Additional file 1

Supplementary analyses for:
Multimodal Deep Learning for Automated Assessment of Periodontal Bone Resorption Severity Using Panoramic Radiographs and Structured Metadata

Adnan Salman and Ibtehal Raghib Bishtawi

This document provides supplementary analyses and material referenced in the main manuscript: the metadata ablation quantifying the contribution of age (Section S8), the comparison of late fusion with joint multimodal fusion (Section S6), full results of the exploratory probability-level ensemble (Section S7), per-class performance metrics and pairwise statistical comparisons for all models, an example of the manual dental-arch cropping (Section S2), representative training curves (Section S3), and Grad-CAM activation maps for all nine reference–prediction combinations (Section S10).

Contents

S1 Comparison with the Published Image-Only Baseline 1

S2 Dental-Arch Cropping 3

S3 Training Dynamics 4

S4 Complete Performance Metrics and Statistical Comparisons 5

S5 Within-Split Comparison with the Metadata-Only Model 7

S6 Fusion Strategy: Late Fusion versus Joint Multimodal Fusion 8

S7 Exploratory Probability-Level Ensemble 10

S8 Metadata Ablation and the Contribution of Age 11

S9 Representative-Split Classification Performance 16

S10 GradCAM Activation Maps for All Reference–Prediction Combinations 17

S1 Comparison with the Published Image-Only Baseline

For a descriptive comparison with the previously published image-only approach, the publicly available implementation of [1] was reproduced unmodified on the original full panoramic radiographs over 20 repeated random runs. Since that implementation generates its own data partitions, these runs are unpaired with those used for the present study's models and are reported descriptively (Table S1).

Table S1: Comparison with the image-only baseline of [1]. *Reported*: originally published single-split result. *Reproduced*: re-evaluated across 20 repeated random runs using the authors' implementation without modification on full panoramic radiographs. *Proposed*: our image-only model trained on cropped radiographs with fourth-root class weighting across 20 repeated stratified random splits. Because the two implementations generate their own data partitions, the reproduced and proposed results are not paired.

Metric	**Reported Single split**	**Reproduced 20 runs**	**Proposed 20 runs**
Accuracy	0.633	0.54 ± 0.06	0.62 ± 0.04
Macro F1	–	0.51 ± 0.05	0.59 ± 0.04
L1 Recall	0.448	0.55 ± 0.21	0.53 ± 0.10
L1 F1	0.510	0.46 ± 0.08	0.54 ± 0.09
L2 Recall	0.835	0.55 ± 0.17	0.68 ± 0.07
L2 F1	0.720	0.58 ± 0.11	0.68 ± 0.04
L3 Recall	0.348	0.51 ± 0.19	0.56 ± 0.09
L3 F1	0.450	0.48 ± 0.09	0.55 ± 0.04

S2 Dental-Arch Cropping

All 988 panoramic radiographs were manually cropped before model training to retain the dental arch and supporting alveolar bone while excluding peripheral craniofacial structures and image borders. Cropping was performed by the first author on anatomical landmarks alone, without reference to the BRAR severity labels.

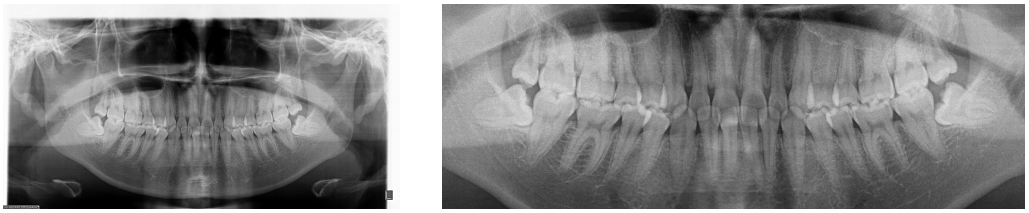


Figure S1: Representative panoramic radiograph before (left) and after (right) manual cropping. Images are shown at equal height, so the cropped radiograph appears at a larger scale.

S3 Training Dynamics

Figure S2 presents the training and validation loss and accuracy curves across the three training stages for a representative run. During Stage 1, with the convolutional backbone frozen, training and validation loss generally decreased, accompanied by an increase in both training and validation accuracy.

At the transition to Stage 2, when the final backbone blocks were unfrozen, training loss temporarily increased before continuing to decrease. Validation loss subsequently remained relatively stable, while training loss continued to decline. A corresponding train–validation gap became more apparent during this stage.

Following full unfreezing in Stage 3, validation loss and accuracy showed limited additional improvement, while training performance continued to fluctuate. Early stopping was applied within each training stage, and the checkpoint with the lowest validation loss across all stages was retained for test evaluation. Thus, the reported test performance was obtained from the validation-selected checkpoint rather than from the final training epoch.

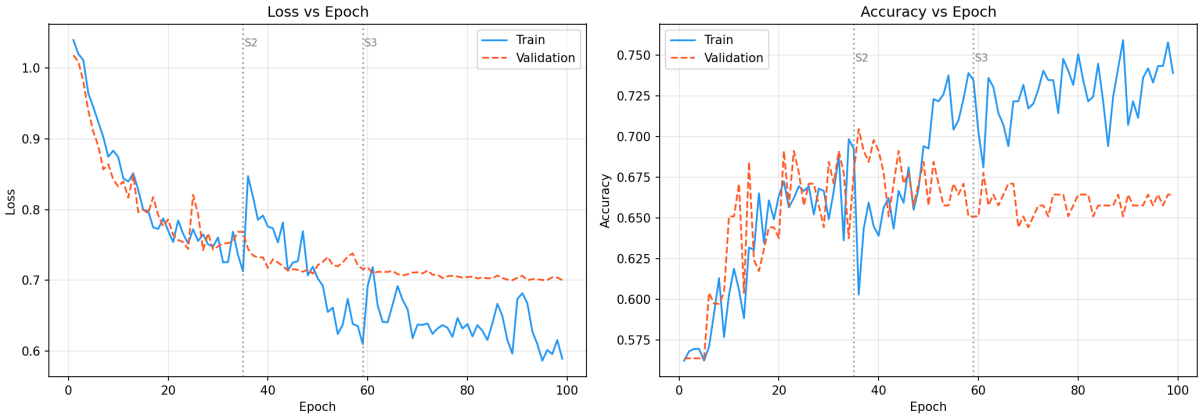


Figure S2: Training and validation loss (left) and accuracy (right) for a representative multimodal training run. Vertical dotted lines indicate transitions between the three training stages (S2 and S3), corresponding to partial and full unfreezing of the DenseNet-121 backbone, respectively.

S4 Complete Performance Metrics and Statistical Comparisons

This section provides the complete metric set underlying the summary results reported in the main manuscript. Table S2 reports all overall and per-class metrics for the image-only, metadata-only, multimodal, and exploratory ensemble models across the 20 repeated stratified random splits; the main manuscript reports accuracy, the three prespecified primary outcomes, and mean absolute grade error from this table. Table S3 reports pairwise Wilcoxon signed-rank comparisons for all metrics. Primary outcomes are shown with Holm–Bonferroni-adjusted p-values; secondary metrics are reported descriptively without multiplicity correction and should not be used to draw confirmatory conclusions. Paired effect estimates with 95% confidence intervals for the primary outcomes are shown in Fig. 4 of the main manuscript.

Table S2: Performance comparison across 20 repeated stratified random splits (mean $\pm$ standard deviation). The exploratory ensemble combines the multimodal, image-only, and metadata-only model probabilities using weights of 0.8, 0.1, and 0.1, respectively, selected by grid search on 10 separate tuning seeds not included among the 20 evaluation seeds. Metrics are not independent, and differences between models reflect trade-offs between class-wise sensitivity and overall ranking quality rather than uniform improvement; see the Discussion in the main manuscript.

Metric	Image-only	Meta-only	Multimodal	Ensemble
Accuracy	0.62 ± 0.04	0.62 ± 0.04	0.67 ± 0.03	0.68 ± 0.02
Macro F1	0.59 ± 0.04	0.51 ± 0.05	0.65 ± 0.03	0.65 ± 0.03
Macro Precision	0.60 ± 0.05	0.60 ± 0.05	0.66 ± 0.03	0.66 ± 0.03
Macro Recall	0.59 ± 0.04	0.53 ± 0.05	0.65 ± 0.04	0.64 ± 0.03
Macro AUC-ROC	0.77 ± 0.03	0.73 ± 0.04	0.81 ± 0.03	0.81 ± 0.03
Micro AUC-ROC	0.81 ± 0.02	0.80 ± 0.03	0.84 ± 0.02	0.85 ± 0.02
Macro AP	0.63 ± 0.04	0.60 ± 0.04	0.69 ± 0.04	0.69 ± 0.04
Micro AP	0.65 ± 0.03	0.65 ± 0.04	0.71 ± 0.04	0.72 ± 0.04
L1 Precision	0.56 ± 0.11	0.55 ± 0.09	0.68 ± 0.09	0.69 ± 0.09
L1 Recall	0.53 ± 0.10	0.59 ± 0.12	0.62 ± 0.10	0.62 ± 0.09
L1 F1-Score	0.54 ± 0.09	0.57 ± 0.10	0.64 ± 0.08	0.65 ± 0.07
L1 AP	0.59 ± 0.10	0.62 ± 0.07	0.70 ± 0.09	0.71 ± 0.09
L2 Precision	0.68 ± 0.04	0.64 ± 0.03	0.72 ± 0.03	0.72 ± 0.02
L2 Recall	0.68 ± 0.07	0.85 ± 0.05	0.72 ± 0.06	0.74 ± 0.06
L2 F1-Score	0.68 ± 0.04	0.73 ± 0.03	0.72 ± 0.03	0.73 ± 0.03
L2 AP	0.71 ± 0.03	0.73 ± 0.05	0.76 ± 0.05	0.77 ± 0.05
L3 Precision	0.56 ± 0.06	0.62 ± 0.16	0.57 ± 0.05	0.59 ± 0.05
L3 Recall	0.56 ± 0.09	0.14 ± 0.07	0.60 ± 0.07	0.57 ± 0.07
L3 F1-Score	0.55 ± 0.04	0.23 ± 0.09	0.58 ± 0.03	0.57 ± 0.04
L3 AP	0.57 ± 0.03	0.45 ± 0.06	0.60 ± 0.06	0.60 ± 0.05
MAGE	0.39 ± 0.04	0.41 ± 0.03	0.35 ± 0.03	0.34 ± 0.03
Errors > 1 grade	0.016 ± 0.013	0.023 ± 0.010	0.015 ± 0.012	0.012 ± 0.009

Table S3: Pairwise model comparisons using the Wilcoxon signed-rank test across 20 repeated stratified random splits. Primary metrics (above rule) are reported using Holm–Bonferroni-adjusted p-values. Secondary metrics (below rule) are reported descriptively using unadjusted p-values and should not be interpreted as multiplicity-corrected inferential tests. Values below 0.001 are reported as $p < 0.001$. † indicates a statistically significant difference favoring the second-named model. Paired effect estimates with 95% confidence intervals for the primary outcomes are shown in Fig. 4 of the main manuscript.

Metric	Multi vs Image	Multi vs Meta	Ensemble vs Multi
Primary metrics (Holm–Bonferroni adjusted)			
Macro F1	< 0.001	< 0.001	0.546
Macro AUC-ROC	< 0.001	< 0.001	0.080
L3 Recall	0.036	< 0.001	$< 0.001^{\dagger}$
Secondary metrics (descriptive; unadjusted)			
Accuracy	< 0.001	< 0.001	0.111
Macro Precision	< 0.001	< 0.001	0.105
Macro Recall	< 0.001	< 0.001	0.126
Micro AUC	< 0.001	< 0.001	< 0.001
Micro AP	< 0.001	< 0.001	< 0.001
Macro AP	< 0.001	< 0.001	0.189
L1 Precision	< 0.001	< 0.001	0.238
L1 Recall	< 0.001	0.369	0.657
L1 F1	< 0.001	< 0.001	0.269
L1 AP	< 0.001	< 0.001	0.231
L2 Precision	< 0.001	< 0.001	$0.005^{\dagger}$
L2 Recall	0.079	$< 0.001^{\dagger}$	< 0.001
L2 F1	0.001	0.641	0.035
L2 AP	< 0.001	< 0.001	0.005
L3 Precision	0.294	0.296	0.002
L3 F1	0.014	< 0.001	0.105
L3 AP	0.011	< 0.001	0.674

S5 Within-Split Comparison with the Metadata-Only Model

Figure S3 shows the within-split comparison of the multimodal and metadata-only models, corresponding to Fig. 3 of the main manuscript, which presents the same comparison against the image-only model. The multimodal model achieved higher values on all three primary outcomes in every one of the 20 splits.

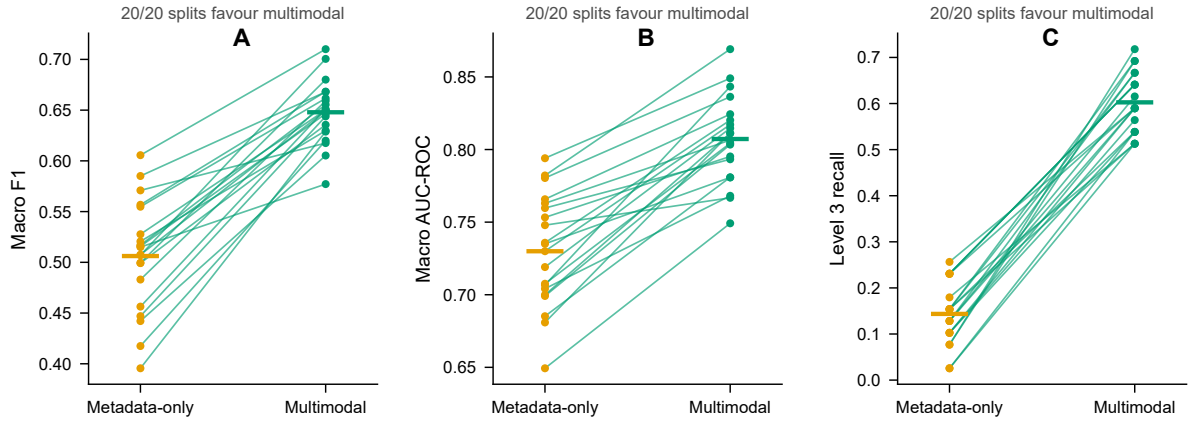


Figure S3: Within-split comparison of the multimodal and metadata-only models across the 20 repeated stratified random splits. **(A)** Macro F1-score. **(B)** Macro AUC-ROC. **(C)** Level 3 recall. Each line joins the two models evaluated on the same data partition; green indicates a higher value for the multimodal model on that split, vermillion a higher value for the metadata-only model, and grey an identical value. Horizontal bars show group means.

S6 Fusion Strategy: Late Fusion versus Joint Multimodal Fusion

To examine whether the observed multimodal benefit arises from learning a joint representation or simply from access to both information sources, a late-fusion variant was evaluated. In this configuration, the separately trained image-only and metadata-only models were combined by weighted averaging of their predicted class probabilities, without any shared representation. Both configurations therefore use identical input information and differ only in the stage at which the two modalities are combined.

Combination weights were selected by grid search in steps of 0.1 using 10 tuning seeds that were not included among the 20 evaluation seeds, with macro F1-score as the optimization criterion (Table S4). The selected weights were 0.7 for the image-only model and 0.3 for the metadata-only model. These weights were then fixed and applied across the 20 evaluation splits.

The grid search indicates that combining the two unimodal models does improve on either alone: the best combination reached a macro F1-score of 0.577, compared with 0.555 for the image-only model and 0.503 for the metadata-only model in isolation. The four best combinations differ by 0.010, indicating a relatively flat objective surface; the selected weights should therefore be regarded as one of several near-equivalent choices rather than a sharply defined optimum.

Table S4: Grid search for late-fusion combination weights on the 10 tuning seeds not included among the 20 evaluation seeds, ordered by macro F1-score. The selected combination is shown in the first row. The final two rows correspond to the unimodal models in isolation.

Image-only weight	Metadata-only weight	Macro F1
0.7	0.3	0.5769
0.6	0.4	0.5718
0.8	0.2	0.5689
0.9	0.1	0.5669
1.0	0.0	0.5546
0.5	0.5	0.5500
0.4	0.6	0.5265
0.3	0.7	0.5149
0.2	0.8	0.5127
0.1	0.9	0.5064
0.0	1.0	0.5029

Table S5 reports the full comparison across all evaluated metrics. Late fusion performed significantly worse than joint multimodal fusion on all three primary metrics after Holm–Bonferroni correction, with the largest difference in Level 3 recall (median paired difference -0.179; interquartile range $[-0.256, -0.147]$). Late fusion achieved higher Level 2 recall (0.83 versus 0.72) and higher Level 3 precision (0.65 versus 0.57), but substantially lower Level 3 recall (0.40 versus 0.60). These findings are consistent with an advantage of learning a joint multimodal representation over combining predictions from separately trained unimodal models.

Table S5: Late fusion versus joint multimodal fusion across the 20 repeated stratified random splits. Values are mean ± standard deviation. Median paired differences and interquartile ranges are computed across the 20 shared splits (late fusion minus multimodal); negative values favour joint fusion. Metrics marked * are the three prespecified primary outcomes, for which p-values are Holm–Bonferroni corrected across those three comparisons; all other p-values are uncorrected and descriptive.

Metric	Late fusion	Multimodal	Median diff.	IQR	p
Accuracy	0.67 ± 0.02	0.67 ± 0.03	-0.007	$[-0.020, +0.015]$	0.454
Macro F1*	0.61 ± 0.04	0.65 ± 0.03	-0.038	$[-0.066, -0.011]$	<0.001
Macro precision	0.66 ± 0.04	0.66 ± 0.03	$+0.007$	$[-0.015, +0.024]$	0.452
Macro recall	0.59 ± 0.04	0.65 ± 0.04	-0.061	$[-0.082, -0.027]$	<0.001
Micro AUC-ROC	0.84 ± 0.02	0.84 ± 0.02	-0.006	$[-0.017, +0.002]$	0.044
Macro AUC-ROC*	0.79 ± 0.03	0.81 ± 0.03	-0.015	$[-0.031, -0.003]$	<0.001
Micro AP	0.70 ± 0.03	0.71 ± 0.04	-0.012	$[-0.028, -0.002]$	0.007
Macro AP	0.67 ± 0.04	0.69 ± 0.04	-0.020	$[-0.041, -0.008]$	<0.001
L1 precision	0.66 ± 0.09	0.68 ± 0.09	-0.013	$[-0.076, +0.021]$	0.261
L1 recall	0.54 ± 0.09	0.62 ± 0.10	-0.077	$[-0.125, -0.038]$	<0.001
L1 F1	0.59 ± 0.07	0.64 ± 0.08	-0.045	$[-0.087, -0.023]$	<0.001
L1 AP	0.67 ± 0.10	0.70 ± 0.09	-0.033	$[-0.054, -0.012]$	<0.001
L2 precision	0.68 ± 0.03	0.72 ± 0.03	-0.052	$[-0.059, -0.021]$	<0.001
L2 recall	0.83 ± 0.05	0.72 ± 0.06	$+0.107$	$[+0.080, +0.137]$	<0.001
L2 F1	0.74 ± 0.02	0.72 ± 0.03	$+0.025$	$[+0.009, +0.036]$	<0.001
L2 AP	0.75 ± 0.03	0.76 ± 0.05	-0.016	$[-0.029, -0.003]$	0.003
L3 precision	0.65 ± 0.08	0.57 ± 0.05	$+0.068$	$[+0.035, +0.126]$	<0.001
L3 recall*	0.40 ± 0.11	0.60 ± 0.07	-0.179	$[-0.256, -0.147]$	<0.001
L3 F1	0.49 ± 0.10	0.58 ± 0.03	-0.073	$[-0.107, -0.042]$	<0.001
L3 AP	0.58 ± 0.05	0.60 ± 0.06	-0.014	$[-0.054, +0.014]$	0.105

S7 Exploratory Probability-Level Ensemble

Table S6 reports the full comparison between the exploratory probability-level ensemble and the single multimodal model across the same 20 splits. Ensemble weights (0.8 multimodal, 0.1 image-only, 0.1 metadata-only) were selected by grid search on the 10 tuning seeds that were not included among the 20 evaluation seeds using macro F1-score, as described in the main manuscript.

The ensemble did not differ significantly from the multimodal model on macro F1 ($p = 0.55$) or macro AUC-ROC ($p = 0.08$ after correction), and showed significantly lower Level 3 recall (median paired difference -0.026; interquartile range $[-0.051, -0.026]$). Significant differences favouring the ensemble were confined to micro-averaged ranking metrics and to Level 2, with median differences between 0.003 and 0.024.

Level 1 predictions were identical in 13 of the 20 splits (median paired difference and interquartile range both zero), consistent with the ensemble being dominated by its multimodal component at a weight of 0.8. The single multimodal model was therefore retained as the principal model in the main manuscript.

Table S6: Exploratory probability-level ensemble versus the single multimodal model across the 20 repeated stratified random splits. Values are mean ± standard deviation. Median paired differences and interquartile ranges are computed across the 20 shared splits (ensemble minus multimodal); positive values favour the ensemble. Metrics marked * are the three primary outcomes, for which p-values are Holm–Bonferroni corrected across those three comparisons; all other p-values are uncorrected and descriptive.

Metric	Ensemble	Multimodal	Median diff.	IQR	p
Accuracy	0.68 ± 0.02	0.67 ± 0.03	+0.007	[−0.002, +0.013]	0.111
Macro F1*	0.65 ± 0.03	0.65 ± 0.03	+0.001	[−0.004, +0.007]	0.546
Macro precision	0.66 ± 0.03	0.66 ± 0.03	+0.005	[−0.002, +0.012]	0.105
Macro recall	0.64 ± 0.03	0.65 ± 0.04	−0.003	[−0.005, +0.000]	0.126
Micro AUC-ROC	0.85 ± 0.02	0.84 ± 0.02	+0.004	[+0.002, +0.005]	<0.001
Macro AUC-ROC*	0.81 ± 0.03	0.81 ± 0.03	+0.001	[−0.000, +0.003]	0.080
Micro AP	0.72 ± 0.04	0.71 ± 0.04	+0.006	[+0.005, +0.009]	<0.001
Macro AP	0.69 ± 0.04	0.69 ± 0.04	+0.001	[−0.001, +0.006]	0.189
L1 precision	0.69 ± 0.09	0.68 ± 0.09	+0.000	[−0.002, +0.024]	0.238
L1 recall	0.62 ± 0.09	0.62 ± 0.10	+0.000	[+0.000, +0.000]	0.657
L1 F1	0.65 ± 0.07	0.64 ± 0.08	+0.000	[−0.003, +0.017]	0.269
L1 AP	0.71 ± 0.09	0.70 ± 0.09	+0.001	[−0.002, +0.007]	0.231
L2 precision	0.72 ± 0.02	0.72 ± 0.03	−0.006	[−0.015, −0.002]	0.005
L2 recall	0.74 ± 0.06	0.72 ± 0.06	+0.024	[+0.012, +0.048]	<0.001
L2 F1	0.73 ± 0.03	0.72 ± 0.03	+0.011	[−0.005, +0.016]	0.035
L2 AP	0.77 ± 0.05	0.76 ± 0.05	+0.003	[+0.000, +0.005]	0.005
L3 precision	0.59 ± 0.05	0.57 ± 0.05	+0.020	[+0.007, +0.034]	0.002
L3 recall*	0.57 ± 0.07	0.60 ± 0.07	−0.026	[−0.051, −0.026]	<0.001
L3 F1	0.57 ± 0.04	0.58 ± 0.03	−0.005	[−0.010, +0.003]	0.105
L3 AP	0.60 ± 0.05	0.60 ± 0.06	+0.001	[−0.004, +0.006]	0.674

S8 Metadata Ablation and the Contribution of Age

Because age is incorporated directly into the BRAR target definition and is also supplied to the multimodal model as an input feature, an ablation of the metadata inputs was performed using four configurations: image-only, image plus age (age as the sole metadata variable), image plus the five non-age structured variables, and the full multimodal model. All configurations were trained and evaluated using the same 20 repeated stratified random splits, the same data partitions, and the same training procedure. Table S7 summarizes absolute performance for the four configurations.

Table S7: Metadata ablation across the 20 repeated stratified random splits. Values are mean [95% confidence interval]. Image + age uses age as the only metadata variable; image + non-age uses the remaining five structured metadata variables. Paired differences between these configurations are shown in Fig. 3 of the main manuscript; full pairwise comparisons are given in Tables S8–S11 below.

Metric	Image-only	Image + age	Image + non-age	Full multimodal
Accuracy	0.62 [0.61, 0.64]	0.64 [0.62, 0.66]	0.64 [0.62, 0.66]	0.67 [0.66, 0.68]
Macro F1	0.59 [0.57, 0.61]	0.61 [0.59, 0.63]	0.61 [0.58, 0.63]	0.65 [0.63, 0.66]
Macro AUC-ROC	0.77 [0.76, 0.78]	0.79 [0.77, 0.80]	0.79 [0.77, 0.80]	0.81 [0.79, 0.82]
Level 3 recall	0.56 [0.52, 0.60]	0.55 [0.51, 0.59]	0.53 [0.48, 0.58]	0.60 [0.57, 0.63]

Four pairwise comparisons are reported. Table S8 compares the full multimodal model with the no-age variant, Table S9 compares the full model with the image-plus-age configuration, and Tables S10 and S11 compare each single-subset configuration with the image-only model.

Removing either metadata subset from the full model reduced performance significantly on all three primary outcomes after Holm–Bonferroni correction. Relative to the no-age variant, the full model improved macro F1 by a median of $+0.040$ ($p < 0.001$), macro AUC-ROC by $+0.019$ ($p < 0.001$), and Level 3 recall by $+0.051$ ($p = 0.009$). Relative to the image-plus-age configuration, the corresponding improvements were $+0.024$ ($p < 0.001$), $+0.016$ ($p = 0.003$), and $+0.051$ ($p = 0.015$).

Neither subset alone, however, differed significantly from the image-only model on macro F1 or Level 3 recall. For the image-plus-age configuration the median paired differences were $+0.017$ ($p = 0.307$) and -0.013 ($p = 0.627$); for the image-plus-non-age configuration they were $+0.013$ ($p = 0.330$) and -0.013 ($p = 0.432$). Both configurations improved only macro AUC-ROC ($+0.026$, $p = 0.011$ and $+0.018$, $p = 0.005$, respectively).

These results indicate that age contributes to BRAR classification but does not independently account for the performance advantage of the full multimodal model. The image-plus-age and image-plus-non-age configurations showed similar performance, and neither significantly improved macro F1 or Level 3 recall relative to the image-only model, although both significantly improved macro AUC-ROC. In contrast, the full multimodal model significantly outperformed the image-plus-age configuration on all three primary outcomes, while removal of age from the full model also significantly reduced all three outcomes. This pattern is consistent with complementary contributions from age and the remaining structured metadata rather than the multimodal improvement being explained by age alone. Because age is mathematically incorporated into the BRAR definition, however, this ablation does not eliminate the inherent relationship between age and the prediction target.

Table S8: Full multimodal model versus the no-age multimodal variant across the 20 repeated stratified random splits. Values are mean ± standard deviation. Median paired differences and interquartile ranges are computed across the 20 shared splits (full minus no-age); positive values favour the full model. Metrics marked * are the three primary outcomes, for which p-values are Holm–Bonferroni corrected across those three comparisons; all other p-values are uncorrected and descriptive.

Metric	Full	No age	Median diff.	IQR	p
Accuracy	0.67 ± 0.03	0.64 ± 0.04	+0.027	[+0.000, +0.059]	0.003
Macro F1*	0.65 ± 0.03	0.60 ± 0.05	+0.040	[+0.013, +0.055]	<0.001
Macro precision	0.66 ± 0.03	0.62 ± 0.04	+0.042	[+0.002, +0.072]	0.002
Macro recall	0.65 ± 0.04	0.60 ± 0.05	+0.038	[+0.018, +0.063]	<0.001
Micro AUC-ROC	0.84 ± 0.02	0.83 ± 0.02	+0.013	[+0.006, +0.024]	<0.001
Macro AUC-ROC*	0.81 ± 0.03	0.79 ± 0.03	+0.019	[+0.008, +0.026]	<0.001
Micro AP	0.71 ± 0.04	0.68 ± 0.03	+0.032	[+0.015, +0.039]	<0.001
Macro AP	0.69 ± 0.04	0.65 ± 0.04	+0.029	[+0.016, +0.047]	<0.001
L1 precision	0.68 ± 0.09	0.63 ± 0.11	+0.067	[−0.026, +0.145]	0.058
L1 recall	0.62 ± 0.10	0.55 ± 0.13	+0.077	[+0.029, +0.125]	0.002
L1 F1	0.64 ± 0.08	0.58 ± 0.11	+0.061	[+0.039, +0.120]	<0.001
L1 AP	0.70 ± 0.09	0.65 ± 0.10	+0.046	[+0.028, +0.085]	<0.001
L2 precision	0.72 ± 0.03	0.69 ± 0.04	+0.039	[+0.008, +0.059]	<0.001
L2 recall	0.72 ± 0.06	0.72 ± 0.07	−0.012	[−0.048, +0.051]	0.823
L2 F1	0.72 ± 0.03	0.70 ± 0.04	+0.018	[−0.006, +0.045]	0.033
L2 AP	0.76 ± 0.05	0.73 ± 0.04	+0.030	[+0.007, +0.049]	<0.001
L3 precision	0.57 ± 0.05	0.55 ± 0.05	+0.012	[−0.016, +0.039]	0.294
L3 recall*	0.60 ± 0.07	0.53 ± 0.10	+0.051	[+0.000, +0.115]	0.009
L3 F1	0.58 ± 0.03	0.54 ± 0.07	+0.030	[+0.005, +0.071]	0.015
L3 AP	0.60 ± 0.06	0.58 ± 0.06	+0.021	[−0.006, +0.050]	0.040

Table S9: Full multimodal model versus the image-plus-age configuration, in which age is the only structured metadata variable, across the 20 repeated stratified random splits. Values are mean ± standard deviation. Median paired differences and interquartile ranges are computed across the 20 shared splits (full minus image-plus-age); positive values favour the full model. Metrics marked * are the three primary outcomes, for which p-values are Holm–Bonferroni corrected across those three comparisons; all other p-values are uncorrected and descriptive.

Metric	Full	Image + age	Median diff.	IQR	p
Accuracy	0.67 ± 0.03	0.64 ± 0.04	+0.017	[+0.000, +0.075]	0.006
Macro F1*	0.65 ± 0.03	0.61 ± 0.04	+0.024	[+0.015, +0.081]	<0.001
Macro precision	0.66 ± 0.03	0.61 ± 0.05	+0.030	[+0.006, +0.094]	<0.001
Macro recall	0.65 ± 0.04	0.61 ± 0.04	+0.030	[+0.011, +0.070]	0.002
Micro AUC-ROC	0.84 ± 0.02	0.83 ± 0.03	+0.015	[+0.003, +0.027]	0.001
Macro AUC-ROC*	0.81 ± 0.03	0.79 ± 0.03	+0.016	[+0.007, +0.033]	0.003
Micro AP	0.71 ± 0.04	0.68 ± 0.04	+0.027	[+0.004, +0.046]	0.001
Macro AP	0.69 ± 0.04	0.66 ± 0.05	+0.032	[+0.012, +0.053]	0.002
L1 precision	0.68 ± 0.09	0.58 ± 0.08	+0.076	[+0.028, +0.160]	<0.001
L1 recall	0.62 ± 0.10	0.57 ± 0.08	+0.019	[−0.010, +0.086]	0.116
L1 F1	0.64 ± 0.08	0.58 ± 0.07	+0.071	[+0.027, +0.123]	0.001
L1 AP	0.70 ± 0.09	0.63 ± 0.10	+0.086	[+0.044, +0.109]	<0.001
L2 precision	0.72 ± 0.03	0.70 ± 0.03	+0.030	[+0.007, +0.046]	0.004
L2 recall	0.72 ± 0.06	0.71 ± 0.05	+0.036	[−0.036, +0.051]	0.389
L2 F1	0.72 ± 0.03	0.70 ± 0.03	+0.017	[−0.001, +0.041]	0.035
L2 AP	0.76 ± 0.05	0.74 ± 0.05	+0.029	[+0.007, +0.035]	0.003
L3 precision	0.57 ± 0.05	0.56 ± 0.07	−0.002	[−0.039, +0.068]	0.546
L3 recall*	0.60 ± 0.07	0.55 ± 0.09	+0.051	[+0.000, +0.103]	0.015
L3 F1	0.58 ± 0.03	0.55 ± 0.06	+0.017	[−0.011, +0.078]	0.070
L3 AP	0.60 ± 0.06	0.61 ± 0.05	−0.009	[−0.039, +0.047]	0.869

Table S10: No-age multimodal variant versus the image-only model across the 20 repeated stratified random splits. Values are mean ± standard deviation. Median paired differences and interquartile ranges are computed across the 20 shared splits (no-age minus image-only); positive values favour the no-age multimodal model. Metrics marked * are the three primary outcomes, for which p-values are Holm–Bonferroni corrected across those three comparisons; all other p-values are uncorrected and descriptive.

Metric	No age	Image-only	Median diff.	IQR	p
Accuracy	0.64 ± 0.04	0.62 ± 0.04	+0.007	[−0.002, +0.040]	0.096
Macro F1*	0.60 ± 0.05	0.59 ± 0.04	+0.013	[−0.015, +0.037]	0.330
Macro precision	0.62 ± 0.04	0.60 ± 0.05	+0.019	[−0.010, +0.058]	0.076
Macro recall	0.60 ± 0.05	0.59 ± 0.04	+0.007	[−0.021, +0.029]	0.368
Micro AUC-ROC	0.83 ± 0.02	0.81 ± 0.02	+0.016	[+0.004, +0.035]	<0.001
Macro AUC-ROC*	0.79 ± 0.03	0.77 ± 0.03	+0.018	[+0.001, +0.036]	0.005
Micro AP	0.68 ± 0.03	0.65 ± 0.03	+0.032	[+0.006, +0.049]	0.001
Macro AP	0.65 ± 0.04	0.63 ± 0.04	+0.025	[−0.005, +0.056]	0.004
L1 precision	0.63 ± 0.11	0.56 ± 0.11	+0.082	[−0.018, +0.115]	0.014
L1 recall	0.55 ± 0.13	0.53 ± 0.10	+0.000	[−0.038, +0.086]	0.454
L1 F1	0.58 ± 0.11	0.54 ± 0.09	+0.032	[−0.015, +0.081]	0.036
L1 AP	0.65 ± 0.10	0.59 ± 0.10	+0.057	[+0.027, +0.103]	0.001
L2 precision	0.69 ± 0.04	0.68 ± 0.04	+0.002	[−0.024, +0.028]	0.648
L2 recall	0.72 ± 0.07	0.68 ± 0.07	+0.012	[−0.027, +0.083]	0.184
L2 F1	0.70 ± 0.04	0.68 ± 0.04	+0.005	[−0.014, +0.054]	0.202
L2 AP	0.73 ± 0.04	0.71 ± 0.03	+0.028	[+0.010, +0.042]	0.002
L3 precision	0.55 ± 0.05	0.56 ± 0.06	+0.006	[−0.034, +0.043]	0.701
L3 recall*	0.53 ± 0.10	0.56 ± 0.09	−0.013	[−0.058, +0.051]	0.432
L3 F1	0.54 ± 0.07	0.55 ± 0.04	−0.016	[−0.048, +0.030]	0.498
L3 AP	0.58 ± 0.06	0.57 ± 0.03	−0.000	[−0.035, +0.033]	0.985

Table S11: Image-plus-age configuration versus the image-only model across the 20 repeated stratified random splits. Values are mean ± standard deviation. Median paired differences and interquartile ranges are computed across the 20 shared splits (image-plus-age minus image-only); positive values favour the image-plus-age configuration. Metrics marked * are the three primary outcomes, for which p-values are Holm–Bonferroni corrected across those three comparisons; all other p-values are uncorrected and descriptive.

Metric	Image + age	Image-only	Median diff.	IQR	p
Accuracy	0.64 ± 0.04	0.62 ± 0.04	$+0.020$	$[-0.022, +0.062]$	0.179
Macro F1*	0.61 ± 0.04	0.59 ± 0.04	$+0.017$	$[-0.016, +0.053]$	0.307
Macro precision	0.61 ± 0.05	0.60 ± 0.05	$+0.023$	$[-0.023, +0.053]$	0.261
Macro recall	0.61 ± 0.04	0.59 ± 0.04	$+0.016$	$[-0.018, +0.043]$	0.133
Micro AUC-ROC	0.83 ± 0.03	0.81 ± 0.02	$+0.021$	$[+0.002, +0.034]$	0.004
Macro AUC-ROC*	0.79 ± 0.03	0.77 ± 0.03	$+0.026$	$[-0.001, +0.040]$	0.011
Micro AP	0.68 ± 0.04	0.65 ± 0.03	$+0.036$	$[+0.003, +0.063]$	0.002
Macro AP	0.66 ± 0.05	0.63 ± 0.04	$+0.040$	$[-0.001, +0.059]$	0.008
L1 precision	0.58 ± 0.08	0.56 ± 0.11	$+0.028$	$[-0.031, +0.080]$	0.173
L1 recall	0.57 ± 0.08	0.53 ± 0.10	$+0.077$	$[+0.000, +0.077]$	0.056
L1 F1	0.58 ± 0.07	0.54 ± 0.09	$+0.037$	$[+0.002, +0.078]$	0.040
L1 AP	0.63 ± 0.10	0.59 ± 0.10	$+0.035$	$[+0.001, +0.075]$	0.014
L2 precision	0.70 ± 0.03	0.68 ± 0.04	$+0.006$	$[-0.005, +0.050]$	0.113
L2 recall	0.71 ± 0.05	0.68 ± 0.07	$+0.042$	$[-0.039, +0.086]$	0.179
L2 F1	0.70 ± 0.03	0.68 ± 0.04	$+0.021$	$[-0.024, +0.071]$	0.131
L2 AP	0.74 ± 0.05	0.71 ± 0.03	$+0.019$	$[+0.008, +0.059]$	0.001
L3 precision	0.56 ± 0.07	0.56 ± 0.06	$+0.004$	$[-0.028, +0.062]$	0.490
L3 recall*	0.55 ± 0.09	0.56 ± 0.09	-0.013	$[-0.077, +0.051]$	0.627
L3 F1	0.55 ± 0.06	0.55 ± 0.04	-0.010	$[-0.046, +0.034]$	0.869
L3 AP	0.61 ± 0.05	0.57 ± 0.03	$+0.032$	$[-0.007, +0.055]$	0.070

S9 Representative-Split Classification Performance

This section provides the receiver operating characteristic curves, precision–recall curves, and confusion matrix for the multimodal model on a single representative evaluation split. These illustrate model behaviour on one partition; performance across all 20 splits is reported in the main manuscript.

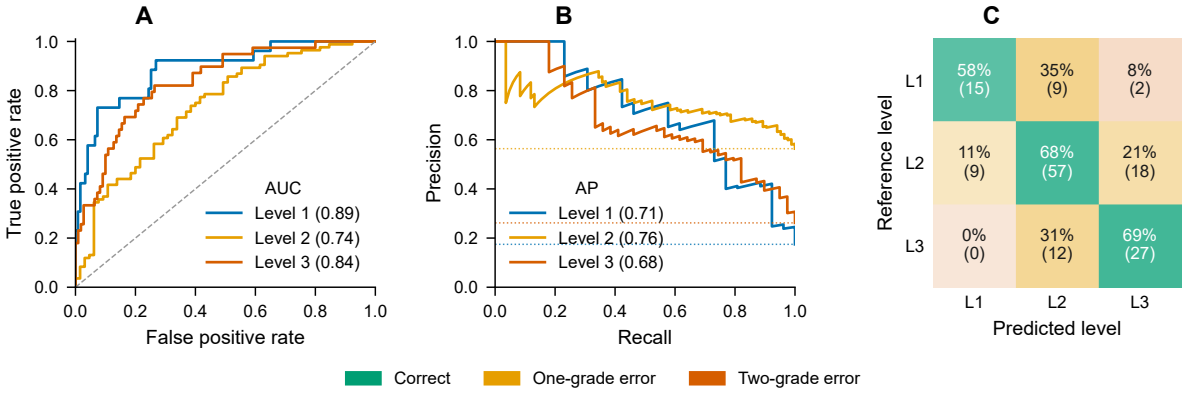


Figure S4: Classification performance of the multimodal model on a representative evaluation split (seed 2024). **(A)** Receiver operating characteristic curves for each severity level, with per-class area under the curve; the dashed diagonal indicates chance performance. **(B)** Precision–recall curves with per-class average precision; dotted horizontal lines indicate class prevalence in the test set, which is the expected precision of a random classifier. **(C)** Confusion matrix showing the percentage of cases in each reference level assigned to each predicted level, with counts in parentheses; rows sum to 100% and diagonal values correspond to per-class recall. Percentages may not sum to exactly 100 due to rounding. Cell colour indicates the ordinal distance between the reference and predicted grades — correct classification, one-grade error, or two-grade error — rather than frequency magnitude. Results across all 20 repeated stratified random splits are reported in the main manuscript.

S10 GradCAM Activation Maps for All Reference–Prediction Combinations

This section provides the complete set of GradCAM activation maps covering all nine combinations of reference and predicted severity grade; the main manuscript shows three of these. Each case was selected as the highest-confidence example in its reference–prediction category from predictions pooled across the 20 repeated stratified random splits, and the maps should be interpreted as qualitative visualizations of model activation rather than as evidence of anatomical localization.

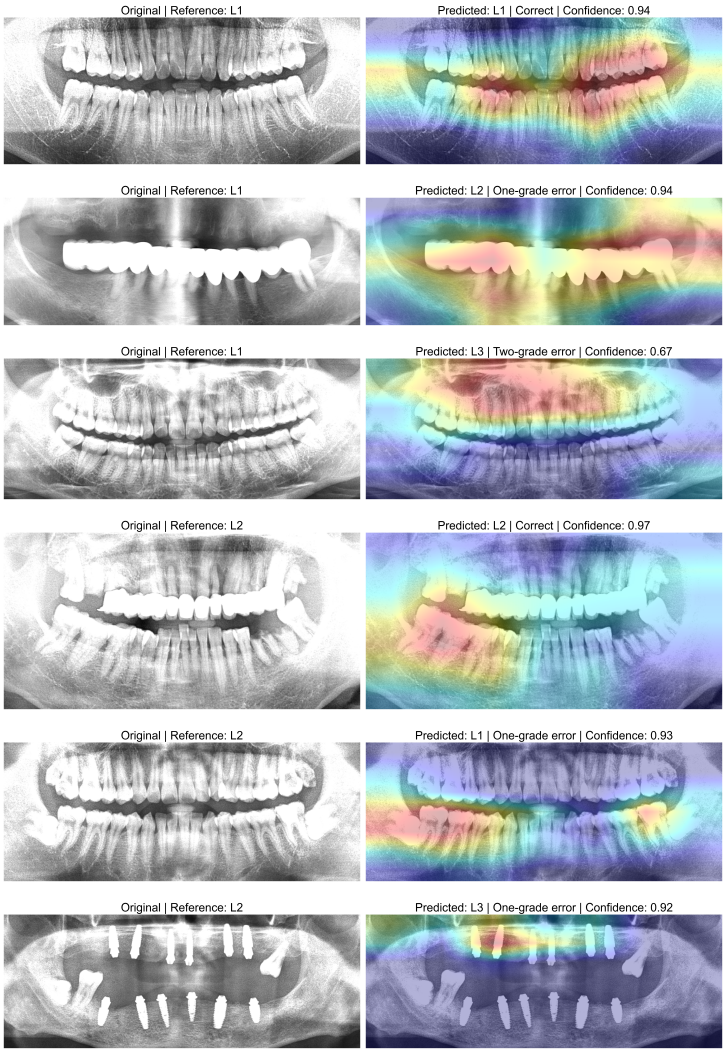


Figure S5: GradCAM activation maps for all nine reference–prediction combinations, ordered by reference grade (Levels 1–3) and, within each grade, by correct classification followed by misclassifications. Each row shows the cropped panoramic radiograph (left) alongside the activation overlay (right); the reference grade, predicted grade, and model confidence in the predicted class are indicated above each panel. Maps are normalized independently for each image, so colour intensity is comparable within but not across panels.

References

[1] Yiru Xia, Zhi Li, Zhikai Lin, Shuai Wang, Yaqi Wang, and Yufeng Xie. BRAR-anchored multimodal dataset of panoramic radiographs for periodontal bone resorption grading. *Scientific Data*, 13(1):89, December 2025.