

Iterative Development of an AI Prototype for Dental Caries Detection and Localization with Informal Clinical Feedback

Chinmay Karthik Amancha
Chelmsford High School
Chelmsford, MA, USA
amancha.chinmay.usa@gmail.com

Abstract

This paper documents the iterative development of a student-built dental caries detection prototype and compares its final version, V4.1, with the preceding V4 system and a fresh factory-default YOLO11n baseline, testing sensitivity to validation selection and reader choice. The final phase used 100 public permanent-dentition bitewing radiographs with annotations from eight dentists and a fixed 60/20/20 image split. V4.1 selected a two-model fusion from 75 validation configurations and was evaluated exploratorily on a previously viewed 20-image holdout. Against reader E0 on this holdout, V4.1 achieved precision 0.415, recall 0.449, F1 0.431, AP50 0.491, and 1.55 false-positive boxes per image; V4 achieved 0.412, 0.286, 0.337, 0.271, and 1.00, and the factory-default baseline achieved 0.750, 0.306, 0.435, 0.458, and 0.25. Only three of the 75 validation configurations were within 0.05 F1 of the winner, and at threshold 0.16 V4.1 F1 ranged from 0.140 to 0.448 across the eight reader references. A precision-oriented threshold of 0.37 reduced V4.1 false positives to 0.50 per image with recall 0.367, and a conditional three-fold out-of-fold stress test produced F1 0.538 and AP50 0.510. V4.1 thus improved recall and AP50 over both comparators, but its F1 was essentially the same as the factory-default baseline and it produced more false-positive boxes. Performance depended strongly on reader choice and threshold, and the small reused holdout, concentrated validation winner, non-nested cross-validation, and absence of patient identifiers prevent clinical or causal claims.

Keywords

Artificial intelligence, dental caries, bitewing radiography, YOLO, object detection, multi-reader evaluation, precision-recall

ACM Reference Format:

Chinmay Karthik Amancha. 2026. Iterative Development of an AI Prototype for Dental Caries Detection and Localization with Informal Clinical Feedback. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 3, ISSUE 4)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.67149/yhjs2024.5/03kanv0l>

1 Introduction

Bitewing radiographs support detection of proximal caries, but interpretation varies among dentists and lesion boundaries are not perfectly reproducible. Deep-learning systems have reported promising results, yet reviews continue to identify small datasets,

heterogeneous reference standards, limited external validation, and risk of bias [1–4]. These concerns are especially important for a 100-image pilot.

This project began with broad pediatric panoramic experiments and later narrowed to one task—caries localization—on permanent-dentition bitewings after informal consultations with dentists. In particular, Dr. Mark Y. Wang, DMD, DMSc, a prosthodontist practising in Concord, Massachusetts, reviewed early prototype demonstrations and discussed missed or ambiguous caries findings. His feedback informed the research questions and focus on permanent-dentition caries; it did not constitute a clinical reader study or validation, and this observational history cannot establish that feedback caused any measured performance improvement.

The revised analysis asks a narrower question: how did a validation-locked V4.1 detector compare with the preceding V4 detector on the same reused holdout, and how stable were the conclusions across candidate configurations, thresholds, reader references, and an internal resampling analysis?

2 Materials and Methods

2.1 Study design and data

Versions 1–3 were formative experiments on a separate 100-image pediatric panoramic subset of the public Children’s Dental Panoramic Radiographs Dataset [8] and are summarized in Supplement S1. The quantitative study concerns V4 and V4.1, which used the public Dental Caries in Bitewing Radiographs dataset [6]. It contains 100 bitewings annotated independently by E0, E1–E4, and N1–N3. E0 supplied the primary training labels. The fixed split contained 60 training images (176 E0 boxes), 20 validation images (44 boxes), and 20 reused-holdout images (49 boxes). The release lacked patient identifiers, so patient-grouped independence could not be verified.

2.2 Comparators

V4 used a COCO-pretrained YOLO11n detector, 768-pixel inputs, batch size 4, AdamW with initial learning rate 0.001, seed 42, deterministic mode, a maximum of 40 epochs, early-stopping patience 12, horizontal flip probability 0.5, rotation up to 5 degrees, translation 0.05, scale 0.20, and brightness augmentation 0.10. Mosaic, mixup, copy-paste, and vertical flipping were disabled. The validation-selected threshold was 0.24 for the full-frame 1024-pixel inference comparator used in the paired report. This comparator was not an untouched factory-default YOLO11 training run.

The fresh factory-default comparator started from the official generic YOLO11n checkpoint and used Ultralytics defaults: 100

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

IJSCAR VOL. 3, ISSUE 4

© 2026 The Author(s). Published by IJSCAR.

epochs, 640-pixel inputs, batch size 16, seed 0, automatic optimizer selection, standard default augmentation, and best validation-fitness checkpoint selection. Only the dataset path, CPU device, zero data-loader workers, and output location were specified. A confidence threshold of 0.35 was locked on the validation set by maximum lesion F1, then precision, then the higher threshold; the holdout was evaluated only after lock.

2.3 V4.1 training and selection

V4.1 trained YOLO11n at 768 and 1024 pixels for up to 80 epochs and YOLO11s at 768 pixels for up to 60 epochs. The YOLO11n runs used batches of 4 and 2 and learning rate 0.001; YOLO11s used batch 2 and learning rate 0.0005. All used COCO-pretrained weights, AdamW, seed 42, and patience 20. The reported software environment and training settings are listed in the Supplement. Checkpoint identities were recorded by SHA-256 in the archived provenance files.

The validation grid included 72 single-checkpoint candidates and three two-model fusion variants. Ranking used lesion F1 at IoU 0.50, then AP50, precision, and lower complexity. The selected fusion combined a YOLO11n-768 checkpoint with full-frame 1024-pixel inference and a YOLO11n-1024 checkpoint with horizontally flipped 1024-pixel inference. The system and confidence threshold 0.16 were locked before holdout scoring.

2.4 Evaluation and sensitivity analyses

Lesion predictions were matched one-to-one to reference boxes at IoU at least 0.50. We report precision, recall, F1, false-positive boxes per image, 101-point AP50, and mAP50–95. Image-level flags used the highest detection score per image and are reported separately. Paired image bootstrap intervals used 5,000 resamples with seed 42. Saved V4.1 predictions were rescored separately against each of the eight readers at thresholds 0.16 and 0.37. Detections unmatched to E0 were also checked for overlap with each alternative reader. These reader-specific comparisons measure agreement, not diagnostic truth.

A completed three-fold out-of-fold stress test trained the selected recipe within the original 60-image training subset and applied the fixed threshold 0.16. Because model selection had already seen the full training subset, this is conditional recipe stability, not nested cross-validation or independent validation.

3 Results

3.1 Comparator results

3.2 Validation leaderboard

The selected validation F1 was 0.695. Three candidates tied at F1 0.695; all three were the same two-model fusion with different final suppression. The next-best single-checkpoint candidate had F1 0.643, 0.052 below the winner. The median candidate F1 was 0.556. This pattern is not a broadly flat leaderboard and reinforces the risk that a narrow winning cluster was selected from a small validation sample.

Table 1: Lesion-level results on the same 20-image reused holdout. Thresholds were selected on validation: V4 0.24, factory-default 0.35, and V4.1 0.16. The paired V4.1-minus-V4 bootstrap interval for F1 was 0.009–0.172; no bootstrap comparison was prespecified for the newly trained baseline.

Metric	V4	Factory default	V4.1
True positive lesions	14	15	22
False-positive boxes	20	5	31
Missed lesions	35	34	27
Precision	0.412	0.750	0.415
Recall	0.286	0.306	0.449
F1	0.337	0.435	0.431
AP50	0.271	0.458	0.491
mAP50–95	0.100	0.187	0.222
False positives per image	1.00	0.25	1.55

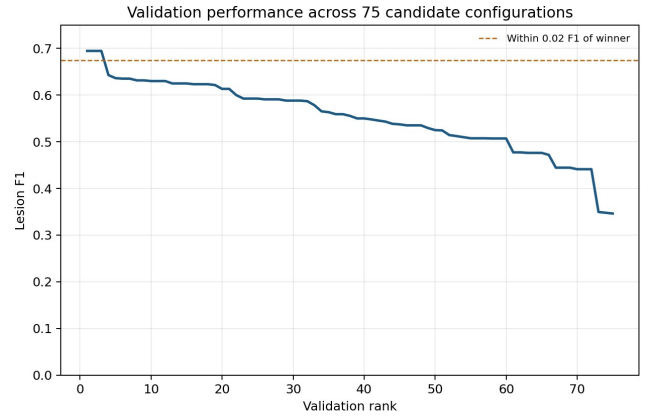


Figure 1: Lesion F1 for all 75 validation configurations, ordered by validation rank.

Table 2: Two validation-selected operating points applied unchanged to the reused holdout. The higher threshold substantially reduced false-positive boxes but missed more E0 lesions.

Threshold	Precision	Recall	F1	FP per image
0.16 balanced	0.415	0.449	0.431	1.55
0.37 precision-oriented	0.643	0.367	0.468	0.50

3.3 Threshold trade-off

At the image level, the balanced threshold produced sensitivity 0.867, specificity 0.800, accuracy 0.850, balanced accuracy 0.833, and ROC AUC 0.933. Specificity was based on only five E0-negative images; one image changes the estimate by 0.20. An image-level flag indicates that at least one detection exceeded threshold and does not establish that every lesion was localized.

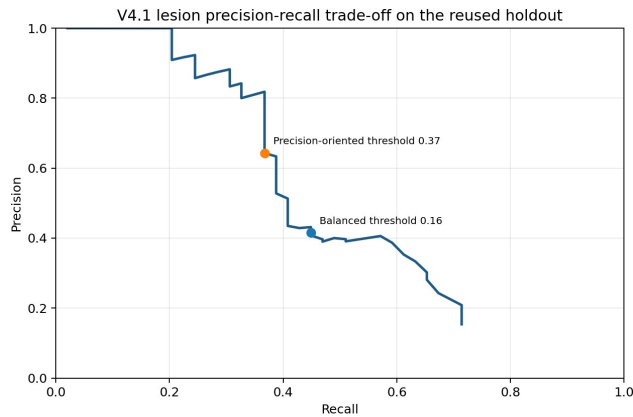


Figure 2: Precision-recall trade-off reconstructed from saved V4.1 holdout predictions and E0 boxes. AP50 was 0.491.

Table 3: V4.1 at the locked threshold 0.16, rescored separately against all eight readers. The 20 holdout images contained 47–85 boxes depending on reader, and F1 ranged from 0.140 to 0.448.

Reader	Boxes	Precision	Recall	F1	FP/image
E0	49	0.415	0.449	0.431	1.55
E1	85	0.302	0.188	0.232	1.85
E2	50	0.283	0.300	0.291	1.90
E3	49	0.358	0.388	0.373	1.70
E4	47	0.132	0.149	0.140	2.30
N1	81	0.566	0.370	0.448	1.15
N2	67	0.321	0.254	0.283	1.80
N3	63	0.434	0.365	0.397	1.50

3.4 Multi-reader sensitivity

Of 31 detections unmatched to E0, 13 overlapped at least one other reader’s box, 9 overlapped at least two, 5 overlapped at least three, and 3 overlapped at least four. The remainder cannot be assumed to be true disease or true error without adjudication. These results show that the reported false-positive count is reference-dependent.

3.5 Conditional cross-validation stress test

Across three out-of-fold partitions of the original 60-image training subset, the fixed-threshold results were 82 true positives, 47 false positives, and 94 missed lesions among 176 E0 boxes: precision 0.636, recall 0.466, F1 0.538, and 0.783 false positives per image. AP50 was 0.510 and mAP50–95 was 0.188. Image-level accuracy was 0.950 and ROC AUC 0.995, but these image-level estimates again do not measure complete lesion localization. Because recipe selection was not nested, the stress test is supportive rather than confirmatory.

3.6 Evaluated output

4 Discussion

V4.1 improved recall, F1, AP50, and mAP50–95 relative to the documented V4 comparator. Against the fresh factory-default baseline, V4.1 improved recall and AP50 but had nearly the same F1 (0.431 versus 0.435) and substantially more false-positive boxes (1.55 versus 0.25 per image). The precision-oriented V4.1 threshold illustrates the trade-off: reducing alerts to 0.50 per image increased precision but reduced recall. A clinical operating point cannot be chosen from this pilot; that choice would require a specified workflow, error costs, calibration, and independent validation.

The multi-reader analysis materially changes interpretation. Agreement was strongest with E0 and N1 and weakest with E4, and 42% of E0-unmatched detections overlapped at least one alternative reader. Therefore, “false positive” means unmatched to a selected reader under the specified IoU rule, not proven absence of caries. Consensus or adjudicated labels are needed for stronger claims.

The validation leaderboard was concentrated rather than flat. Although the configuration was locked before holdout scoring, selecting a narrow winning fusion from 75 candidates on 20 images creates a genuine validation-overfitting risk. The three-fold analysis is reassuring about recipe stability but is not nested and cannot remove this bias.

Published context also limits interpretation. Tichý and colleagues evaluated the ensemble developed in the companion Part I study [5], trained on 3,989 separate bitewings, against the original annotator on the same 100-image evaluation dataset and reported precision 0.78, recall 0.73, and F1 0.75 [7]. Those values were verified directly in the primary Part II paper. The system used far more training data and multiple detector families, so it provides context rather than a fair head-to-head benchmark for this 60-image training split.

The reviewer-requested factory-default comparison was completed as a fresh 100-epoch YOLO11n run. Its high precision and low false-positive burden accompanied low recall, while V4.1 recovered more E0 lesions and achieved higher AP50. Because all systems used the same small image split and reused holdout, this comparison does not solve the more important limitations of sample size, holdout reuse, and reference uncertainty.

The informal consultations influenced project framing and visualization questions, but neither clinician annotated the evaluation set or participated in a controlled experiment. The manuscript therefore does not claim that clinical feedback caused the measured gain. The insurance idea is limited to one future-work sentence because no insurer, claim outcome, or workflow endpoint was studied.

5 Limitations

This study used 100 images, a 20-image validation set, and a previously examined 20-image holdout. Patient identifiers were unavailable. E0 labels were used for training and primary selection, candidate search was extensive relative to validation size, and cross-validation was not nested. The factory-default comparator was a single run rather than a repeated-seed estimate. Confidence scores were not calibrated probabilities, and robustness tests were artificial. The consultations were informal, no diagnostic reader study was performed, and no external cohort, histologic reference, or adjudicated consensus was available.

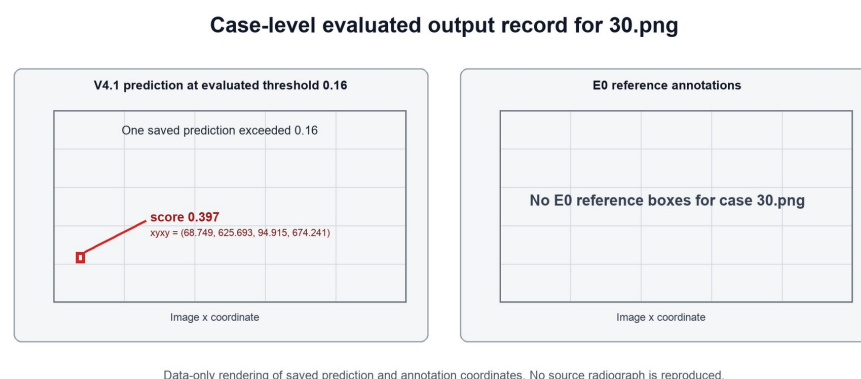


Figure 3: Data-only case record for 30.png at the evaluated threshold 0.16. The single saved V4.1 prediction (score 0.397) was unmatched because E0 supplied no reference box. Coordinates are shown without reproducing the source radiograph.

6 Conclusion

V4.1 improved recall and AP50 over both the documented V4 comparator and the fresh factory-default baseline, but its F1 was essentially tied with the default baseline and it produced more false-positive boxes. Results varied substantially with reader choice and operating threshold. The evidence supports continued research, not clinical use, and does not establish that clinician feedback caused the observed differences.

7 Future Work

Priorities are a larger patient-grouped cohort, a locked independent external test, multi-expert consensus or adjudication, repeated-seed comparator runs, nested model-selection assessment, score calibration, and a formal reader study. Visualization styles and insurance-documentation hypotheses should be tested separately from detector accuracy.

Acknowledgments

The author thanks Dr. Mark Y. Wang, DMD, DMSc, a prosthodontist practising in Concord, Massachusetts, for reviewing early prototype demonstrations and offering informal advice about missed or ambiguous caries findings. Dr. Wang did not annotate the evaluation set, participate in blinded reader testing, or clinically validate the system.

Ethics, Data, Funding, and Disclosures

Ethics: The project used public de-identified datasets; no new participants were recruited or imaged.

Data availability and attribution: the V4 and V4.1 analyses used Dental caries in bitewing radiographs, Version 1, contributed by Jan Kybic, Antonín Tichý, and Lukáš Kunt, Mendeley Data, DOI 10.17632/4fbdxs7s7w.1, licensed CC BY-NC 3.0. The fixed analysis split and derived outputs used in this study are described in the Supplement. Figure 3 is an original data-only visualization created for this study from saved model output; it does not reproduce source radiograph pixels. Its coordinates and E0 reference status derive from the licensed dataset.

Funding: none.

Conflict of interest: the author reports no financial interest in dental-AI companies mentioned.

AI use: OpenAI ChatGPT assisted with technical explanation, scripted analysis checks, organization, drafting, and language refinement. The author reviewed the materials and retains responsibility for accuracy and conclusions; no independent human statistician or external analyst performed a separate validation.

References

- [1] Nour Ammar and Jan Kühnisch. 2024. Diagnostic performance of artificial intelligence-aided caries detection on bitewing radiographs: a systematic review and meta-analysis. *Japanese Dental Science Review* 60 (2024), 128–136. <https://doi.org/10.1016/j.jdsr.2024.02.001>
- [2] Ibrahim Sevki Bayraktar, Kaan Orhan, Serdar Akarsu, Özer Çelik, Samet Atasoy, Adem Pekince, Yasin Yasa, Elif Bilgir, Hande Sağlam, Ahmet Faruk Aslan, and Alper Odabaş. 2022. Deep-learning approach for caries detection and segmentation on dental bitewing radiographs. *Oral Radiology* 38 (2022), 468–479. <https://doi.org/10.1007/s11282-021-00577-9>
- [3] Anselmo Garcia Cantu, Sascha Gehrun, Joachim Krois, Akhilanand Chaurasia, Jesus Gomez Rossi, Robert Gaudin, Karim Elhennawy, and Falk Schwendicke. 2020. Detecting caries lesions of different radiographic extension on bitewings using deep learning. *Journal of Dentistry* 100 (2020), 103425. <https://doi.org/10.1016/j.jdent.2020.103425>
- [4] Kristof Sebastian Hansson Horvath, Nils Roar Gjerdet, and Xie-Qi Shi. 2026. Advancements in caries diagnostics using bitewing radiography: a systematic review of deep learning approaches. *Caries Research* 60 (2026), 127–151. <https://doi.org/10.1159/000546448>
- [5] Lukáš Kunt, Jan Kybic, Valéria Nagyová, and Antonín Tichý. 2023. Automatic caries detection in bitewing radiographs: Part I – deep learning. *Clinical Oral Investigations* 27 (2023), 7463–7471. <https://doi.org/10.1007/s00784-023-05335-1>
- [6] Jan Kybic, Antonín Tichý, and Lukáš Kunt. 2023. Dental caries in bitewing radiographs. Mendeley Data, Version 1. <https://doi.org/10.17632/4fbdxs7s7w.1>
- [7] Antonín Tichý, Lukáš Kunt, Valéria Nagyová, and Jan Kybic. 2024. Automatic caries detection in bitewing radiographs – Part II: experimental comparison. *Clinical Oral Investigations* 28 (2024), 133. <https://doi.org/10.1007/s00784-024-05528-2>
- [8] Yifan Zhang, Fan Ye, Lingxiao Chen, Feng Xu, Xiaodiao Chen, Hongkun Wu, Mingguo Cao, Yunxiang Li, Yaqi Wang, and Xingru Huang. 2023. Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. *Scientific Data* 10 (2023), 380. <https://doi.org/10.1038/s41597-023-02237-5> Dataset: doi:10.6084/m9.figshare.21621705 (CC0 1.0).

Supplementary Methods and Results

Supporting material for: Iterative Development of an AI Prototype for Dental Caries Detection and Localization with Informal Clinical Feedback

S1 Formative Versions 1 to 3

Versions 1–3 used a 100-image subset of the Children’s Dental Panoramic Radiographs Dataset (Zhang et al., *Scientific Data*, 2023;10:380 [8]; article DOI: 10.1038/s41597-023-02237-5). The source dataset is hosted by Springer Nature Figshare (dataset DOI: 10.6084/m9.figshare.21621705) under the CC0 1.0 Universal Public Domain Dedication. Version 1 tested binary image classification; Version 2 expanded to six labels but several labels had very few positives; Version 3 attempted six-class localization and performed weakly. These experiments motivated narrowing to caries on permanent-dentition bitewings. Informal consultation with Dr. Mark Y. Wang, DMD, DMSc, a prosthodontist practising in Concord, Massachusetts, included review of early prototype demonstrations and discussion of missed or ambiguous findings; his input was design feedback rather than expert reference annotation, independent reader testing, or clinical validation. The author confirms that Dr. Wang consented to his name, credentials, and this description appearing in the manuscript and supplement. They do not enter the V4 versus V4.1 quantitative comparison. The V4 and V4.1 work used the Dental caries in bitewing radiographs, Version 1, contributed by Jan Kybic, Antonín Tichý, and Lukáš Kunt, Mendeley Data, DOI 10.17632/4fbdxs7s7w.1, under CC BY-NC 3.0. The release contains 100 bitewings with bounding-box annotations from eight dentists. Figure 3 in the main manuscript is an original data-only visualization from saved model output and does not reproduce radiograph pixels; the displayed coordinates and E0 reference status derive from this dataset.

S2 Reproducible Training Settings

Table S1: Training settings for each run.

Run	Model	Pixels	Epochs	Batch	LR	Optimizer	Pretraining
V4	YOLO11n	768	40	4	0.001	AdamW	COCO
V4.1 n768	YOLO11n	768	80	4	0.001	AdamW	COCO
V4.1 n1024	YOLO11n	1024	80	2	0.001	AdamW	COCO
V4.1 s768	YOLO11s	768	60	2	0.0005	AdamW	COCO
Factory default	YOLO11n	640	100	16	Auto	Auto to AdamW	COCO

V4 and V4.1 used seed 42, deterministic mode, no vertical flip, no mosaic, no mixup, no copy-paste, horizontal flip 0.5, rotation 5 degrees, translation 0.05, scale 0.20, and brightness augmentation 0.10. V4 used patience 12; V4.1 used patience 20. The fresh factory-default run used Ultralytics defaults, including seed 0, patience 100, mosaic 1.0, translation 0.10, scale 0.50, HSV augmentation, and automatic optimizer selection, which resolved to AdamW. Only the dataset path, CPU device, zero workers, and output path were specified. Software versions were Python 3.13.9, PyTorch 2.13.0, Ultralytics 8.4.128, NumPy 2.3.5, pandas 2.3.3, and scikit-learn 1.7.2.

S3 Validation Candidate Analysis

Table S2: Top ten validation candidates by lesion F1 (A, B, and C defined in the text below).

Rank	Candidate	F1	P	R	AP50
1	Fusion A + B (NMS .50)	0.695	0.647	0.750	0.651
2	Fusion A + B (NMS .30)	0.695	0.647	0.750	0.650
3	Fusion A + B (NMS .50 and nested .80)	0.695	0.647	0.750	0.650
4	A (NMS .30)	0.643	0.675	0.614	0.567
5	B (NMS .50 and nested .80)	0.636	0.636	0.636	0.606
6	A (NMS .50)	0.635	0.659	0.614	0.574
7	A (NMS .50 and nested .80)	0.635	0.659	0.614	0.563
8	B (NMS .50)	0.632	0.750	0.545	0.610
9	B (NMS .30)	0.632	0.750	0.545	0.609
10	C (NMS .50)	0.630	0.793	0.523	0.578

Display key: A is YOLO11n-768 with full-frame 1024-pixel inference; B is YOLO11n-1024 with horizontally flipped 1024-pixel inference; C is YOLO11n-768 with horizontally flipped 1024-pixel inference. Among 75 candidates, best F1 was 0.695 and median F1 was 0.556. Three candidates were within 0.05 F1 of the winner; all were final-suppression variants of the same A plus B fusion.

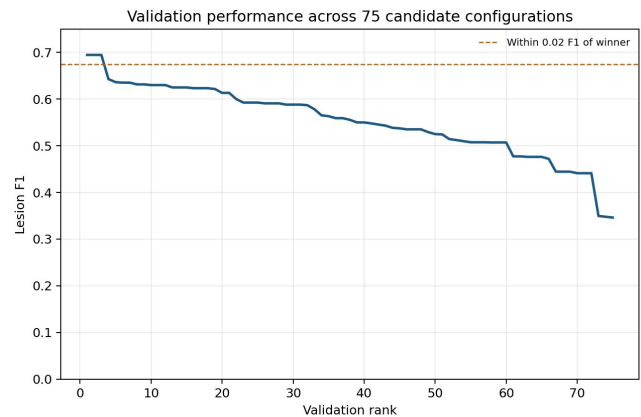


Figure S1: Full validation leaderboard.

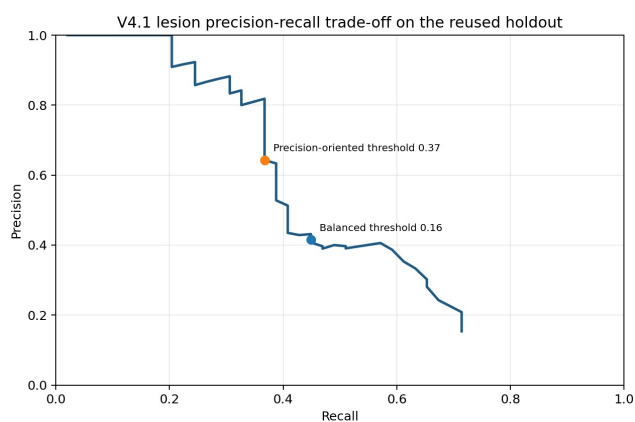
S4 Multi-reader Results

The eight independent references are not interchangeable ground truths. At threshold 0.16, 31 predictions were unmatched to E0; 13 overlapped at least one other reader, 9 at least two, 5 at least three, and 3 at least four.

Table S3: V4.1 rescored against each of the eight readers at thresholds 0.16 and 0.37.

Reader	Boxes	P .16	R .16	F1 .16	P .37	R .37	F1 .37	AP50
E0	49	0.415	0.449	0.431	0.643	0.367	0.468	0.491
E1	85	0.302	0.188	0.232	0.500	0.165	0.248	0.138
E2	50	0.283	0.300	0.291	0.464	0.260	0.333	0.251
E3	49	0.358	0.388	0.373	0.536	0.306	0.390	0.279
E4	47	0.132	0.149	0.140	0.179	0.106	0.133	0.042
N1	81	0.566	0.370	0.448	0.786	0.272	0.404	0.388
N2	67	0.321	0.254	0.283	0.464	0.194	0.274	0.172
N3	63	0.434	0.365	0.397	0.607	0.270	0.374	0.386

S5 Threshold Results

**Figure S2: Holdout precision-recall trade-off against E0.**

At 0.16: 22 TP, 31 FP, 27 FN, precision 0.415, recall 0.449, F1 0.431, and 1.55 FP/image. At 0.37: 18 TP, 10 FP, 31 FN, precision 0.643, recall 0.367, F1 0.468, and 0.50 FP/image. ROC AUC for the image-level maximum-score flag was 0.933; it is threshold-independent and based on 20 images.

S6 Conditional Three-fold Stress Test

The three folds covered the original 60 training images. At threshold 0.16, pooled out-of-fold detection yielded precision 0.636, recall 0.466, F1 0.538, 0.783 FP/image, AP50 0.510, and mAP50–95 0.188. This is not nested cross-validation because the recipe was informed by development on the same 60-image pool.

S7 Factory-default Comparator and Verification Status

The fresh factory-default YOLO11n run completed 100 epochs from the official generic checkpoint using unmodified training defaults. Validation selected threshold 0.35 before holdout evaluation. On the reused holdout it produced 15 TP, 5 FP, and 34 FN: precision 0.750, recall 0.306, F1 0.435, 0.25 FP/image, AP50 0.458, and mAP50–95 0.187. This single-run comparison is exploratory and does not correct the small reused holdout or missing patient identifiers. The available project package describes raw predictions, annotations,

evaluation scripts, training logs, cross-validation assignments, configuration files, and checkpoint hashes. The reported values were recomputed through scripted checks prepared with AI assistance and reviewed by the author. This was not an independent human statistical audit or external validation.