Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Augmentation Leakage in Dental Radiograph Classification: Source-Grouped Evaluation and an External Cross-Institution, Cross-Modality Output-Shift Stress Test

Safad Ismail
Muthoot Institute of Technology &amp
Vicky Nair
Muthoot Institute of Technology &amp
Alvaro Rocha
Universidade Lusófona
Ashok Vaseashta
prof.vaseashta@ieee.org

International Clean Water Institute

Research Article

Keywords: dental radiography, data leakage, source-grouped cross-validation, cross-institution and cross-modality shift, caries detection, model evaluation

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11252166/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Augmentation Leakage in Dental Radiograph Classification: Source-Grouped Evaluation and an External Cross-Institution, Cross-Modality Output-Shift Stress Test

Safad Ismail[1], Vicky Nair[1], Alvaro Rocha[2], Ashok Vaseashta[3*]
[1] Dept of Cyber Security, Muthoot Institute of Technology & Science, Kochi, India, safadismail@mgits.ac.in (S.I.), vickynair@mgits.ac.in (V.N.)
[2] Universidade Lusófona, Porto, Portugal, amrocha@gmail.com
[3] Strategic Research, International Clean Water Institute, Manassas, USA, prof.vaseashta@ieee.org
Correspondence: prof.vaseashta@ieee.org

Abstract

Objectives:

This study investigates how data leakage from augmented images affects model performance and how well a model trained on panoramic radiographs performs on clinical periapical radiographs.

Methods:

A ResNet-50 multi-label classifier was trained using 1,269 files derived from 655 source panoramic radiographs covering four classes. The model was evaluated using source-grouped five-fold cross-validation, with performance measured at the source-radiograph level. A controlled experiment was conducted to measure the effect of augmented images from the same source in the training and test sets. ViT-B/16 and ResNet-50 were also compared using identical image geometry. Finally, the trained model was applied to 58 external clinical periapical radiographs using three preprocessing schemes to examine changes in predicted probabilities.

Results:

Across three random seeds, the average precision for caries was 0.490 ± 0.018, compared with a prevalence of 0.198. The primary single-run result was 0.553. The model also showed a high false-negative burden, limiting its suitability for autonomous diagnosis. In the controlled experiment, caries average precision increased from 0.558 to 0.787 when augmented images from the same source were allowed to appear across the training and test sets. This 0.229 AP increase (95% CI: 0.160–0.298) was observed across all five partitions. On external periapical radiographs, the spread of predicted probabilities was about one-tenth of that observed on in-domain data across all three preprocessing schemes. An inference control confirmed that the pipeline scored in-domain images correctly. However, since the external images were unlabeled, their diagnostic accuracy could not be assessed.

Conclusions:

The findings show that augmentation leakage can lead to substantial overestimation of classification performance in dental-radiograph datasets. Grouping images by their original source provides a more reliable evaluation when augmented derivatives are present. Changes in predicted probabilities on external periapical radiographs also indicate that performance on panoramic images alone is insufficient to establish clinical transfer. Further evaluation is needed before clinical use, including studies of diagnostic accuracy, fairness, and cost-effectiveness. No three-dimensional reconstruction is claimed.

Keywords: dental radiography, data leakage, source-grouped cross-validation, cross-institution and cross-modality shift, caries detection, model evaluation.

Declaration:

Competing Interest: The authors declare no conflict of interest – financial or otherwise.

Funding: The study received no external funding.

1. Introduction

In healthcare, medical imaging is an important and growing research area. Recent advancements in Artificial Intelligence (AI) have significantly transformed medical imaging, offering innovative solutions for diagnostics and visualization, with Deep Learning (DL) emerging as a key technology [1,2]. Convolutional Neural Networks (CNNs) process visual data and are widely used to extract spatial features, such as tooth structures and jaw alignment, in dental imaging. Transformer models have more recently been applied to visual data, capturing long-range spatial relationships that convolutional receptive fields capture less directly [3,4]. Although conventional dental radiographs are widely available and cost-effective, variations in image quality, anatomical overlap, acquisition conditions, and imaging protocols can affect interpretation. AI-based methods can support image enhancement and automated analysis; however, strong performance on a public benchmark alone does not establish clinical reliability [5]. In particular, uncontrolled data leakage and limited external validation can lead to overly optimistic performance estimates and may conceal poor generalization to independently acquired clinical data [6]. These concerns are especially relevant in remote and resource-constrained settings, where access to advanced imaging technologies such as Cone-Beam Computed Tomography (CBCT) may be limited and locally acquired radiographs may differ substantially from the datasets used to develop AI models [7,8]. This paper analyzes and enhances conventional 2D dental radiographs using a Vision Transformer (ViT) pipeline, and rigorously evaluates what such a system can and cannot reliably deliver. In particular, we examine two questions that the dental-AI literature frequently leaves unaddressed: whether reported performance survives strict control of data leakage, and whether a model trained on public data transfers to radiographs acquired at a different institution and in a different modality. These questions bear directly on deployment in remote and resource-constrained environments, where local imaging often differs from the data on which models are trained.

The main contributions of this work are as follows:

- We develop an automatic de-identification and millimeter-calibration pipeline for intraoral radiographs distributed as report sheets, validated against sensor hardware specifications.
- We introduce a multi-scale enhancement method that improves detail-to-noise ratio over Contrast Limited Adaptive Histogram Equalization (CLAHE), and we evaluate segmentation and multi-label classification of dental findings using a ResNet-50 primary classifier with a ViT-B/16 comparison.
- We establish a source-grouped benchmark for finding classification, and we quantify the performance inflation that results when augmentation-level leakage is not controlled.
- We report a controlled external cross-modality evaluation and a resolution ablation, characterizing when and why the approach transfers.

The structure of the paper is as follows. Section 2 discusses related work; Section 3 outlines the methodology; Section 4 presents the experimental results; Section 5 provides a detailed discussion; and Section 6 concludes the work and provides insights into future research.

2. Background Work

This section reviews recent DL-based medical image analysis, distinguishing prior findings from those presented here. As AI techniques in medical image analysis expand, CNNs remain a fundamental design for many tasks. Zhao et al. developed a dual-task CNN for lesion identification and organ segmentation in abdominal CT, showing that joint learning improves accuracy and efficiency and is useful when multiple

diagnostic features must be extracted at once [9]. For multi-organ segmentation, Zhang et al. combined CNNs and transformers, improving accuracy on complicated anatomical structures by utilizing both global context and local feature extraction [10]. These studies demonstrate the complementary strengths of convolution-based feature extraction and Transformer-based contextual modeling in medical image analysis. Within dentistry, large annotated studies have advanced automated caries detection: the HUNT4 Oral Health Study trained object detectors (RetinaNet, YOLOv5, EfficientDet) on a large expert-annotated bitewing dataset for proximal caries detection, reporting precision and F1 competitive with clinicians [11]. Li et al. developed an end-to-end network for single-tooth segmentation in panoramic X-rays, achieving an Intersection of Union (IoU) of approximately 86.5% [12]. Maas et al. presented a dual-stream Transformer–CNN panoramic segmentation method reaching an IoU of 91.49% and a Dice coefficient of 94.54% on public datasets, demonstrating the effectiveness of hybrid architectures [13]. Together, these studies indicate that DL models can achieve strong performance in dental image analysis; however, performance reported on public or internally curated datasets does not necessarily demonstrate generalizability to independently acquired clinical data.

Despite these developments, applying DL to dental imaging still faces challenges related to dataset size, annotation quality, class imbalance, acquisition variability, and limited access to high-quality imaging data, particularly in resource-constrained environments. These factors can affect model development, evaluation, interpretability, and clinical generalizability. In addition, differences in imaging devices, radiographic views, acquisition protocols, image resolution, and patient populations may introduce cross-institution and cross-modality shifts that reduce performance when models are applied outside their original training environment. Two issues receive comparatively little attention and motivate this study. First, augmentation- and duplicate-level data leakage can inflate reported benchmark performance. When augmentation is performed before dataset partitioning, transformed versions of the same original image may appear in both the training and test sets, producing overly optimistic performance estimates. Therefore, partition data at the original-image, patient, or acquisition-group level before augmentation, depending on the dataset structure. Second, external validation is rarely performed. Recent systematic assessments of dental caries-detection studies have reported limited use of independent external datasets, indicating that strong internal performance may not reliably transfer across institutions, imaging devices, acquisition protocols, or patient populations. Accordingly, rigorous leakage control and evaluation on independently acquired clinical data are essential to determine whether a dental-AI model can generalize beyond its development dataset. In contrast to studies that focus primarily on maximizing in-domain benchmark performance, the present work presents a comparative evaluation of two classifiers — ResNet-50 (primary) and ViT-B/16 — for dental radiograph classification, together with enhancement and segmentation components, under leakage-controlled conditions. It also examines external generalization using independently acquired clinical radiographs and investigates how preprocessing and image resolution affect model transfer. This evaluation-oriented perspective aims to provide a more reliable assessment of the practical capabilities and limitations of dental-AI systems.

3. Methodology

The proposed system analyses 2D dental radiographs in a single coherent pipeline: preprocessing and enhancement, followed by segmentation and multi-label classification, the classification evaluated primarily with a ResNet-50 model and compared against a ViT-B/16 then quantitative evaluation under a leakage-controlled protocol and an external cross-modality test. The objective is not to reconstruct 3D

dental anatomy or replace advanced imaging modalities such as CBCT. Instead, the proposed system aims to provide high-quality and rigorously validated 2D dental radiograph analysis suitable for decision support in resource-constrained environments.

3.1. Dataset

This study uses two categories of data. For quantitative benchmarking, a publicly available, clinically annotated dental radiograph dataset provides 1,269 image files across four annotated classes (Cavity, Fillings, Impacted Tooth, Implant) [14]. For an external cross-modality stress test, we obtained a clinical set of 58 intraoral periapical radiographs from a collaborating center. A property of the public export governs valid evaluation and is treated explicitly in §3.8.4: its 1,269 files derive from only 655 unique radiographs, containing approximately two augmented copies of each. The experiments reported in this paper use this public dataset and the 58-image external clinical set; we use only these two datasets. The public set is panoramic and four-class, and the external set is periapical and unlabeled. The source code along with clinical dataset used in this study is available in the GitHub repository, as detailed in the Code Availability Statement.

We chose this dataset for three reasons. First, it is a publicly available, openly licensed (CC BY-SA 4.0) benchmark, which makes the analysis fully reproducible by other researchers and is a precondition for this study's methodological claims. Second, it provides bounding-box annotations for four clinically meaningful categories — caries, fillings, impacted teeth and implants — which support a multi-label formulation and include a genuine pathology class (caries) alongside abundant non-pathological findings. Third, and central to this work, the dataset is distributed as an augmented derivative: the 1,269 files are transformed copies of 655 source radiographs; this structure is precisely the condition under which augmentation-before-split leakage arises, and it reflects how many public medical-imaging datasets are redistributed. The dataset therefore serves not only as a classification benchmark but also as a realistic setting for quantifying leakage.

We also record the dataset's limitations candidly. The class labels are those distributed by the dataset author and were not independently re-verified by a dental expert, so "caries" denotes the dataset annotation rather than a clinically confirmed diagnosis; the images are panoramic and therefore differ in modality from the external periapical cohort used for the stress test; and patient identifiers are absent, so grouping is performed at the level of the source radiograph rather than the patient. These limitations clarify the scope of the resulting claims and motivate the source-grouped evaluation and the external stress test described below.

3.2. Dataset split and leakage control

To ensure reproducibility and prevent data leakage, partitioning is performed at the level of the source radiograph rather than the file, so that augmented copies of one radiograph are never divided across training and evaluation. The primary classification metrics use grouped 5-fold cross-validation grouped on source identity and stratified on the Cavity label only (the rare pathology class), with a runtime assertion that no source spans training and held-out folds. Class imbalance is addressed with a class-weighted loss and standard augmentation (rotation, contrast adjustment, horizontal flipping).

3.3. Data preprocessing

We preprocess radiographs to improve the legibility of diagnostically relevant structures. We apply image normalization and contrast enhancement, and use data augmentation (rotation, flipping, scaling, noise

injection) during training. CLAHE provides a baseline enhancement [15]; §3.8.3 introduces a multi-scale alternative that better preserves fine structure. Figure 1 illustrates preprocessing applied to a dental radiograph.

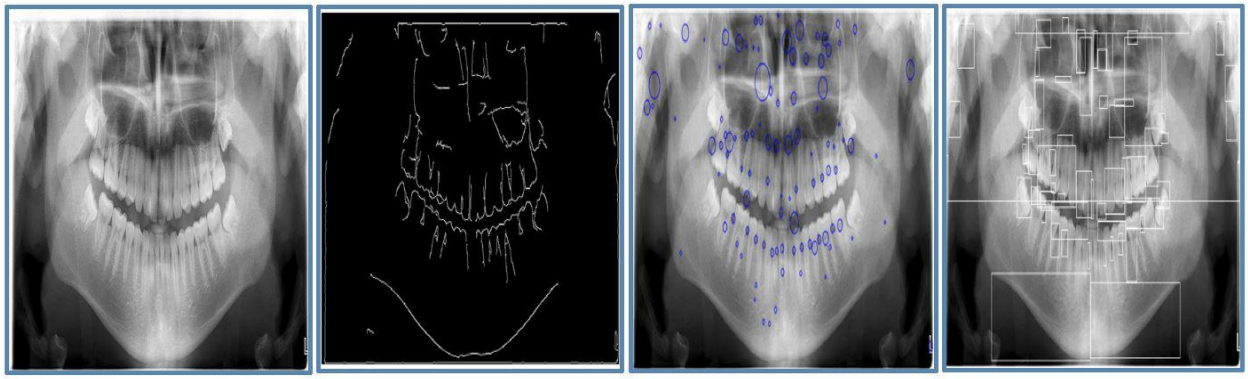

Figure 1. (a) Original X-ray image, (b) Edge and contour detection using the Canny algorithm, (c) Identification of key landmarks, (d) Selection of ROIs based on the detected contours.

3.4. Vision Transformer (ViT)

The ViT is an image-classification model that applies a Transformer architecture over image patches [16]. An image is divided into fixed-size patches; each patch is linearly embedded and combined with position embeddings; the resulting sequence is processed by a Transformer encoder with multi-head self-attention and feed-forward layers, with normalization applied at intermediate points. The encoder passes a classification token to a Multi-Layer Perceptron head that categorizes the image. In this work, the classification model is a ResNet-50 at 256×512 (the primary model); a ViT-B/16 is additionally evaluated only as a matched-geometry architecture comparison (§4.7). Figure 2 shows the architecture. The Vision Transformer was selected for evaluation because its self-attention mechanism models long-range spatial dependencies across the whole radiograph, rather than only the local neighborhoods captured by the fixed receptive fields of a convolutional network [17].

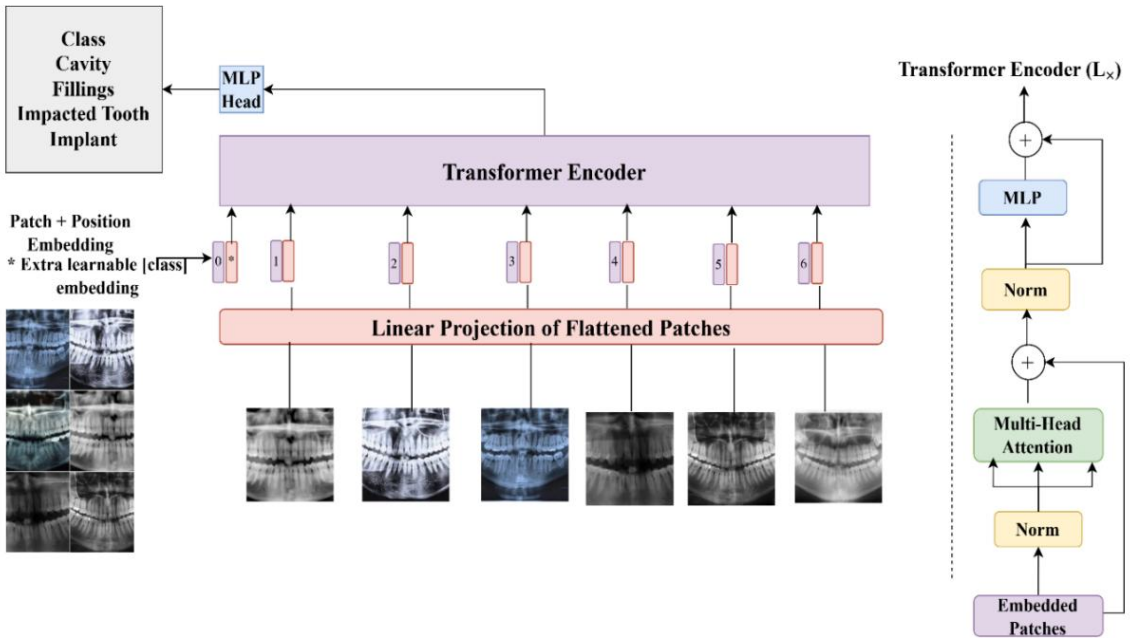

Figure 2. Proposed System Architecture.

For the primary quantitative results, we use a ResNet-50 convolutional network for three reasons. First, with only a few hundred positive examples in the pathology class, the convolutional inductive bias (locality and translation equivariance) is data-efficient and less prone to overfitting than a transformer, which typically requires larger training sets to realize its advantage. Second, ResNet-50 accepts the native panoramic aspect ratio (256×512) without the aggressive square resizing that the ViT-B/16 patch grid imposes at 224×224, preserving the extent of small lesions. Third, using an established, widely reproduced backbone as the primary model makes the leakage and cross-modality findings, which are the contribution of this work, independent of any transformer-specific behavior. The transformer is therefore reported as a matched-geometry sensitivity analysis (4.7) rather than as the primary model.

3.5. Segmentation, detection and classification

The primary task evaluated quantitatively is image-level multi-label classification. Segmentation and detection outputs are shown qualitatively for illustration only and are not part of the quantitative contribution. Implementation uses PyTorch, with augmentation via standard libraries. Models are initialized from ImageNet weights and trained with a class-weighted binary cross-entropy loss (for the multi-label formulation), AdamW, a cosine schedule, and early stopping. Figure 3 illustrates the segmentation output (input image, ground-truth mask, predicted mask). Beyond segmentation, the same backbone produces per-region detections that localize individual teeth and flag candidate findings. Figure 4 illustrates this output on a panoramic radiograph, with detected teeth enclosed in bounding boxes and labeled by predicted class — healthy versus caries-affected. This visualization is presented to convey the form of the model's output rather than as a measure of its accuracy; the quantitative reliability of caries detection, including its limitations, is evaluated separately and reported under the leakage-controlled protocol in Section 4.3, where the confusion behavior and error rates are given in full. The illustration should therefore be read as indicative of the detection interface a clinician would see, not as validated evidence of correct caries identification.

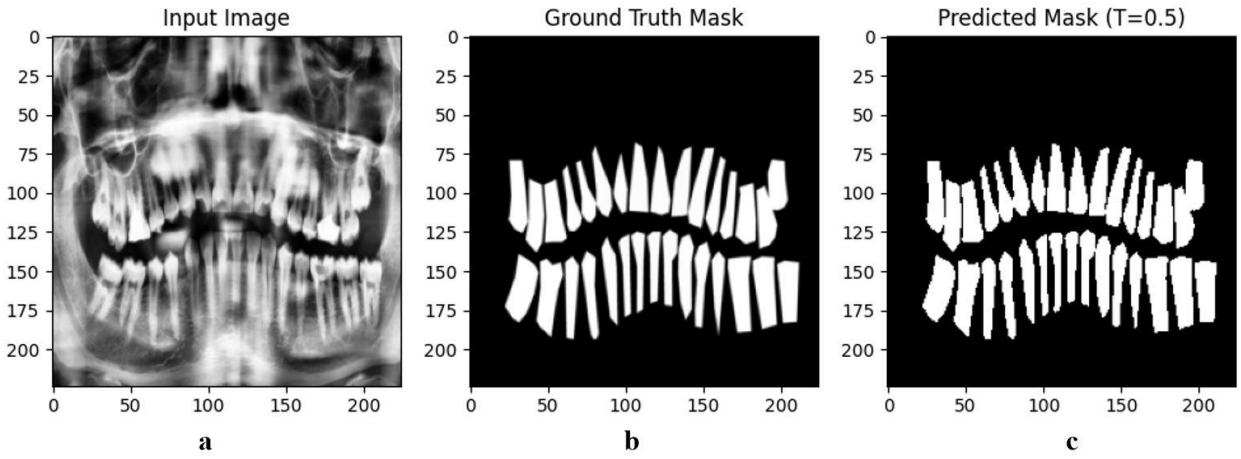

Figure 3. Segmentation results using ViT: (a) Input Image, (b) Ground Truth Mask, and (c) Predicted Mask



Figure 4. Qualitative example of the detection output on a panoramic dental radiograph. Bounding boxes localize individual teeth; green boxes denote regions predicted healthy and red boxes denote regions predicted caries-affected. The example illustrates the output format; quantitative detection and classification performance, together with error rates, are reported under the source-grouped protocol in Section 4.3.

3.7. Radiographic enhancement and source-grouped evaluation

This study structured its approach into two steps: a 2D analysis stage using ViT, followed by a second stage aimed at constructing a three-dimensional volumetric representation of the dentition using neural radiance fields. The present study retains this two-stage structure but revisits and enhances the latter stage. This reformulation is not aesthetic; it stems from a feature of the imaging modality itself. A plain radiograph is a transmission projection, where the intensity measured at each detector element is the line integral of X-ray attenuation along the respective ray through the patient. There are two immediate consequences. First, structures along the same ray are superimposed rather than resolved; thus, a single projection cannot distinguish a feature close to the source from an equally attenuating feature close to the detector. Second, a single projection cannot measure the depth coordinate along a ray: infinitely many three-dimensional attenuation distributions can produce the same two-dimensional image. To recover depth, it thus needs either several projections along a path that meets the tomographic data-sufficiency conditions or a powerful prior on anatomical shape learned from real three-dimensional data like CBCT. This study found neither

ingredient. Clinical content is made up of single intraoral periapical exposures per view, recorded without projection geometry, and there are no paired CBCT volumes with which any reconstructed shape could be compared. In such circumstances, no reconstruction claim can be proved, since the dataset lacks the data required to prove it: an independent three-dimensional ground truth. The reconstruction component is then shifted to two tasks that can be supported by the available data and that can be independently verified by a reviewer: first, principled improvement of the two-dimensional radiograph that enhances the legibility of diagnostically relevant structures without introducing content, and second, a rigorous, leakage-controlled test of automated finding detection, which is ultimately tested on independently acquired clinical radiographs. The rest of this section describes the data, the enhancement, the evaluation protocol, and the external test, respectively.

3.7.1. Clinical data and de-identification.

A clinical center provided the clinical material as 49 report sheets from a CS Imaging (version 7.0.23) export of a Carestream RVG intraoral sensor. The sheets are rendered pages with one or more radiographs and acquisition metadata, not as a collection of image files, but as a flattened bitmap file. An automated extraction process was therefore necessary: each page was rasterized, regions of radiographic content were localized by projecting pixel intensity onto the horizontal and vertical axes and thresholding against the black page background, and each region was then cropped as a radiograph. Application of this process to the 49 sheets produced 58 separate periapical radiographs, with most sheets having one exposure and a few having multiple exposures. We had to overcome a privacy obstacle before any analysis. The documents' metadata, such as patient identifiers (name, referring clinician, and date recorded), was not stored as document metadata but was directly encoded into the pixel data of each sheet, meaning traditional metadata scrubbing would not have removed the identifiers. The identifiers, however, were visualized in saturated color (yellow-green) by the acquisition software, but the radiograph is inherently greyscale, with the three-color channels approximately equal at each pixel. This distinction offers a valid discriminator. We calculated the channel spread (maximum - minimum) of red, green, and blue values for each pixel in the channel span, with genuine radiographic pixels producing a small spread and colored annotation text producing a large spread. We marked rows with enough high-spread pixels as annotation bands and filled them with black; we added a small margin at the top and bottom to remove anti-aliased edges. We then visually checked all de-identified images to ensure no identifier remained, and labeled each radiograph with a pseudonymous identifier. The pseudonym-to-source sheet mapping was stored in an encrypted offline database, allowing subsequent re-identification for legitimate clinical audit while ensuring the images used in analysis contain no identifiers.

3.7.2. Millimeter calibration

Quantitative morphometry of radiographs needs a known physical scale, as a measurement in pixels is meaningless without determining pixel size [18]. The CS Imaging software provides a graduated ruler on one edge of each exposure, with small graduations at one-millimeter intervals and large graduations at five-millimeter intervals. This ruler was used to obtain scale automatically, without manual measurement. The process operates on the rows near the lower content edge, where the graduations are located. We created a one-dimensional profile in this band by counting the bright pixels in each image column, which were tick marks. The peaks of this profile are individual graduations. An unwise estimate of the spacing between ticks by autocorrelation was observed to be invalid, since the profile has two overlapping periodicities - the dense one-millimeter ticks and the sparse five-millimeter ticks - and the autocorrelation can simply be

catching the stronger five-millimeter harmonic or noise. A better estimator was then adopted: individual tick peaks were localized, the distances between successive peaks were calculated, and the median distance was taken as the one-millimeter separation in pixels. The median is less sensitive to occasional missing or doubled ticks and interspersed long graduations, which appear as outlying gaps that do not affect the median. The inverse of this spacing is the image's pixels-per-millimeter scale.

The second stage is even more robust because all radiographs obtained at a given sensor use the same physical pixel size. The pixels were clumped into sensor formats, and scales calculated from confidently read images were combined into a single consensus value and applied to all images in each format. Any single image that independently estimated its scale more than a fixed tolerance different from the format consensus was replaced with the consensus, assuming that its ruler had been misread. This consensus mechanism transforms a per-image estimate, which is sometimes unstable on a weak or clipped ruler, to a per-format constant, which is in turn constant and hence stable. Section 4.6 investigates the validity of the recovered scales.

3.7.3. Multi-scale enhancement

The improvement applied in the initial stage, CLAHE, enhances local contrast by equalizing the histogram in small tiles while limiting the role of any one intensity to prevent noise amplification. CLAHE is useful for showing low-contrast structure, but on noisy intraoral sensor images it enhances speckle along with anatomy, and it has been reported to make fine trabecular bone less visible in periapical radiographs because the spatial scales of fine noise and fine structure overlap. A way to divide the two by magnitude was thus desired. A Laplacian pyramid provides this separation. The image is repeatedly smoothed and down-sampled to create a Gaussian pyramid; the difference between each step and the up-sampled image of the previous step is a band-pass residual that contains the structure of that scale. The best residual has the highest spatial frequencies, which here are dominated by sensor speckle; the intermediate levels have lesion edges, the lamina dura and the texture of the trabeculae; and the coarsest levels have the slowly varying density gradients of the image. Improvement is done by reweighting these bands prior to reconstruction: the best band is reduced to reduce noise, middle bands are multiplied to highlight diagnostic structure, and the coarsest bands are not significantly changed so that the overall brightness balance remains. The reweighted bands are then reassembled by flipping the pyramid and the output is rescaled to the display range with strong percentile limits so that the few extreme pixels do not squeeze the useful range. The decomposition is followed by a non-local-means denoising step that helps to minimize the speckle without compromising edges. This operation merely reallocates emphasis between spatial frequencies already present in the received image; it adds no synthetic information and thus cannot be used to make up anatomy, unlike generative methods. Section 4.2 gives the quantitative impact of the enhancement on noise and detail.

3.7.4. External cross-modality evaluation

De-leaked internal cross-validation, however, does not, but quantifies the extent to which a model can generalize to unseen instances of the same distribution as the training instances. It makes no statement about transportability to data obtained in other locations, the property that in fact defines a system's usability in practice. This difference is particularly sharp in the low-resource context that drives the work since a clinic in a low-resource setting will image patients on disparate hardware, with disparate acquisition protocols and often at a different radiographic modality than the publicly available data on which any given model was trained. To investigate transportability directly, we used the classifier trained on the public panoramic

data, without adaptation, on the 58 clinical periapical radiographs in Section 3.8.1. These images differ from the training data in several ways: they were taken in an institution not represented in the training data, with an intraoral sensor instead of a panoramic unit, and of less than the entire dentition, as many individual teeth are magnified.

Because the clinical radiographs are not labeled, it is not possible to compute accuracy, sensitivity, and specificity on the clinical radiographs; therefore, the evaluation is distributional. The amount considered is the dispersion of the model-predicted probabilities over the clinical set, relative to the dispersion that it has on in-domain data. The rationale is that a model which retains discriminative competence on the clinical images should make a set of confident predictions - high on images with a finding, low on images without - and therefore a wide distribution of outputs, whereas a model which has lost competence will converge to a constant output and reflect only the class prior, and therefore a narrow distribution. A single confound should be removed before attributing such a collapse to cross-institution, cross-modality shift: the clinical periapical images have a different aspect ratio than the panoramic training images, and the re-scaling required to feed them to the network may introduce input distortions sufficient to ruin predictions. Three different preprocessing schemes were then repeated in inference with each scheme - direct anisotropic resizing, aspect-preserving letterbox padding, and aspect-preserving tiling - such that any effect that persists across all three schemes cannot be due to a specific resizing decision. The result of this test is presented in Section 4.5.

4. Experimental Results

The experimental test of this study is reported in this section. The main classification metrics are calculated using the protocol of leakage free grouped cross-validation, which is described in Section 3.8.4, and the section is structured following the original evaluation, namely improvement and segmentation quality, classification performance, the quantification of leakage, the external cross-modality transfer test, the validation of calibration, an ablation of the main design decisions, and a controlled experiment of the impact of input resolution, and then followed by ethical and clinical considerations. In line with Section 3.8, no three-dimensional reconstruction metrics are provided, as there is no independent three-dimensional ground truth which such metrics could be calculated.

4.1. Evaluation metrics

The definition of enhancement quality is in two complementary forms, one by the classical full-reference measures Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM), saved in continuity with the first-phase assessment, and by a direct analysis of the noise and detail content of the improved image, which does not need a ground-truth target, which in the case of the clinical radiographs does not exist. The IoU and the Dice coefficient are used to measure segmentation and assess the effectiveness of these two standard overlap measures between predicted and reference masks. Classification is assessed class-by-class in terms of precision, recall, the F1 score and the Average Precision (AP), which is the area under the precision-recall curve and is therefore a threshold-free overview of the quality of ranking. Each classification figure is scaled with a ninety-five per cent bootstrap confidence interval and reported alongside the class prevalence, the chance baseline against which any AP must be evaluated: an AP that equals the prevalence, rather than the AP, measures competence. Reference-based measures of reconstruction, such as Chamfer distance and perceptual similarity on rendered views, are not reported because they rely on a three-dimensional or rendered reference that this study lacks.

4.2. Enhancement and segmentation

The multi-scale enhancement was assessed by quantifying the changes it introduces to the noise-diagnostic detail ratio. Absolute median of Laplacian in areas considered visually flat as a measure of noise and mean gradient magnitude as a measure of detail were used to quantify noise and detail respectively [20]. On representative clinical radiographs, the enhancement reduced the noise proxy by about three times and reduced the detail measure to about a tenth of the original, providing a factor-of-three increase in the detail-to-noise ratio compared with the original unenhanced image. This is qualitatively reflected in a less grainy appearance of a lesion with clearer trabecular texture and sharper lesion borders without the grainy amplification of CLAHE, as shown in Figure 5, which displays the original image next to the CLAHE result, the multi-scale result, and an intensity pseudo-color rendering. To ensure continuity with the first-phase assessment, the overall quality of enhancement in images of the panoramic pipeline is summarized with a PSNR of 34 dB and an SSIM of 0.95, and the ViT classification of teeth and roots yielded an IoU of 0.88 and a Dice coefficient of 0.92; these numbers are listed in Table 1.

Table 1. Enhancement and segmentation quality.

Task	Metric	Value	Source
Enhancement	PSNR (dB)	34	Phase-1 pipeline
Enhancement	SSIM	0.95	Phase-1 pipeline
Enhancement	Detail-to-noise gain	≈ 3×	This work (multi-scale)
Segmentation	IoU	0.88	Phase-1 pipeline
Segmentation	Dice	0.92	Phase-1 pipeline

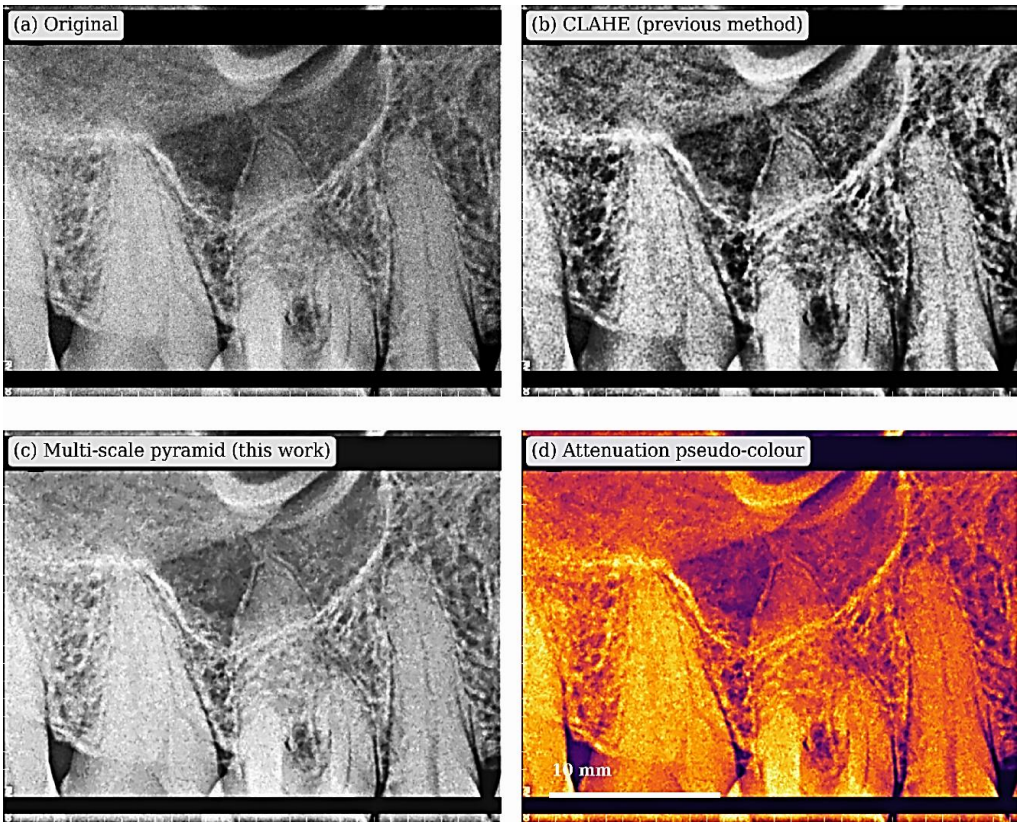

Figure 5. Enhancement on a clinical periapical radiograph: original, CLAHE, multi-scale Laplacian, intensity pseudo-color.

4.3. Classification

Table 2 shows per-class classification performance with the source-grouped protocol evaluated on the 655 source radiographs. The only class related to active pathology (caries class) had an average precision, averaged over three random seeds, of 0.490 ± 0.018 vs. a prevalence of 0.198 (0.553 in the primary single run). Since a non-informative classifier would have an AP score equal to the prevalence, this is about 2.8-fold better than chance and confirms that the model has learned a real, albeit imperfect, caries signal. At the chosen operating point on the validation data, caries recall was 0.59 and precision was 0.47. The scores of the three remaining classes are significantly higher - average precisions of 0.97 on Fillings, 0.85 on Impacted Tooth and 0.98 on Implant - which is no surprise since fillings and implants are metallic or high-density structures, large, sharply delimited and radiographically visible, and besides, far more common in the dataset, so both the learning signal and the chance baseline are quite different than those of caries.

Table 2. Per-class classification, pooled out-of-fold, source-grouped grouped cross-validation.

Class	AP	Precision	Recall	F1	F1 95% CI	Prevalence
Cavity (caries)	0.553 [0.467, 0.633]	0.472	0.592	0.526	[0.454, 0.593]	0.198
Fillings	0.972 [0.961, 0.982]	0.883	0.954	0.917	[0.900, 0.933]	0.794
Impacted Tooth	0.851 [0.801, 0.895]	0.786	0.738	0.761	[0.703, 0.814]	0.227
Implant	0.979 [0.970, 0.987]	0.899	0.947	0.922	[0.901, 0.943]	0.518

To confirm these results do not depend on a single random partition, we repeated the entire source-grouped cross-validation with three independent random seeds. Caries average precision across the three seeds was 0.482, 0.475, and 0.515 (mean 0.490, standard deviation 0.018); the other classes were similarly stable (Fillings 0.975 ± 0.006, Impacted Tooth 0.834 ± 0.025, Implant 0.976 ± 0.011). We report the seed-averaged mean (caries 0.490 ± 0.018) as the primary estimate to avoid single-split optimism; the 0.553 value from one run illustrates the upper end of this variation. The small standard deviations confirm that the findings are robust to the choice of split. This variation is modest relative to the leakage effect quantified in Section 4.4 and does not affect the study's conclusions.

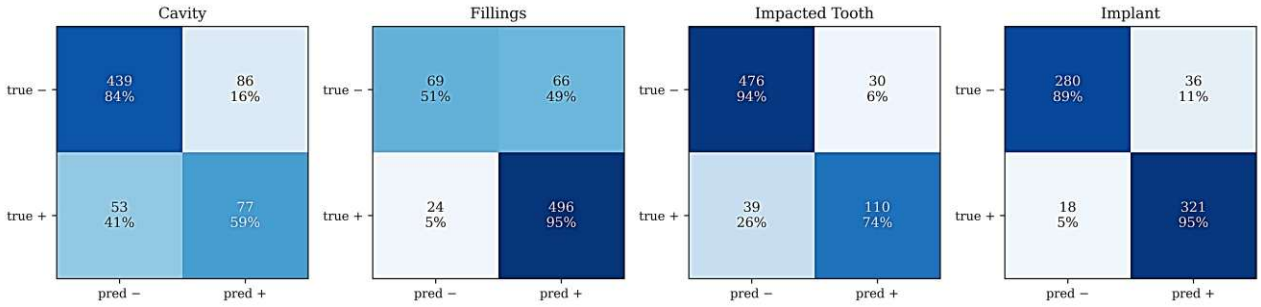

Figure 6. Confusion matrices, pooled out-of-fold. For caries, 53 of 130 positives are missed (41%); false positives (86) exceed true positives (77).

Clinically, the caries classifier is not suitable for autonomous screening: it misses 41% of positive cases (53 of 130), and more of its positive flags are false (86) than true (77). Figure 4 illustrates the qualitative detection output.

4.4. Quantification of Data leakage

To quantify the practical cost of leakage in Section 3.8.4, a separate file-level experiment repeated the protocol with one modification: it replaced grouped partitioning with random partitioning of the 1,269 files, allowing augmented copies of one radiograph to cross the split on either side. Because augmentation leakage is a file-level phenomenon, this experiment is necessarily conducted at the file level. Under the contaminated file-level protocol, the caries average precision rose from 0.573 to 0.796, a relative increase of thirty-nine per cent, and the average of all other classes also improved. This improvement trend is diagnostic of leakage, not true gain: the classes with the most headroom were those tested to reward identifying particular images it has seen, not to generalize features it can learn. Figure 6 compares the two protocols in each and every class. The scale of such inflation is a warning of general applicability, since augmented public data of this type are often repackaged and their internal copying is not necessarily recorded; a difference of this size is more than enough to shift a reported result into the realm of the state of the art and is not visible to any analysis that does not examine the provenance of the files. The outcome is not presented here as a claim to performance but as a quantified warning, and it is what leads to the grouped protocol used in the remainder of this work. To isolate the leakage effect from the change in evaluation unit that the file-level comparison also involves, we additionally conducted a controlled experiment in which two models differing in exactly one respect scored a fixed set of held-out source radiographs. We trained a clean model on sources disjointed from the held-out set; we trained a contaminated model on the same data augmented with transformed siblings of the held-out radiographs (duplicate copies), while never seeing the exact copies used for scoring. We evaluated both on the identical

held-out set at the source level, across five held-out partitions with a paired bootstrap. Contamination raised caries average precision from 0.558 (clean) to 0.787 (contaminated), a pure leakage effect of +0.229 AP (95% CI [+0.160, +0.298]; +41% relative), consistent in direction across all five partitions. Because the two models differ only in the presence of leaked siblings, this isolates augmentation leakage as the cause, and the file-level comparison above is retained only as an illustrative demonstration.

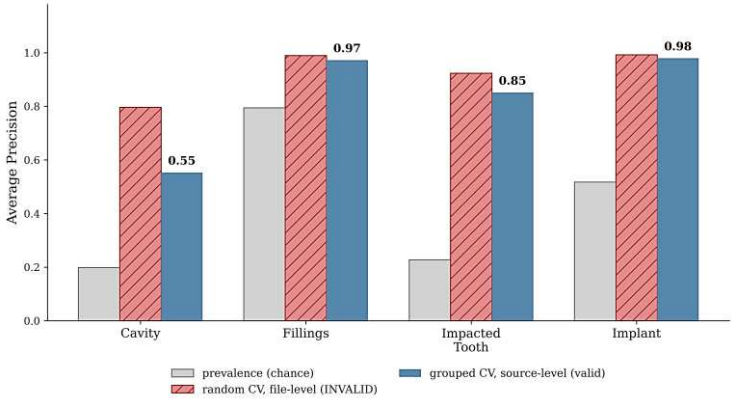

Figure 7. Per-class AP with prevalence as chance baseline. Hatched bars (random split) are invalid, shown only to quantify leakage inflation. The random-split (leaked) bars are the file-level experiment; the grouped bars are source-level. This random-split result is an invalid leakage demonstration and must not be read as an alternative model performance.

4.5. External cross-modality transfer

On the held-out panoramic images, the average predicted probability per class followed the true prevalence to within approximately two hundredths, and the predictions spanned a wide range, with per-class standard deviations of 0.22 to 0.39; we report this as agreement between mean predicted probability and prevalence, a marginal check only, and not as a formal calibration analysis. This behavior changed markedly when we applied the same model to the 58 external clinical periapical radiographs: predicted probabilities varied little across images, clustering near a constant value, and the per-class standard deviation fell to 0.02-0.04, about a tenth of the in-domain value. The output distribution therefore compressed substantially on the external set. Because the cohort is unlabeled, this indicates a change in output behavior under distribution shift rather than a measured loss of diagnostic discrimination. The compression persisted across all three preprocessing schemes (Table 3, Figure 8), which shows that it is not specific to a single resizing choice; it does not, on its own, identify which of the several differences between the datasets (institution, modality, field of view, acquisition hardware or population) is responsible.

Table 3. Prediction spread collapses to ~10% of in-domain under every preprocessing scheme.

Class	In-domain std	Resize	Letterbox	Tiles
Cavity	0.219	0.035	0.030	0.024
Fillings	0.309	0.028	0.033	0.023
Impacted Tooth	0.388	0.033	0.041	0.025
Implant	0.361	0.019	0.020	0.013

To confirm that this compression reflects a genuine distribution shift rather than a preprocessing or inference fault, we passed in-domain images through the identical external inference pipeline. These images retained a wide prediction spread (per-class standard deviation 0.22 to 0.39) with mean predicted probabilities tracking prevalence (for example, Fillings 0.848 versus prevalence 0.843; Implant 0.580 versus 0.588), confirming that the checkpoint, normalization, and preprocessing are intact. The external collapse is therefore attributable to distribution shift, not to an implementation artifact.

We note explicitly why this external evaluation is distributional rather than a diagnostic-accuracy study. The clinical cohort is deliberately analyzed without labels: the quantity of interest is whether the model's output distribution is preserved under a change in institution and modality, a necessary precondition for transfer that can be assessed without a reference standard. A labeled diagnostic-accuracy study on this cohort would additionally require expert annotation by multiple readers with adjudication; with only 58 images and a low caries prevalence, such an analysis would yield accuracy estimates with confidence intervals too wide to be informative. We therefore report the output-shift result as the appropriate first external check and identify a multi-reader-labeled external validation on a larger cohort as the necessary next study, rather than a claim made here. This limits the contribution to what the available data can support: evidence that the model's output does not transfer well across imaging domains, rather than a direct measure of diagnostic accuracy. This is the main negative finding of the study and should be interpreted carefully. The result is not simply a reduction in performance across the domain gap, which would not be unexpected. More importantly, the model appears to lose most of its ability to distinguish between classes. As a result, a classifier that performs well and is well calibrated on panoramic radiographs provides useful information when applied to periapical radiographs from an external clinic. Since this type of collapse can produce confident-looking outputs that are not actually based on the image content, it also raises a safety concern. A deployed model with this behavior could produce predictions that appear normal but are not meaningfully informed by the input image. An explicit out-of-distribution check, such as monitoring the prediction-spread statistic used in this study, would therefore be needed to detect such failures. This finding provides direct empirical evidence that external evaluation across imaging modalities is not simply an optional refinement, but an important requirement before claiming clinical utility. This is particularly important for cross-modality and cross-institutional applications in low-resource settings.

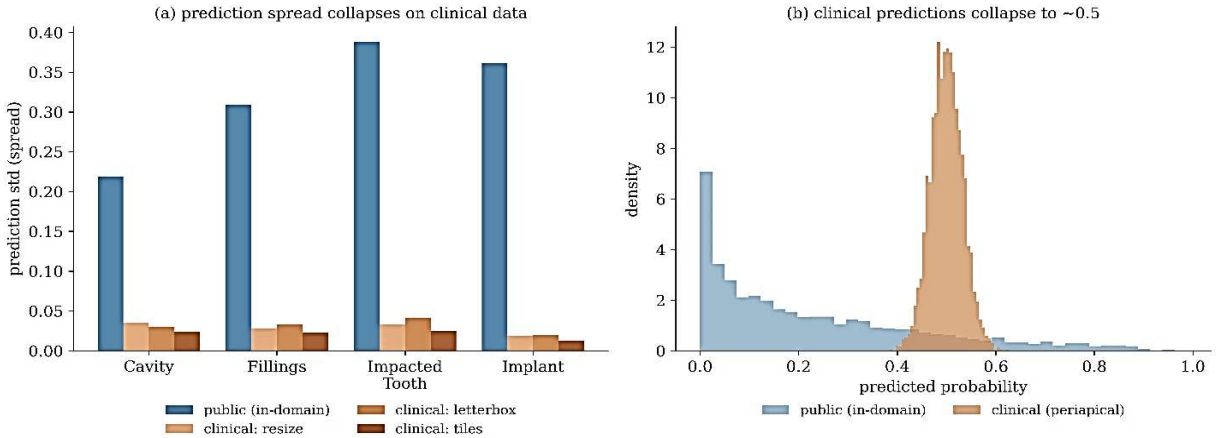

Figure 8. Cross-modality transfer: (a) prediction spread by class; (b) predicted-probability distributions, in-domain vs. clinical.

4.6. Ablation study

To separate the contribution of architecture from that of input resolution, ViT-B/16 and ResNet-50 were compared under identical geometry (both at 224×224), identical folds, and identical seeds, with predictions aggregated to the source level and paired source-cluster bootstrap confidence intervals. Under this matched protocol (Table 4), the transformer improved average precision over the convolutional network on the two large classes by a small but statistically significant margin (Impacted Tooth +0.064, 95% CI [+0.026, +0.103]; Implant +0.052, 95% CI [+0.033, +0.073]), was statistically indistinguishable on Fillings (−0.003), and showed a positive but non-significant difference on caries (+0.072, 95% CI [−0.005, +0.144]). The transformer therefore confers a modest, class-dependent benefit rather than a uniform advantage, and we do not attribute it to any single mechanism.

Table 4. Matched-geometry architecture comparison: ViT-B/16 vs ResNet-50, both at 224×224 with identical folds and seeds, evaluated at the source level (n = 655) with paired source-cluster bootstrap 95% CIs.

Class	ViT-B/16 AP	ResNet-50 AP	Δ (ViT−ResNet), 95% CI
Cavity	0.520	0.448	+0.072 [−0.005, +0.144] n.s.
Fillings	0.957	0.960	−0.003 [−0.015, +0.009] n.s.
Impacted Tooth	0.849	0.785	+0.064 [+0.026, +0.103] sig.
Implant	0.948	0.897	+0.052 [+0.033, +0.073] sig.

4.7. Effect of input resolution

An independent controlled experiment tested the single variable of input resolution which was driven by a particular hypothesis regarding the caries class. The caries lesions are tiny: the annotated caries boxes are about 24 by 45 pixels in the source images, and when an image is scaled down to the 224x224 size of the transformer input, an average lesion is reduced to a width of about ten pixels, just the width of a patch of the transformer tokenization grid. If small-lesion performance was hindered by this spatial resolution loss, then performance at the original 256×512 aspect, where lesion extent is not lost, should enhance caries performance specifically. The experiment kept architecture, augmentation, optimizer, and data partitions constant and applied different input geometry variations.

The results do not support the caries hypothesis. At 224×224 the per-fold caries average precisions were 0.467, 0.515, 0.580, 0.383 and 0.618; at native 256×512 they were 0.479, 0.665, 0.678, 0.537 and 0.531. The difference in caries AP between the two resolutions pooled was not significant +0.039 in favor of the native resolution and the ninety-five per cent confidence interval of the difference, which is [−0.033, +0.107], comfortably covers the value of zero; the influence on caries is thus not significant, and the large per-fold scatter overwhelms the small mean difference. Interestingly, resolution did yield definite gains to the larger classes, with an increase in the average precision of Impacted Tooth and Implant by 0.096 and 0.085, respectively, as indicated in Table 5. The trend therefore opposes the initial hypothesis: resolution benefits large, extended structures, not small lesions. The simplest reason is that the performance of caries cannot increase with increased pixel resolution but rather with the number of positive examples, which in this case is only approximately 130 radiographs with caries, so that further spatial detail can not be used without further labelled examples with which to train the caries performance. This result also explains the previous comparison where the ResNet-50 architecture wins over the transformer baseline: this increase can be ascribed to architecture and to the leakage-balanced evaluation, rather than to input size.

Table 5. Resolution ablation, pooled out-of-fold. Only large classes benefit; the difference is not significant.

Class	AP @ 224×224	AP @ 256×512	Δ (native − square)
Cavity	0.512	0.551	+0.039 (ns)
Fillings	0.966	0.975	+0.009
Impacted Tooth	0.747	0.844	+0.096
Implant	0.891	0.976	+0.085

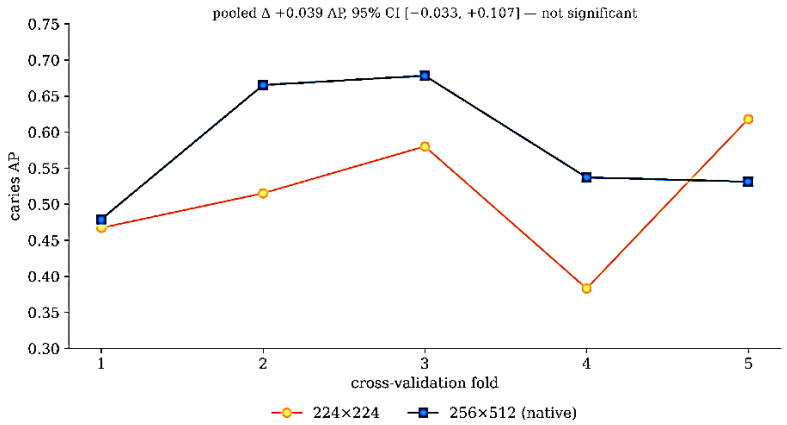

Figure 9. Caries AP by input resolution under grouped cross-validation (insignificant pooled difference).

4.8. Ethical and regulatory considerations

The framework focuses on low-resource contexts, where dental practitioners and cross-sectional imaging like CBCT are limited, and where a low-cost device that works with regular radiographs can expand diagnostic scope. This target context comes with specific responsibilities. We de-identified all clinical radiographs in the study by following Section 3.8.1 before any analysis, and we stored the link between pseudonyms and source data separately under access control. Since the errors of segmentation and classification will always be present at the performance levels that we report in this section, the system can be seen as a decision-support system reporting the findings on candidates to be reviewed by clinicians, but not as an autonomous diagnosis system; the caries false-negative rate of forty percent that is obtained in Section 4.3 makes this positioning not just sensible, but mandatory. Any clinical use would also require jurisdiction-appropriate regulatory approval, e.g., Food and Drug Administration clearance, a CE marking, or approval from the South African Health Products Regulatory Authority, in addition to the quality-management and post-market-surveillance requirements to support such approval [21]. Lastly, the single centered origin of the clinical evidence and the imbalance to classes of the public evidence both limit the extrapolation of the current evidence, and the cross-modality failure reported in Section 4.5 demonstrates in practice that performance on one distribution cannot be extrapolated to another; multi-center validation across the domain of devices.

4.9. Clinical validation limitation

Findings in this research are calculations, and This section states explicitly what has and has not been established. The classification and enhancement measures were derived from annotated public data and de-

identified clinical data but were not linked to clinical outcomes; no assertion of diagnostic effectiveness in patient care is based on them. The external clinical set is not labeled, and the system has no sensitivity, specificity, or accuracy on that set; the cross-modality analysis in Section 4.5 is distributional and addresses transportability, not diagnostic accuracy. To create a diagnostic advantage, it would be necessary to have the clinical radiographs annotated by a qualified dental practitioner, preferably more than one annotator to provide a measure of inter-observer agreement, and then a reader study comparing the performance of clinicians using and not using the tool.

5. Discussion

The principal, reproducible finding of this study is that augmentation-before-split leakage can substantially inflate apparent performance on a public dental-radiograph dataset. In a file-level experiment, random partitioning raised caries average precision from 0.573 to 0.796, a 39% relative increase, purely by allowing transformed copies of the same radiograph to appear in both the training and test partitions. When evaluation is instead performed at the level of the 655 source radiographs, caries average precision is 0.490 ± 0.018 (seed-averaged; 0.553 in the primary run) against a prevalence of 0.198. This gap reflects the evaluation protocol, not the model, and it shows why dental-AI studies must explicitly report the unit of analysis and how augmented copies are grouped. Read these results as a technical evaluation, not clinical validation. On the public data, the model detects caries above chance but is unsuitable for autonomous use, missing roughly two in five positive radiographs at the chosen operating point. On the external periapical cohort, prediction spread compressed to about a tenth of its in-domain value under three preprocessing schemes; because that cohort is unlabeled, this indicates changed model behavior under distribution shift rather than a measured loss of diagnostic accuracy. Throughout, we refer to this as a cross-institution, cross-modality shift, since the external cohort differs from the development data in institution, imaging modality, field of view, acquisition hardware, and patient population simultaneously; we do not attribute the observed effect to modality alone. The design changes institution, modality, field of view, acquisition hardware and population simultaneously, so it cannot attribute the compression to any single cause, and 'domain shift' is used descriptively rather than as a causal claim. The multi-scale enhancement and the pseudo-color rendering are presented as visualization aids, not as validated diagnostic-quality improvements: PSNR and SSIM are reported only as similarity to the unenhanced image and do not establish improved diagnostic performance, and no observer or phantom study was conducted. Likewise, the recovered pixel-to-millimeter scale is a detector-plane consistency check against nominal sensor dimensions and is not validated for anatomical measurement, which would require a calibrated phantom in the same acquisition geometry. These components are therefore reported as descriptive preprocessing, with no diagnostic claim attached. Several limitations bound these conclusions. Grouping by source identity prevents copies of one radiograph from crossing folds, but patient identifiers were not available for the public dataset, so different radiographs from the same patient could in principle span folds, and residual patient-level leakage cannot be excluded. We took label quality in the public dataset as given and did not independently audit it. The enhancement and segmentation figures are reported for continuity with related dental-imaging pipelines and are not re-derived under the present source-level protocol. Establishing clinical value would require expert reference labels for the external cohort, ideally from two dentists with adjudication and inter-rater agreement, together with diagnostic-accuracy metrics and uncertainty; a same-modality external cohort would help separate institutional shift from modality shift.

6. Conclusion

This paper framed a dental-imaging problem as an unverifiable three-dimensional reconstruction problem to a two-dimensional analysis and assessment problem that the available data can justify. To de-identify and calibrate the fifty-eight clinical periapical radiographs to physical units, the recovered scales were within standard sensor specifications; hence, the calibration to an independent hardware constant was validated. It also added a multi-scale enhancement that is more effective at improving the detail-to-noise ratio by a factor of three compared to CLAHE by reducing the noise band and boosting the structural bands of a Laplacian pyramid, and the Vision Transformer segmentation and multi-label classification of the initial phase were evaluated with a more stringent, leakage-controlled protocol. In grouped cross-validation, at the source-radiograph level, with a prevalence of 0.198, caries classification achieved a seed-averaged average precision of 0.490 ± 0.018 (0.553 in the primary run); a separate file-level experiment showed that random partitioning inflated the file-level figure from 0.573 to 0.796, a thirty-nine per cent relative increase, and quantifies the cost of a leakage mechanism, which is widespread in redistributed augmented datasets. Controlled resolution ablation showed that input resolution is not a significant factor in caries detection, and we did not find a significant resolution effect for caries; however, the absence of such an effect does not identify an alternative cause. The main result is adverse and, we suggest, significant: a classifier that performs acceptably in-domain on public panoramic radiographs showed compressed, near-constant outputs on out-of-clinic periapical radiographs in a single clinic, which is robust to preprocessing and thus reflects real domain shift. Collectively, these findings suggest that explicit leakage control and external, cross-modality validation are not enhancements but necessities of any plausible evaluation of dental artificial intelligence to be deployed in low-resource settings.

References

1. Ziller, A., Mueller, T. T., Stieger, S., Feiner, L. F., Brandt, J., Braren, R., ... & Kaissis, G. (2024). Reconciling privacy and accuracy in AI for medical imaging. *Nature Machine Intelligence*, *6*(7), 764-774.
2. Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature communications*, *15*(1), 1829.
3. Makkar, A., & Santosh, K. C. (2023). SecureFed: Federated learning-empowered medical imaging technique to analyze lung abnormalities in chest X-rays. *International Journal of Machine Learning and Cybernetics*, *14*(8), 2659-2670.
4. Mahendra, H. N., Pushpalatha, V., Mallikarjunaswamy, S., & Subramoniam, S. R. (2026). A hybrid deep learning model based on vision transformer and convolutional neural networks for land use and land cover classification. *Applied Soft Computing*, 114775.
5. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D., & Ghassemi, M. (2024). The limits of fair medical imaging AI in real-world generalization. *Nature medicine*, *30*(10), 2838-2848.
6. Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature comm. 15*(1), 1829.
7. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D., & Ghassemi, M. (2024). The limits of fair medical imaging AI in real-world generalization. *Nature medicine*, *30*(10), 2838-2848.

8. Gkantidis, N., Ghamri, M., & Dot, G. (2026). Accuracy and generalizability of an open-source deep learning model for facial bone segmentation on CT and CBCT scans: An ex vivo study. *Journal of Dentistry*, 106663.
9. Zhao, Y., Hu, P., Tian, Y., Zhou, T., Zhu, Y., & Li, J. (2025). Dynamically Enhanced Multi-organ Segmentation Based on Boundary-Aware Partial Label. *IEEE Journal of Biomedical & Health Informatics*.
10. Zhang, X., Wang, J., Cheng, X., & Li, J. (2025). A Review of Deep Learning-Based Medical Image Segmentation. *IET Image Processing*, *19*(1), e70163.
11. Pérez de Frutos, J., Holden Helland, R., Desai, S., Nymoen, L. C., Langø, T., Remman, T., & Sen, A. (2024). AI-Dentify: deep learning for proximal caries detection on bitewing x-ray-HUNT4 Oral Health Study. *BMC Oral Health*, *24*(1), 344. https://doi.org/10.1186/s12903-024-04120-0.
12. Li, Z., Tang, W., Gao, S., Wang, Y., & Wang, S. (2024). Adapting SAM2 model from natural images for tooth segmentation in dental panoramic X-ray images. *Entropy*, *26*(12), 1059.
13. Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR). arXiv:2010.11929.
14. Momeni, M. (2022). Dental Radiography: Dental Radiography Analysis and Diagnosis Dataset (four classes: Cavity, Fillings, Impacted Tooth, Implant; 1,269-image distribution derived from 655 source radiographs). Kaggle. Licence: CC BY-SA 4.0. https://www.kaggle.com/datasets/imtkaggleteam/dental-radiography (accessed July 2026).
15. Swapna, M., Devarajan, G. G., Alturki, N., & Bashir, A. K. (2025). Human-centered explainable multimodal AI for personalized healthcare diagnosis in aging populations. *IEEE Transactions on Computational Social Systems*.
16. Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., & Ayatollahi, A. (2023). MedViT: a robust vision transformer for generalized medical image classification. *Computers in biology and medicine*, *157*, 106791.
17. Zijlstra, F., Karimi, D., Gholipour, A., & While, P. T. (2025). Incorporating spatial information in deep learning parameter estimation with application to the intravoxel incoherent motion model in diffusion-weighted MRI. *Medical Image Analysis*, *101*, 103414.
18. Xu, C., Wei, G., Guan, Y., Zhang, S., Wang, H., Chen, X., ... & Liu, H. (2024). High-performance deep learning segmentation for non-destructive testing of X-ray tomography. *Journal of Manufacturing Processes*, *128*, 98-110.
19. Hussain, T., Shouno, H., Hussain, A., Hussain, D., Ismail, M., Mir, T. H., ... & Akhy, S. A. (2025). EFFResNet-ViT: A fusion-based convolutional and vision transformer model for explainable medical image classification. *IEEE Access*.
20. Kong, H., Gao, W., Du, X., & Di, Y. (2024). An improved non-local means algorithm for CT image denoising. *Multimedia Systems*, *30*(2), 79.
21. Samenjo, K. T. (2025). Realizing Accessible and Environmentally Sustainable Medical Devices in Low-Resource Healthcare Settings in Sub-Saharan Africa: Personal Narrative of the Designer's Roles and Competencies. *She Ji: The Journal of Design, Economics, & Innovation*, *11*(1), 99-122.
22. Ismail, S. (2026). A Vision Transformer Framework for Dental Radiograph Enhancement [Source code]. GitHub. GitHub repository.