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.

DINOv2 versus Supervised Visual Representations for Tooth-Level Multi- Label Diagnosis Classification on Panoramic Radiographs: A Frozen-Feature Sample-Efficiency Study

Panagoula Dimitrelou
National and Kapodistrian University of Athens
Efstathios Alonaris
National and Kapodistrian University of Athens
Athanasia Sergounioti
athanasia.sergounioti@ac.eap.gr

Hellenic Open University

Research Article

Keywords: Panoramic radiography, Deep learning, Self-supervised learning, Multi-label classification, Dental diagnosis, Sample efficiency

Posted Date: August 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10744021/v1

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

Additional Declarations: No competing interests reported.

Abstract

Objectives. To compare frozen DINOv2-small with supervised ImageNet-pretrained visual representations for tooth-level multi-label diagnosis classification on panoramic radiographs, emphasizing performance under constrained annotation budgets.

Methods. We analysed 705 fully annotated DENTEX panoramic radiographs using 224×224 tooth-level regions of interest for four diagnosis labels and a radiograph-grouped 65/15/20 train/validation/test split. Frozen DINOv2-small, ImageNet-pretrained ResNet50, and random ResNet50 representations were evaluated with per-label logistic-regression linear probes. Fifty matched few-shot repetitions used annotation budgets of k = 5–100 positive examples per label with matched negatives. A secondary sensitivity analysis added a supervised ImageNet-pretrained ViT-S/16 comparator.

Results. In the prespecified full-data comparison, DINOv2 and ImageNet ResNet50 achieved similar macro-F1 (0.6243 vs 0.6181; 95% CI for the difference, -0.057 to 0.068; permutation p = 0.88), while random ResNet50 achieved 0.4635. Across few-shot budgets, DINOv2 exceeded ImageNet ResNet50 in all 50 repetitions (mean paired difference + 0.0488 to + 0.0732). Supervised ViT achieved full-data macro-F1 0.6088; DINOv2 also exceeded it on average at every few-shot budget (mean paired difference + 0.0139 to + 0.0331).

Conclusions. DINOv2 showed no statistically robust full-data advantage over ImageNet ResNet50 but provided the strongest frozen representation under constrained annotation budgets. Its few-shot advantage persisted against a supervised Vision Transformer, indicating that the finding was not solely a CNN-versus-transformer effect. Differences in pretraining data, objective, patch size, and training recipe prevent attribution of the advantage specifically to self-supervision.

Introduction

Panoramic radiography is widely used for initial dental assessment, but interpretation is subject to observer variability and has recognized limitations for subtle caries and periapical findings [1, 14]. These limitations have motivated substantial work on computer-assisted analysis of dental radiographs. Deep-learning systems have been evaluated for caries detection and staging [15–18], periapical lesion detection and segmentation [19, 20], impacted-tooth detection [21, 22], and automated tooth detection and numbering [23]. The Dental Enumeration and Diagnosis on Panoramic X-rays (DENTEX) challenge subsequently established a public, expert-annotated benchmark for tooth-level detection and diagnosis of four abnormal findings: caries, deep caries, periapical lesion, and impacted tooth [2].

A recurring practical constraint in dental AI, and in medical imaging AI more broadly, is the scarcity and cost of expert-annotated training data. This annotation bottleneck is a central motivation for self-supervised, foundation-model, and other label-efficient learning approaches in medical imaging [24–27, 30–33]. Large-scale radiology foundation models increasingly rely on self-supervised encoders and transferable representations, with evaluation on downstream tasks using limited supervision becoming

an important methodological question [30–33]. In dentistry, this issue has become especially current: a DINOv2-initialized, domain-adapted self-supervised framework recently achieved strong segmentation performance on periapical films using only 229 labelled images after pretraining on 74,292 unlabelled films [34], while an unsupervised few-shot approach has also been evaluated for periodontal diagnosis on panoramic radiographs [35]. Dental-specific panoramic foundation models are now emerging at substantially larger scale [36], and recent benchmarking work has evaluated DINO-family encoders across multiple dental image-analysis tasks [4]. These developments strengthen the case for explicitly evaluating not only absolute downstream accuracy but also how efficiently alternative pretrained representations use scarce annotations.

Recent panoramic-radiography AI studies also illustrate how rapidly the task-specific literature is expanding. Contemporary work includes multistage and transformer-based caries detection [37–39, 49, 56], periapical lesion classification and segmentation [40–43], impacted-tooth detection and classification with hybrid CNN-transformer, YOLO-based, and embedding-based approaches [44, 45, 53, 54], and newer deep-learning systems for tooth identification and numbering [46, 47]. At the other end of the scale, 2026 systems such as DentFound and DentVLM illustrate the rapid move toward large multi-centre and multimodal dental foundation models [59, 60]. These systems address comprehensive diagnosis and report generation rather than the controlled frozen-representation question studied here. Recent evidence syntheses further document the breadth and heterogeneity of this field: systematic reviews/meta-analyses have evaluated tooth detection and segmentation [50], periapical radiolucency detection [51, 52], and CNN-versus-Vision-Transformer approaches across dental imaging tasks [55]. The latter review found that panoramic radiography was the most common imaging modality among included CNN-versus-ViT comparisons, while also emphasizing that neither architecture is uniformly superior across tasks [55]. This literature demonstrates increasingly strong task-specific performance, but it predominantly optimizes individual end-to-end systems. It does not directly answer the representation-level question addressed here: how self-supervised and supervised pretrained visual features compare under identical downstream classifiers and deliberately matched annotation budgets.

The primary study compared frozen DINOv2-small self-supervised features with frozen, supervised ImageNet-pretrained ResNet50 features for tooth-level multi-label diagnosis classification on DENTEX, using the established linear-evaluation paradigm [28, 29]. These models were selected as practical off-the-shelf representatives of two widely used transfer-learning regimes: a modern self-supervised Vision Transformer and a conventional supervised convolutional ImageNet backbone. A randomly initialized ResNet50 served as a negative control. Because this primary comparison confounds pretraining regime with backbone family, we additionally performed a secondary architecture-matched sensitivity analysis using a supervised ImageNet-pretrained ViT-S/16 of similar model scale. This additional comparator was not treated as a perfectly controlled pretraining-objective ablation, because patch size, pretraining corpus, and training recipe still differ from DINOv2. The central question was whether representation-level differences persist when expert-labelled training examples are deliberately constrained.

Materials and methods

Dataset and prediction unit

The fully annotated subset of the public DENTEX dataset (705 panoramic radiographs, three institutions, heterogeneous acquisition equipment) was used [2]. Ground-truth diagnosis annotations are provided for four labels — Caries, Deep Caries, Periapical Lesion, and Impacted — at the level of individual anatomical teeth, defined by the triplet (image identifier, quadrant, tooth position). The analytic manifest used in this study therefore treats the source radiograph as the highest available grouping unit; patient-level demographic variables were not incorporated into model development or evaluation. This study addresses diagnosis classification conditional on an already-localized, annotated abnormal-tooth ROI; it does not address end-to-end detection or screening of unselected radiographs.

Region-of-interest construction

For each anatomical tooth, all diagnosis-specific bounding boxes assigned to the same (image identifier, quadrant, tooth position) triplet were combined into a single union bounding box. Local context was added symmetrically by extending each horizontal side by 10% of the union-box width and each vertical side by 10% of its height, with coordinates clipped to the original image boundaries. The resulting region was cropped from the original-resolution RGB radiograph. Each crop was then resized with aspect ratio preserved using Lanczos resampling so that it fit within 224×224 pixels and was centered on a 224×224 canvas with black padding (letterboxing); no geometric cropping was used during this resizing step. For teeth carrying more than one diagnosis label, the same union ROI was used for all labels.

Sample size. No a priori sample-size calculation was performed because the analysis used the complete eligible fully annotated DENTEX subset available for the study. Statistical uncertainty and robustness were instead characterized using grouped resampling and cross-validation.

Data split

A radiograph-grouped split (train 65% / validation 15% / test 20%) was used, so that no radiograph — and therefore no tooth from that radiograph — appeared in more than one partition. Among 5000 candidate group-shuffle splits, the split minimizing a combined score of (i) deviation of each label's prevalence in every partition from its overall prevalence and (ii) deviation of partition size from its target fraction was selected and frozen for all subsequent analyses.

Visual representations

Four frozen image representations were evaluated across the complete analysis: (1) DINOv2-small [3] (Hugging Face checkpoint “facebook/dinov2-small”), a self-supervised Vision Transformer with a 384-dimensional output; (2) ResNet50 pretrained on ImageNet-1k under supervised classification (torchvision IMAGENET1K_V2 weights; 2048-dimensional global-average-pooled output) [8, 9]; (3) a randomly initialized ResNet50 with the same architecture and no pretraining, used as a negative control; and (4), in a secondary architecture-matched sensitivity analysis, a supervised ViT-S/16 (timm checkpoint “vit_small_patch16_224.augreg_in21k_ft_in1k”), a 22.1-million-parameter Vision Transformer

with 16×16 patches and 384-dimensional features, pretrained on ImageNet-21k and fine-tuned on ImageNet-1k using the AugReg training recipe [57, 58]. All backbones were frozen and only label-specific linear classifiers were trained. For the supervised ViT, model-specific inference preprocessing was obtained programmatically with timm.data.resolve_model_data_config and timm.data.create_transform (is_training=False), corresponding to 224×224 input, bicubic interpolation, center cropping, and channel-wise normalization with mean and standard deviation of 0.5. The supervised ViT comparator was introduced to test whether the DINOv2 few-shot finding persisted against another transformer backbone of similar model scale; it was not considered a causal isolation of self-supervised pretraining because the two ViTs differ in patch size, pretraining corpus, objective, and training recipe.

Classifier and model selection

For each representation and each of the four labels, a one-vs-rest logistic regression linear probe was fit on the training partition. Regularization strength ($C \in \{0.01, 0.1, 1, 10, 100\}$), class weighting (none or “balanced”), and the decision threshold (grid search over 0.05–0.95 in steps of 0.01, maximising validation F1) were selected using the validation partition only, independently for every label and representation. The test partition was not used for any model or threshold selection and was evaluated exactly once per representation for the primary analysis.

Full-data evaluation

Using the validation-selected configuration, each representation's linear probes were refit on the training partition and evaluated once on the locked test partition. Macro-averaged F1 across the four labels was the primary outcome measure, given the substantially different prevalence of the four labels; micro-averaged F1 and per-label sensitivity, specificity, precision, and F1 are reported as secondary measures.

Robustness and sensitivity analyses

Radiograph-cluster bootstrap. Because multiple teeth can originate from the same radiograph and are therefore not independent observations, uncertainty was quantified with a cluster bootstrap resampling whole radiographs (with replacement) rather than individual teeth, using 2000 resamples, both for overall macro/micro-F1 and, in a supplementary analysis, for each label individually.

Cluster-level permutation test. As an independent, non-parametric complement to the bootstrap confidence interval, a paired sign-flip permutation test (5000 permutations) was performed for the full-data DINOv2-versus-ImageNet macro-F1 difference: in each permutation, the two representations' predictions were exchanged, radiograph by radiograph, with probability 0.5, and the resulting null distribution of macro-F1 differences was compared with the observed difference.

Nested radiograph-grouped cross-validation. As a complementary estimate of performance independent of the single frozen split, the complete labelled pool was partitioned into five outer folds using GroupKFold with source image_id as the grouping variable, ensuring that teeth from the same radiograph could not occur in both outer-training and outer-test data. Within each outer-training pool, a single

radiograph-grouped 80:20 inner train/validation split was generated with GroupShuffleSplit. For every representation and diagnosis label, the inner validation set was used to select C ∈ {0.01, 0.1, 1, 10, 100}, class_weight ∈ {None, balanced}, and a decision threshold from 0.05 to 0.95 in increments of 0.05 by maximizing label-specific F1. The selected configuration was then refitted on the complete outer-training pool and evaluated once on the held-out outer-test fold. This analysis was supplementary to, and did not replace or modify, the prespecified locked train/validation/test analysis.

Few-shot sample-efficiency analysis. The repeated analysis used label-specific balanced training subsets at k ∈ {5, 10, 20, 50, 100}. For each one-vs-rest diagnosis task, up to k positive and k negative training teeth were sampled. Sampling preferentially selected teeth from distinct source radiographs, with additional teeth from represented radiographs used only when required to reach k. Multi-label teeth were handled independently for each label. Fifty independently seeded repetitions were performed at every k. Identical label-specific sampled indices were reused across DINOv2, ImageNet-pretrained ResNet50, random ResNet50, and, in the secondary architecture-matched analysis, supervised ViT-S/16. For the three original representations, previously validation-selected hyperparameters and thresholds were held fixed. The supervised ViT underwent the same validation-only selection procedure before locked-test and few-shot evaluation. All repetitions used the same locked test partition.

Secondary grouped-split robustness. As a further check on dependence on the single frozen split, the full-data comparison was repeated across 20 independently resampled radiograph-grouped splits.

Statistical analysis

Ninety-five percent confidence intervals for the primary locked-test analyses were computed as radiograph-cluster bootstrap percentile intervals unless otherwise stated. For each original DINOv2-versus-ImageNet few-shot budget, the 50 matched macro-F1 differences were summarized by mean and standard deviation, and a 95% bootstrap interval for the mean paired difference was obtained from 5,000 bootstrap resamples. Architecture-matched DINOv2-versus-supervised-ViT comparisons were summarized as paired mean differences, standard deviations, and the number of DINOv2-favouring repetitions. Because all few-shot repetitions share the same locked test partition, these summaries quantify training-sample sensitivity rather than uncertainty from independent datasets. Per-label and architecture-matched analyses were secondary/exploratory; the prespecified primary endpoint remained four-label macro-F1 for DINOv2 versus ImageNet-pretrained ResNet50.

Reporting

This study is reported with reference to the TRIPOD + AI statement for studies developing or evaluating a prediction model that uses regression or machine-learning methods [12], insofar as its items apply to a representation-comparison design rather than a single deployable model.

Results

Full training-set comparison

On the locked test partition of the frozen radiograph-grouped split, DINOv2 achieved a macro-F1 of 0.624 (micro-F1 0.748), ImageNet-pretrained ResNet50 achieved a macro-F1 of 0.618 (micro-F1 0.756), and the randomly initialized ResNet50 achieved a macro-F1 of 0.464 (micro-F1 0.658) (Table 1; per-label positive counts are reported in Table 2). The radiograph-cluster bootstrap 95% confidence interval for the DINOv2-minus-ImageNet macro-F1 difference was − 0.057 to 0.068 (bootstrap probability of a positive difference, 0.61), indicating no statistically robust full-data advantage for either representation. This was corroborated by an independent cluster-level permutation test (observed difference + 0.006; two-sided p = 0.88) and by nested 5-fold radiograph-grouped cross-validation, in which mean macro-F1 was 0.633 (SD 0.025) for DINOv2 and 0.617 (SD 0.014) for ImageNet, with overlapping fold-level ranges. Both pretrained representations substantially and robustly exceeded the randomly initialized control (DINOv2 minus random, + 0.161, 95% CI 0.107–0.212; ImageNet minus random, + 0.153, 95% CI 0.101–0.209), confirming that the pipeline captured genuine, pretraining-dependent representation quality rather than an artefact of the classification procedure.

Table 1
Prespecified primary full-data test-set performance for DINOv2, ImageNet-pretrained ResNet50, and the random-feature negative control (locked test partition; validation-selected hyperparameters and thresholds). The supervised ViT was introduced only as a secondary architecture-matched sensitivity comparator and is reported separately.

Representation	Macro-F1	Micro-F1	Macro precision	Macro recall
DINOv2-small (frozen)	0.624	0.748	0.579	0.705
ResNet50, ImageNet-pretrained (frozen)	0.618	0.756	0.562	0.687
ResNet50, randomly initialized (negative control)	0.464	0.658	0.410	0.551
Per-label full-data performance				

Per-label F1 varied considerably with label prevalence (Table 2). Caries (observed test prevalence 64.4%) and Impacted (17.5%) were classified with high F1 by all pretrained representations. Deep Caries (15.5%) and, in particular, Periapical Lesion (4.6%; 33 positive teeth in the test partition) were classified with substantially lower and less stable F1. Radiograph-cluster bootstrap confidence intervals computed separately for each label showed that, for Periapical Lesion, the point-estimate ranking between DINOv2 (F1 0.385, 95% CI 0.204–0.561) and ImageNet (F1 0.436, 95% CI 0.255–0.607) reversed relative to the other labels, but the two confidence intervals overlapped almost completely, indicating that neither representation's apparent advantage on this label is statistically distinguishable from the other given the small number of positive test examples. The randomly initialized control performed near floor on this label (F1 0.049, 95% CI 0-0.133), confirming that both pretrained representations retain genuine, if limited, discriminative information for this rarer finding. Accordingly, label-specific comparisons for

Periapical Lesion are considered exploratory and should not be used to infer superiority of either pretrained representation.

Table 2
Full-data per-label F1 (95% radiograph-cluster bootstrap confidence interval in parentheses).

Label	DINOv2	ImageNet ResNet50	Random ResNet50 (control)
Caries (n = 457 positive)	0.853 (0.828–0.878)	0.857 (0.832–0.883)	0.821
Deep Caries (n = 110)	0.391 (0.319–0.464)	0.339 (0.241–0.425)	0.293
Impacted (n = 124)	0.873 (0.817–0.918)	0.833 (0.771–0.883)	0.686
Periapical Lesion (n = 33)	0.385 (0.204–0.561)	0.436 (0.255–0.607)	0.049 (0.000-0.133)
Sample-efficiency (few-shot) comparison			

Across all five constrained annotation budgets, the 50-repetition analysis showed a consistent ordering of DINOv2 > ImageNet-pretrained ResNet50 > random ResNet50 (Table 3; Fig. 1). DINOv2 mean macro-F1 increased from 0.4298 at k = 5 to 0.5317 at k = 100, compared with 0.3606 to 0.4830 for ImageNet ResNet50 and 0.2725 to 0.3847 for random ResNet50. The paired DINOv2-minus-ImageNet advantage was + 0.0692, + 0.0732, + 0.0668, + 0.0577, and + 0.0488 at k = 5, 10, 20, 50, and 100, respectively; DINOv2 was favoured in all 50 repetitions at every budget. Bootstrap intervals for the mean paired differences excluded zero at all five budgets.

Table 3
Paired few-shot sample-efficiency comparison: DINOv2 minus ImageNet-pretrained ResNet50 macro-F1 across 50 matched training-set resampling repetitions per budget. Each k denotes up to k positive plus k negative examples for each label-specific binary probe.

Up to k positive and k negative examples per label	Mean paired Δ macro-F1	95% bootstrap interval for mean Δ	DINOv2 better (of 50 repetitions)
5	+ 0.0692	0.0619 to 0.0767	50/50
10	+ 0.0732	0.0665 to 0.0796	50/50
20	+ 0.0668	0.0626 to 0.0713	50/50
50	+ 0.0577	0.0534 to 0.0619	50/50
100	+ 0.0488	0.0448 to 0.0527	50/50

Architecture-matched supervised-ViT sensitivity

A supervised ImageNet-pretrained ViT-S/16 was added as a secondary architecture-matched comparator. Its full-data macro-F1 was 0.6088 (Caries 0.8557; Deep Caries 0.3417; Impacted 0.8456; Periapical Lesion 0.3922). Under matched few-shot sampling, mean macro-F1 ranged from 0.4159 at k = 5 to 0.4986 at k = 100. DINOv2 exceeded the supervised ViT on average at every budget, with paired mean differences of + 0.0139, + 0.0220, + 0.0262, + 0.0317, and + 0.0331 at k = 5, 10, 20, 50, and 100, respectively. DINOv2 was favoured in 36/50, 42/50, 49/50, 50/50, and 50/50 matched repetitions. This secondary result indicates that the sample-efficiency advantage was not confined to comparison with a convolutional supervised backbone.

Table 4
Architecture-matched few-shot sensitivity: DINOv2 versus supervised ViT-S/16 across 50 matched training-set resampling repetitions.

k	DINOv2 mean (SD)	Supervised ViT mean (SD)	Mean paired Δ	DINOv2 better
5	0.4298 (0.0268)	0.4159 (0.0285)	+ 0.0139	36/50
10	0.4623 (0.0182)	0.4403 (0.0188)	+ 0.0220	42/50
20	0.4855 (0.0133)	0.4594 (0.0140)	+ 0.0262	49/50
50	0.5139 (0.0126)	0.4822 (0.0113)	+ 0.0317	50/50
100	0.5317 (0.0089)	0.4986 (0.0078)	+ 0.0331	50/50
Sensitivity to threshold and hyperparameter choice				

Full-data macro-F1 was similar whether the validation-selected, per-label threshold or a fixed threshold of 0.5 was used for all three representations (DINOv2: 0.624 vs 0.636; ImageNet: 0.618 vs 0.619; random control: 0.463 vs 0.459), indicating that the primary comparison was not an artefact of threshold selection. A denser, purely descriptive grid over the regularization parameter C (1 to 200, thresholds held fixed) showed a non-monotonic, sign-unstable DINOv2-minus-ImageNet difference (range − 0.050 to + 0.030) with no consistent plateau, reinforcing rather than contradicting the conclusion, from the bootstrap and permutation analyses, that no robust full-data advantage exists for either representation.

Discussion

In this tooth-level multi-label classification task, frozen DINOv2 and supervised ImageNet-pretrained ResNet50 features were statistically indistinguishable on the prespecified full-data split, whereas DINOv2 showed a consistent advantage when labelled training examples were constrained. The 50-repetition analysis demonstrated DINOv2 superiority over ImageNet ResNet50 in all 50 matched repetitions at every annotation budget, while the random-feature control remained substantially lower. This pattern is directionally concordant with recent dental studies specifically addressing annotation scarcity: domain-adapted DINOv2 enabled segmentation of seven structures from only 229 labelled periapical films [34], and an unsupervised few-shot framework has been evaluated for periodontal classification from panoramic radiographs [35]. More broadly, contemporary reviews identify label efficiency as a central

challenge for medical imaging and highlight foundation-model transfer as an increasingly important strategy [26, 30–33]. The present study nevertheless addresses a narrower representation-level question by keeping the downstream classifier framework fixed. Accordingly, the primary interpretation is comparative representation quality under constrained annotation, rather than a general claim that one learning paradigm is intrinsically superior.

The secondary supervised-ViT analysis materially refines interpretation of the few-shot result. The supervised ViT-S/16 achieved full-data macro-F1 0.6088, close to the two original pretrained representations, but its few-shot performance remained below DINOv2 at every budget: mean paired DINOv2-minus-ViT differences were + 0.0139, + 0.0220, + 0.0262, + 0.0317, and + 0.0331 from k = 5 through k = 100, with DINOv2 favoured in 36/50, 42/50, 49/50, 50/50, and 50/50 repetitions, respectively. Thus, the DINOv2 sample-efficiency advantage was not restricted to comparison with a convolutional supervised backbone. This is important because a recent systematic review of dental image processing found frequent task-dependent differences between CNN and ViT models rather than uniform superiority of one architecture [55]. However, the sensitivity analysis does not establish that self-supervision itself caused the advantage: the supervised comparator was pretrained on ImageNet-21k and fine-tuned on ImageNet-1k with AugReg [58], whereas DINOv2 was pretrained self-supervised on a much larger curated image corpus [3]. Patch size and training recipe also differ. The observed differences should therefore be interpreted as properties of the complete pretrained representations—including architecture, pretraining objective, corpus, and optimization recipe—rather than as an isolated causal estimate of self-supervision. The magnitude and budget dependence of the advantage also differed by comparator: the DINOv2-ResNet50 gap attenuated with increasing k, whereas the DINOv2-supervised-ViT gap increased modestly.

From a practical model-development perspective, the few-shot gains quantify annotation efficiency rather than demonstrated clinical benefit. For example, at k = 50, mean locked-test macro-F1 was 0.5139 for DINOv2, 0.4822 for the supervised ViT, and 0.4562 for ImageNet ResNet50. No validated clinical-utility threshold exists for macro-F1 in this experimental, ROI-conditioned task, so these differences should not be labelled clinically meaningful. Their practical relevance is that, for a fixed small expert-annotation budget, the choice of frozen representation changed downstream performance consistently. More generally, representation choice became consequential precisely when annotation was constrained, even though the pretrained representations approached similar performance in the full-data setting. Frozen-feature workflows also front-load computation: embeddings can be extracted once and reused while the logistic probes are inexpensive to retrain. We did not benchmark inference latency, memory use, energy consumption, or carbon footprint, however, and therefore cannot claim that the observed accuracy gains are computationally or economically cost-effective.

Several limitations should be considered when interpreting these results. First, the dataset originates from a single challenge cohort (three institutions, but a single curated release) [2], and generalisability to different equipment or patient populations was not assessed. Recent panoramic foundation and vision-language systems have been developed and evaluated at substantially larger multi-centre scale [36, 59,

60], underscoring that the present results require independent replication rather than further internal resampling to establish transportability. These systems address broader end-to-end diagnostic or multimodal objectives than the present frozen-representation experiment. The current findings therefore define a complementary research setting: when extensive task-specific annotation or development of a domain-specific foundation model is not feasible, a frozen general-purpose representation may provide a strong starting point for label-efficient downstream modelling; this should not be interpreted as equivalence to, or replacement of, domain-specific foundation models. Second, Periapical Lesion had only 33 positive test examples and unstable per-label estimates. Its full-data point estimate even reversed the ranking seen for Deep Caries and Impacted teeth (DINOv2 F1 0.3793 versus ImageNet ResNet50 0.4384), illustrating how a rare label can dominate a four-label macro-average. In a secondary sensitivity analysis excluding Periapical Lesion, the DINOv2-minus-ImageNet difference increased from + 0.0062 to + 0.0279 macro-F1, but the radiograph-cluster bootstrap interval still included zero (-0.0066 to + 0.0652). We therefore retain the prespecified four-label endpoint and interpret the Periapical result as an underpowered rare-label stress case rather than using post-hoc label exclusion to redefine the main task. Third, the study evaluated frozen linear probes rather than end-to-end fine-tuning. Fourth, the task was diagnosis classification conditional on an already-localized annotated abnormal-tooth ROI and should not be extrapolated to end-to-end screening. Fifth, the supervised ViT comparator reduces, but does not eliminate, architecture-related confounding: DINOv2-small and supervised ViT-S/16 still differ in patch size, pretraining corpus, objective, and training recipe. Sixth, probability calibration was not evaluated. Finally, patient-level demographic and detailed acquisition metadata were not incorporated, limiting subgroup and equipment-specific generalizability assessments.

In conclusion, frozen DINOv2-small did not show a statistically robust advantage over ImageNet-pretrained ResNet50 on the prespecified full-data comparison, but it consistently provided the strongest performance under constrained annotation budgets. The few-shot advantage persisted against a supervised Vision Transformer of similar model scale, reducing concern that the original result was solely a CNN-versus-transformer effect. These findings support DINOv2 as a strong candidate frozen representation for annotation-limited dental imaging research, while not isolating self-supervised pretraining as the causal source of the advantage. External validation, calibration assessment, and end-to-end fine-tuning studies remain necessary before broader generalisation or clinical-use claims can be made.

Declarations

Funding

This work received no external funding.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics approval

This study was conducted in accordance with the ethical principles of the World Medical Association Declaration of Helsinki. No new human participant data were collected; all panoramic radiographs analysed were obtained from the publicly available, de-identified DENTEX dataset [2], and no identifiable patient information was accessed by the authors. Consistent with precedent for secondary analyses of public, de-identified imaging datasets in this literature [13], no additional institutional ethics committee approval or informed consent was required for this analysis.

Consent to participate / Consent to publish

Not applicable — this study did not involve new data collection from identifiable individuals; all radiographs analysed were obtained from a pre-existing, publicly available, de-identified dataset released under its own data-use terms [2].

Data and code availability

The DENTEX dataset is publicly available (https://dentex.grand-challenge.org/). Derived, non-identifiable result tables and analysis code supporting the findings of this study are available from the corresponding author on reasonable request.

Author contributions

A.S. conceived and designed the study, performed data curation, feature extraction, model development, and statistical analysis, and drafted the manuscript. P.D. and E.A. contributed to study design and critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

During preparation of this work, the authors used ChatGPT (OpenAI) solely for language editing and improvement of readability. The authors reviewed and edited the output and take full responsibility for the content of the manuscript.

References

1. Rushton VE, Horner K. The use of panoramic radiology in dental practice. J Dent. 1996;24(3):185–201.
2. Hamamci IE, Er S, Simsar E, Yuksel AE, Gultekin S, Ozdemir SD et al. DENTEX: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic X-rays. arXiv preprint. 2023; arXiv:2305.19112.
3. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V et al. DINOv2: Learning robust visual features without supervision. Trans Mach Learn Res. 2024. https://openreview.net/forum?id=a68SUt6zFt
4. Tang K, Yang X, Zheng M, Liu X, Li X, Guo X, Chen R, Shen L, Meng H, DinoDental. Benchmarking DINOv3 as a unified vision encoder for dental image analysis. arXiv preprint. 2026; arXiv:2603.28297.
5. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
6. Mureșanu S, Hedeșiu M, Iacob L–M, Eftimie R, Olariu E, Dinu C, et al. Automating dental condition detection on panoramic radiographs: challenges, pitfalls, and opportunities. Diagnostics. 2024;14(20):2336. https://doi.org/10.3390/diagnostics14202336.
7. Li X, Ma X, Zhao Y, Hu J, Liu J, Yang Z, Han F, Zhang J, Liu W, Zhou Z. A multi-center dental panoramic radiography image dataset for impacted teeth, periodontitis, and dental caries: benchmarking segmentation and classification tasks. J Imaging Inf Med. 2024;37(2):831–41. https://doi.org/10.1007/s10278-024-00972-8.
8. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. pp. 770–778.
9. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L, ImageNet. A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. pp. 248–255.
10. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman and Hall/CRC; 1994.
11. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57(1):289–300.
12. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.
13. van Nistelrooij N, El Ghoul K, Xi T, Saha A, Kempers S, Cenci M, et al. Combining public datasets for automated tooth assessment in panoramic radiographs. BMC Oral Health. 2024;24(1):387. https://doi.org/10.1186/s12903-024-04129-5.
14. Molander B, Ahlqwist M, Gröndahl HG, Hollender L. Comparison of panoramic and intraoral radiography for the diagnosis of caries and periapical pathology. Dentomaxillofac Radiol. 1993;22(1):28–32. https://doi.org/10.1259/dmfr.22.1.8508938.
15. Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent. 2018;77:106–11.

https://doi.org/10.1016/j.jdent.2018.07.015.

16. Vinayahalingam S, Kempers S, Limon L, Deibel D, Maal T, Hanisch M, Bergé S, Xi T. Classification of caries in third molars on panoramic radiographs using deep learning. Sci Rep. 2021;11(1):12609. https://doi.org/10.1038/s41598-021-92121-2.
17. Başaran M, Çelik Ö, Bayrakdar IS, Bilgir E, Orhan K, Odabaş A, Aslan AF, Jagtap R. Diagnostic charting of panoramic radiography using deep-learning artificial intelligence system. Oral Radiol. 2022;38(3):363–9. https://doi.org/10.1007/s11282-021-00572-0.
18. 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. https://doi.org/10.1007/s00521-021-06684-2.
19. Çelik B, Savaştaer EF, Kaya HI, Çelik ME. The role of deep learning for periapical lesion detection on panoramic radiographs. Dentomaxillofac Radiol. 2023;52(8):20230118. https://doi.org/10.1259/dmfr.20230118.
20. Boztuna M, Firincioglulari M, Akkaya N, Orhan K. Segmentation of periapical lesions with automatic deep learning on panoramic radiographs: an artificial intelligence study. BMC Oral Health. 2024;24(1):1332. https://doi.org/10.1186/s12903-024-05126-4.
21. Celik ME. Deep Learning Based Detection Tool for Impacted Mandibular Third Molar Teeth. Diagnostics (Basel). 2022;12(4):942. https://doi.org/10.3390/diagnostics12040942.
22. He Z, Wang Y, Li X. Deep Learning-Based Detection of Impacted Teeth on Panoramic Radiographs. Biomed Eng Comput Biol. 2024;15:11795972241288319. https://doi.org/10.1177/11795972241288319.
23. Tuzoff DV, Tuzova LN, Bornstein MM, Krasnov AS, Kharchenko MA, Nikolenko SI, Sveshnikov MM, Bednenko GB. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. Dentomaxillofac Radiol. 2019;48(4):20180051. https://doi.org/10.1259/dmfr.20180051.
24. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digit Med. 2023;6(1):74. https://doi.org/10.1038/s41746-023-00811-0.
25. Shurrab S, Duwairi R. Self-supervised learning methods and applications in medical imaging analysis: a survey. PeerJ Comput Sci. 2022;8:e1045. https://doi.org/10.7717/peerj-cs.1045.
26. Jin C, Guo Z, Lin Y, Luo L, Chen H. Learning with less supervision: a survey of label-efficient learning for medical image analysis. Med Image Anal. 2026;111:104062. https://doi.org/10.1016/j.media.2026.104062.
27. Zhou Z, Sodha V, Pang J, Gotway MB, Liang J. Models Genesis. Med Image Anal. 2021;67:101840. https://doi.org/10.1016/j.media.2020.101840.
28. Chen T, Kornblith S, Norouzi M, Hinton G. A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 37th International Conference on Machine Learning. PMLR. 2020;119:1597–1607.

29. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A. Emerging Properties in Self-Supervised Vision Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021:9650–9660. https://doi.org/10.1109/ICCV48922.2021.00951
30. Paschali M, Chen Z, Blankemeier L, Varma M, Youssef A, Bluethgen C, et al. Foundation Models in Radiology: What, How, Why, and Why Not. Radiology. 2025;314(2):e240597. https://doi.org/10.1148/radiol.240597.
31. de Almeida JG, Cerdá Alberich L, Tsakou G, Marias K, Tsiknakis M, Lekadir K, et al. Foundation models for radiology—the position of the AI for Health Imaging (AI4HI) network. Insights Imaging. 2025;16:168. https://doi.org/10.1186/s13244-025-02056-9.
32. Ryu JS, Kang H, Chu Y, et al. Vision-language foundation models for medical imaging: a review of current practices and innovations. Biomed Eng Lett. 2025;15:809–30. https://doi.org/10.1007/s13534-025-00484-6.
33. Sun Y, Wen X, Zhang Y, Jin L, Yang C, Zhang Q, et al. Visual-language foundation models in medical imaging: A systematic review and meta-analysis of diagnostic and analytical applications. Comput Methods Programs Biomed. 2025;268:108870. https://doi.org/10.1016/j.cmpb.2025.108870.
34. Hu M, Zhang Q, Wei Z, Jia P, Yuan M, Yu H, Yin X, Peng J. Self-supervised learning enhances periapical films segmentation with limited labeled data. J Dent. 2025;163:106150. https://doi.org/10.1016/j.jdent.2025.106150.
35. Kim MJ, Chae SG, Bae SJ, et al. Unsupervised few shot learning architecture for diagnosis of periodontal disease in dental panoramic radiographs. Sci Rep. 2024;14:23237. https://doi.org/10.1038/s41598-024-73665-5.
36. Han X, Wu Z, Wang Z, Dong S, Zhang R, Liu J, Zhang L. PanoFM: An LLM-empowered panoramic foundation model with clinical semantic integration for comprehensive dental disease diagnosis. Pattern Recognit. 2026;180:114411. https://doi.org/10.1016/j.patcog.2026.114411.
37. Pornprasertsuk-Damrongsri S, Vachmanus S, Papasratorn D, et al. Clinical application of deep learning for enhanced multistage caries detection in panoramic radiographs. Sci Rep. 2025;15:33491. https://doi.org/10.1038/s41598-025-16591-4.
38. Biltekin H, Geduk G, Altan A, Karasu S. Evaluation of deep learning systems in detection of dental caries on panoramic radiography. Am J Dent. 2025;38(4):163–8.
39. Biçengil K, Kurt A, Naralan ME. Deep Learning-Based Dental Caries Diagnosis on Panoramic Radiographies: Performance of YOLOv8 Versus Human Observers. Diagnostics (Basel). 2026;16(8):1150. https://doi.org/10.3390/diagnostics16081150.
40. Ver Berne J, Baseri Saadi S, Oliveira-Santos N, Marinho-Vieira LE, Jacobs R. Automated classification of panoramic radiographs with inflammatory periapical lesions using a CNN-LSTM architecture. J Dent. 2025;156:105688. https://doi.org/10.1016/j.jdent.2025.105688.
41. Ketenci Çay F, Yeşil Ç, Çay O, et al. DeepLabv3 + method for detecting and segmenting apical lesions on panoramic radiography. Clin Oral Investig. 2025;29:101. https://doi.org/10.1007/s00784-025-06156-0.

42. Alaqla A, Khanagar SB, Albelaihi AI, Singh OG, Alfadley A. Application and performance of artificial intelligence-based models in the detection, segmentation and classification of periapical lesions: a systematic review. Front Dent Med. 2025;6:1717343. https://doi.org/10.3389/fdmed.2025.1717343.
43. Pul U, Tichy A, Pitchika V, Schwendicke F. Impact of artificial intelligence assistance on diagnosing periapical radiolucencies: A randomized controlled trial. J Dent. 2025;160:105868. https://doi.org/10.1016/j.jdent.2025.105868.
44. Küçük DB, Imak A, Özçelik STA, Çelebi A, Türkoğlu M, Sengur A, Koundal D. Hybrid CNN-Transformer Model for Accurate Impacted Tooth Detection in Panoramic Radiographs. Diagnostics (Basel). 2025;15(3):244. https://doi.org/10.3390/diagnostics15030244.
45. Balel Y, Sağtaş K. Deep learning-based approach to third molar impaction analysis with clinical classifications. Sci Rep. 2025;15:23688. https://doi.org/10.1038/s41598-025-93783-y.
46. Peker RB, Kurtoglu CO. Evaluation of the Performance of a YOLOv10-Based Deep Learning Model for Tooth Detection and Numbering on Panoramic Radiographs of Patients in the Mixed Dentition Period. Diagnostics (Basel). 2025;15(4):405. https://doi.org/10.3390/diagnostics15040405.
47. Ozcelik E, Simsek H, Aktas A, Ilhan HO, et al. Teeth identification and numbering in mixed dentition: evaluating deep learning models for pediatric panoramic radiographs. BMC Oral Health. 2026;26:727. https://doi.org/10.1186/s12903-026-08097-w.
48. Ameli N, Miri Moghaddam M, Lai H, Pacheco-Pereira C. Automated quality evaluation of dental panoramic radiographs using deep learning. Imaging Sci Dent. 2025. https://doi.org/10.5624/isd.20240232.
49. Wang L, Li Z. Transformer-based intelligent detection model for early dental caries in panoramic radiographs. Sci Rep. 2026;16:3507. https://doi.org/10.1038/s41598-025-33391-y.
50. Bonfanti-Gris M, Herrera A, Salido Rodríguez-Manzaneque MP, Martínez-Rus F, Pradíes G. Deep learning for tooth detection and segmentation in panoramic radiographs: a systematic review and meta-analysis. BMC Oral Health. 2025;25:1280. https://doi.org/10.1186/s12903-025-06349-9.
51. Ahmad IA, Ba-Hattab R, Al-Maweri SA, Tamimi F. Deep learning models for detection of periapical radiolucent lesions on panoramic radiographs: a systematic review and meta-analysis. Int Endod J. 2026;1–39. https://doi.org/10.1111/iej.70207.
52. da Silva-Filho JE, da Silva Sousa Z, Caracas-de-Araújo AP, Dos Santos Fornagero L, Machado MP, de Aguiar AWO, Silva CM, de Albuquerque DF, Gurgel-Filho ED. Deep learning for detecting periapical bone rarefaction in panoramic radiographs: a systematic review and critical assessment. Dentomaxillofac Radiol. 2025;54(6):405–19. https://doi.org/10.1093/dmfr/twaf044.
53. Alam SS, Ahad A, Ahmed S, Dudley J, Farook TH. Using deep learning to segment impacted molar teeth from panoramic radiographs. Digit Dentistry J. 2025;1(1):100007. https://doi.org/10.1016/j.ddj.2025.100007.
54. Alenezi O, Bhattacharjee T, Alseed HA, Tosun YI, Chaudhry J, Prasad S. Evaluating the Efficacy of Various Deep Learning Architectures for Automated Preprocessing and Identification of Impacted

Maxillary Canines in Panoramic Radiographs. Int Dent J. 2025;75(5):100940. https://doi.org/10.1016/j.identj.2025.100940.

55. Felek T, Tercanlı H, Şendişçi Gök R. Evaluating vision transformers and convolutional neural networks in the context of dental image processing: a systematic review. BMC Oral Health. 2025;25:1626. https://doi.org/10.1186/s12903-025-07036-5.
56. Lim YS, Chun D, Kim J, Lee JY, Ju MJ, Park JH, Jeon HJ. Dual convolutional neural network framework for segmenting dental caries in panoramic radiographs. J Prosthet Dent. 2026;135(2):403. https://doi.org/10.1016/j.prosdent.2025.09.040. .e1–403.e8.
57. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T et al. An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations. 2021.
58. Steiner A, Kolesnikov A, Zhai X, Wightman R, Uszkoreit J, Beyer L. How to train your ViT? Data, augmentation, and regularization in Vision Transformers. Trans Mach Learn Res. 2022.
59. Zhu Q, Lin Y, Fu W, Tang W, Li J, Zhang Y et al. Towards clinical-level interpretation of dental panoramic radiography using an instance-guided vision-language model. Nat Biomed Eng 2026 Jun 25. https://doi.org/10.1038/s41551-026-01713-8
60. Meng Z, Hao J, Dai X, Feng Y, Liu J, Feng B, et al. A multimodal vision-language model for comprehensive dental diagnosis and enhanced clinical practice. Nat Commun. 2026 Jul;22. https://doi.org/10.1038/s41467-026-75718-x.

Figures

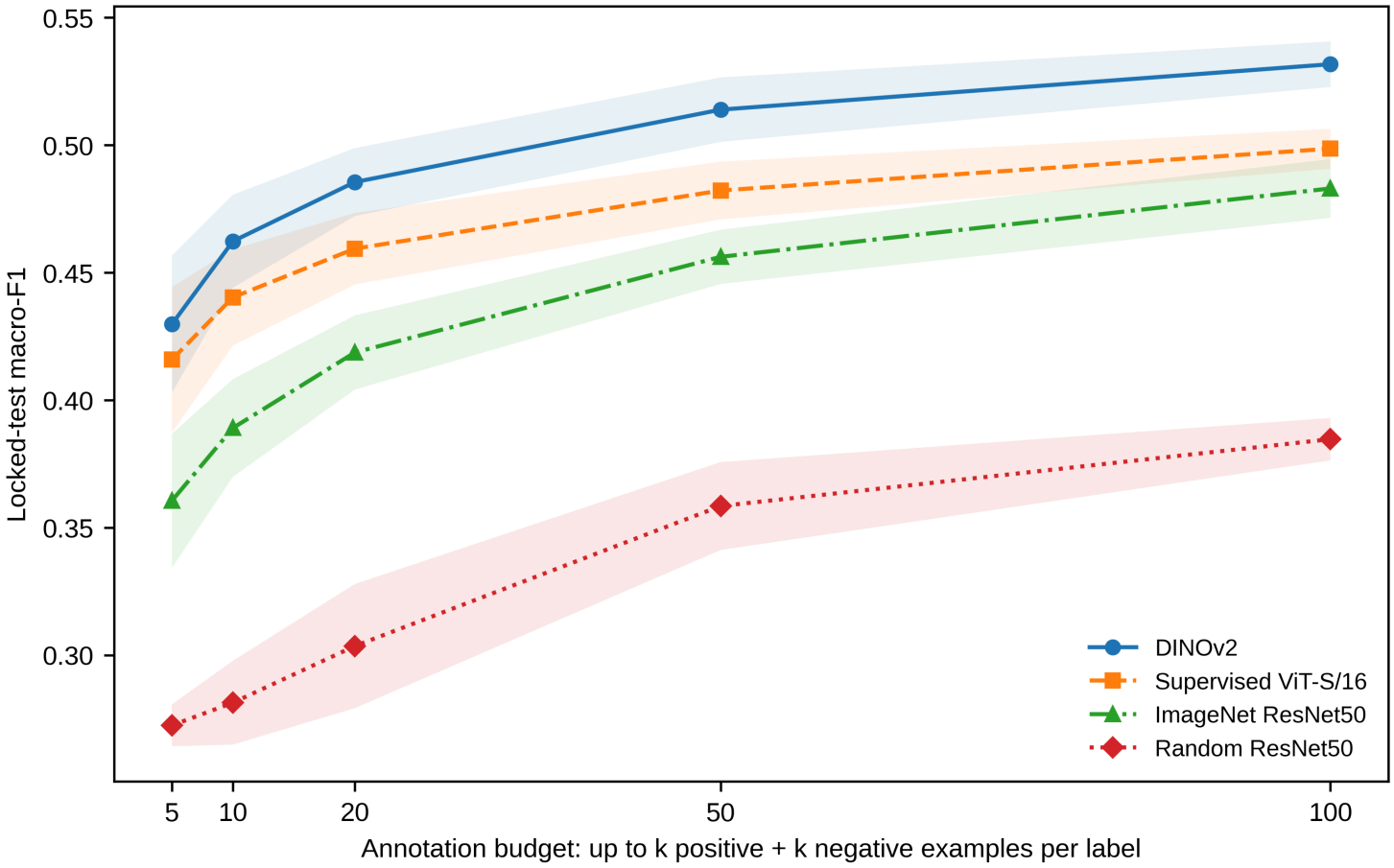


Figure 1

Few-shot locked-test macro-F1 across frozen visual representations. Points show means across 50 matched training-set resampling repetitions; shaded bands show ±1 SD. Identical label-specific sampled indices were used across representations within each repetition. Variability therefore reflects training-sample selection conditional on the same locked test partition