
38. Moran, M.; Faria, M.; Giraldi, G.; Bastos, L.; Oliveira, L.; Conci, A. Classification of Approximal Caries in Bitewing Radiographs Using Convolutional Neural Networks. *Sensors* **2021**, *21*, 5192. [CrossRef]
39. Lee, S.; Oh, S.I.; Jo, J.; Kang, S.; Shin, Y.; Park, J.W. Deep learning for early dental caries detection in bitewing radiographs. *Sci. Rep.* **2021**, *11*, 16807. [CrossRef]
40. Mao, Y.-C.; Chen, T.-Y.; Chou, H.-S.; Lin, S.-Y.; Liu, S.-Y.; Chen, Y.-A.; Liu, Y.-L.; Chen, C.-A.; Huang, Y.-C.; Chen, S.-L.; et al. Caries and Restoration Detection Using Bitewing Film Based on Transfer Learning with CNNs. *Sensors* **2021**, *21*, 4613. [CrossRef] [PubMed]
41. Walsh, T. Fuzzy gold standards: Approaches to handling an imperfect reference standard. *J. Dent.* **2018**, *74*, 47–49. [CrossRef]
42. Jablonski-Momeni, A.; Stachniss, V. Serial sectioning of teeth and microscopy in cariology research. In *Microscopy: Science, Technology, Applications and Education*, 4th ed.; Méndez-Vilas, A., Díaz, J., Eds.; FORMATEX: Badajoz, Spain, 2010; Volume 3, pp. 785–791.
43. Pitts, N.B. Clinical diagnosis of dental caries: A European perspective. *J. Dent. Educ.* **2001**, *65*, 972–978. [CrossRef]
44. Li, G.; Qu, X.-m.; Chen, Y.; Zhang, J.; Zhang, Z.-y.; Ma, X.-c. Diagnostic accuracy of proximal caries by digital radiographs: An in vivo and in vitro comparative study. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodontology* **2010**, *109*, 463–467. [CrossRef]
45. Hintze, H.; Wenze, A. Clinical and laboratory radiographic caries diagnosis. A study of the same teeth. *Dentomaxillofacl Radiol.* **1996**, *25*, 115–118. [CrossRef] [PubMed]
46. Suzuki, T.; Finger, W.J. Dentin adhesives: Site of dentin vs. bonding of composite resins. *Dent. Mater.* **1988**, *4*, 379–383. [CrossRef]
47. O'Brien, J.A., 3rd; Retief, D.H.; Bradley, E.L.; Denys, F.R. Shear bond strength of a new dentin bonding restorative system. *Dent. Mater.* **1988**, *4*, 179–183. [CrossRef]
48. Haller, B.; Hofmann, N.; Klaiber, B.; Bloching, U. Effect of storage media on microleakage of five dentin bonding agents. *Dent. Mater.* **1993**, *9*, 191–197. [CrossRef]
49. Söderholm, K.J.M. Correlation of in vivo and in vitro performance of adhesive restorative materials: A report of the ASC MD156 task group on test methods for the adhesion of restorative materials. *Dent. Mater.* **1991**, *7*, 74–83. [CrossRef]

50. Wenzel, A.; Hintze, H. Comparison of microscopy and radiography as gold standards in radiographic caries diagnosis. *Dentomaxillofacl Radiol.* **1999**, *28*, 182–185. [CrossRef] [PubMed]
51. Rodrigues, J.A.; Neuhaus, K.W.; Diniz, M.B.; Hug, I.; Stich, H.; Karlsson, L.; Lussi, A. Comparison among gold standard techniques used for the validation of methods for occlusal caries detection. *Microsc. Res. Tech.* **2012**, *75*, 605–608. [CrossRef] [PubMed]
52. Schulze, R.K.; Nackat, D.; D'Hoedt, B. In vitro carious lesion detection on D-, E-, and F-speed radiographic films. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodontology* **2004**, *97*, 529–534. [CrossRef]
53. Kay, E.J.; Knill-Jones, R. Variation in restorative treatment decisions: Application of Receiver Operating Characteristic curve (ROC) analysis. *Community Dent. Oral Endodontology* **1992**, *20*, 113–117. [CrossRef]
54. Mileman, P.A.; van der Weele, L.T. Accuracy in radiographic diagnosis: Dutch practitioners and dental caries. *J. Dent.* **1990**, *18*, 130–136. [CrossRef] [PubMed]
55. Peers, A.; Hill, F.J.; Mitropoulos, C.M.; Holloway, P.J. Validity and reproducibility of clinical examination, fibre-optic transillumination, and bite-wing radiology for the diagnosis of small approximal carious lesions: An in vitro study. *Caries Res.* **1993**, *27*, 307–311. [CrossRef]
56. Deprá Lde, C.; Vessoni Iwaki, L.C.; Chicarelli, M.; Takeshita, W.M. Influence of Image Filters and Variation in Horizontal Angle of Incidence of X-ray Beam in Digital Interproximal Radiographs for Diagnosis of Secondary Caries in Esthetic Restorations. *J. Contemp. Dent. Pract.* **2015**, *16*, 805–812. [CrossRef] [PubMed]
57. Chadwick, B.L.; Dummer, P.M.; van der Stelt, P.F. The effect of alterations in horizontal X-ray beam angulation and bucco-lingual cavity width on the radiographic depth of approximal cavities. *J. Oral Rehabil.* **1999**, *26*, 292–301. [CrossRef] [PubMed]
58. Geibel, M.-A.; Carstens, S.; Braisch, U.; Rahman, A.; Herz, M.; Jablonski-Momeni, A. Radiographic diagnosis of proximal caries—Influence of experience and gender of the dental staff. *Clin. Oral Investig.* **2017**, *21*, 2761–2770. [CrossRef]
59. Jordan, A.; Micheelis, W.; Cholmakow-Bodechtel, C.; Füßl-Grünig, E.; Geyer, S.; Hertrampf, K.; Hoffmann, T.; Holtfreter, B.; Kocher, T.; Nitschke, I.; et al. *Fünfte Deutsche Mundgesundheitsstudie (DMS V)*; Deutscher Zahnärzteverlag: Köln, Germany, 2016.
60. Dinga, R.; Penninx, B.W.J.H.; Veltman, D.J.; Schmaal, L.; Marquand, A.F. Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. *bioRxiv* **2019**. [CrossRef]
61. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.