

Multimodal deep learning for automated assessment of periodontal bone resorption severity using panoramic radiographs and structured metadata

Adnan Salman · Ibtehal Raghieb Bishtawi

Received: 18 August 2026 / Accepted: 9 October 2026

© The Author(s) 2026

Abstract

Purpose Assessment of periodontal bone resorption severity from panoramic radiographs can be time-consuming and subject to inter-observer variability. This study evaluates a multimodal deep learning framework for automated Bone Resorption Age Ratio (BRAR)-based severity classification by integrating panoramic radiographic features with structured metadata.

Methods Using 988 annotated panoramic radiographs from a publicly available dataset, image-only, metadata-only, and multimodal models were evaluated across 20 repeated stratified random splits. The multimodal model integrates DenseNet-121 radiographic features with a multilayer perceptron encoding structured metadata through feature-level concatenation and progressive fine-tuning. Performance was assessed using accuracy, macro F1-score, macro area under the receiver operating characteristic curve (macro AUC-ROC), average precision, and class-specific metrics. Paired comparisons of three primary outcomes (macro F1, macro AUC-ROC, and Level 3 recall) used the Wilcoxon signed-rank test with Holm–Bonferroni correction. A four-way metadata ablation assessing the contributions of age and the remaining structured variables, an exploratory probability-level ensemble, qualitative Gradient-weighted Class Activation Mapping (GradCAM), and a targeted expert error analysis on 88 purposively selected cases were also performed.

Results The multimodal model achieved an accuracy of 0.67 ± 0.03 , macro F1-score of 0.65 ± 0.03 , macro AUC-ROC of 0.81 ± 0.03 , and Level 3 recall of 0.60 ± 0.07 . It outperformed the image-only model in macro F1 and macro AUC-ROC (both Holm-adjusted $p < 0.001$) and Level 3 recall ($p = 0.036$), with all three primary outcomes also significantly higher than the metadata-only model ($p < 0.001$), which identified only 0.14 ± 0.07 of severe cases. In an ablation of the metadata inputs, neither age alone nor the five non-age variables alone significantly improved macro F1 or Level 3 recall over the image-only model, whereas the full model significantly outperformed both. In the targeted expert subset, dentist–original agreement was higher among model-correct cases ($\kappa = 0.585$; weighted $\kappa = 0.650$) than among model-misclassified cases ($\kappa = 0.208$; weighted $\kappa = 0.248$).

Conclusions Integrating panoramic radiographs with structured metadata improved prediction of BRAR-derived periodontal severity labels compared with either modality alone. The metadata ablation showed that age alone did not account for the multimodal improvement, with the findings supporting complementary contributions from age and the remaining structured metadata. The targeted expert analysis showed greater dentist–annotation

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



disagreement among selected model-misclassified cases but, because it involved one dentist and a purposively selected subset, does not establish annotation uncertainty as the cause of model errors. External validation, automated localization, and evaluation of direct radiographic bone-resorption estimation followed by age normalization are important directions for future work.

Keywords Multimodal deep learning · Panoramic radiographs · Periodontal disease · Bone resorption · Medical image analysis · Grad-CAM

Introduction

Periodontal disease is a prevalent chronic inflammatory condition affecting over a billion people worldwide [1] and a leading cause of tooth loss [2–4]. Accurate assessment of alveolar bone resorption—the progressive loss of bone supporting the teeth—is central to diagnosis and treatment. If left undetected, advanced bone resorption leads to irreversible tooth loss, functional impairment, and increased healthcare burden [3, 4].

Current clinical assessment relies on interpretation of panoramic radiographs (orthopantomograms, OPGs) by trained clinicians, including estimation of alveolar bone loss in relation to dental structures and consideration of patient-specific factors. Such assessment is subject to inter-observer variability [5]. Limited access to periodontal expertise, particularly in primary-care and resource-constrained settings, motivates the development of computational methods that may support more standardized periodontal assessment [6].

Artificial intelligence (AI), particularly deep learning, has demonstrated promising performance in medical image analysis, including diabetic retinopathy screening [7], chest radiograph interpretation, and skin lesion classification [8]. In dentistry, convolutional neural networks (CNNs) have been applied to tasks including caries detection, tooth segmentation and dental arch detection [9], and implant identification [10]. However, comparatively few studies have investigated periodontal bone-resorption severity classification using both radiographic and structured patient information.

The recently introduced BRAR-anchored multimodal dataset [11] provides a publicly available collection of 988 panoramic radiographs with associated structured metadata and severity labels derived from the Bone Resorption Age Ratio (BRAR). BRAR is an age-adjusted measure based on radiographic bone resorption and patient age, and the resulting labels categorize cases into Level 1 (mild), Level 2 (moderate), and Level 3 (severe). These labels therefore represent BRAR-based severity categories rather than a comprehensive independent clinical diagnosis of periodontal disease. The original study using this dataset reported an image-only EfficientNet-B0 baseline with an accuracy of 63.3% on a single data split and comparatively limited sensitivity for Level 3 cases (recall=0.348) [11]. The availability of associated structured metadata motivates investigation of whether combining radiographic and non-image information can improve BRAR-based classification.

In this study, we evaluate a multimodal deep learning framework that combines radiographic features extracted from panoramic images with structured metadata through feature-level fusion. The framework predicts BRAR-based severity categories without requiring explicit manual measurement of bone loss, although the current pre-processing pipeline includes manual dental-arch cropping. Performance is evaluated across 20 repeated stratified random splits using identical data partitions for paired model comparisons. Primary comparisons focus on three primary performance measures with multiplicity adjustment, while repeated splits are used principally to characterize performance variability rather than as independent experimental replications.

The objectives of this study are to: (i) develop and evaluate a multimodal fusion architecture combining panoramic radiographs with structured metadata; (ii) compare image-only, metadata-only, and multimodal formulations to characterize the contribution of modality fusion; and (iii) investigate model failure patterns using targeted independent expert re-evaluation and Grad-CAM visualization.

The main contributions of this work are as follows:

1. We systematically evaluate feature-level fusion of panoramic radiographic representations and structured metadata for BRAR-based periodontal bone resorption severity classification, comparing image-only, meta-data-only, and multimodal models under a common experimental protocol.
2. We evaluate model performance across 20 repeated stratified random splits using identical partitions for paired model comparisons, with primary performance metrics, confidence intervals, and multiplicity-adjusted statistical testing to provide more robust performance estimates than a single-split evaluation.
3. We investigate the influence of age, which contributes to BRAR label construction, through a four-way ablation comparing image-only, image plus age, image plus non-age metadata, and the full multimodal model, assessing whether age alone accounts for the observed multimodal improvement.
4. We complement quantitative evaluation with a targeted expert error analysis and qualitative Grad-CAM visualization to explore annotation disagreement and plausible image regions associated with model predictions, while recognizing the purposefully selected nature of the expert subset and the absence of localization ground truth.

Related work

Deep learning for dental radiograph analysis

Panoramic radiographs are widely used in dental practice, and deep learning has increasingly been applied to support their automated interpretation. Convolutional neural networks have demonstrated promising performance in tooth detection and numbering [12], caries detection [13], and tooth segmentation and implant classification [14]. More recently, attention-based architectures and transformer models have been explored for dental landmark detection and pathology recognition [15].

For periodontal assessment specifically, several studies have investigated automated bone-level estimation from periapical and panoramic radiographs. Jundaeng et al. [16] proposed a CNN-based approach for detecting alveolar bone-loss patterns, while Putra et al. [17] applied object-detection frameworks to localize regions of bone resorption. Existing approaches have largely focused on radiographic image analysis, with comparatively limited investigation of multimodal approaches combining panoramic radiographs with structured patient information for severity classification.

Multimodal learning in medical imaging

Multimodal learning combines information from complementary data sources, such as medical images and structured clinical variables, with the aim of capturing information that may not be represented by either modality alone. Imaging-clinical fusion has been investigated across radiology [18], pathology [19], and ophthalmology [20]. Common integration strategies include early fusion, intermediate or feature-level fusion, and late decision-level fusion [21–23]. Feature-level fusion, as adopted in the present study, combines learned representations from modality-specific branches before classification, allowing joint modeling while retaining separate feature-extraction pathways [21].

Transfer learning is particularly relevant when medical imaging datasets are limited in size. Progressive fine-tuning of pretrained convolutional backbones can permit task-specific adaptation while limiting abrupt changes to previously learned representations [24, 25].

Recent multimodal studies in radiology and nuclear medicine have demonstrated consistent benefits of imaging-clinical fusion while also highlighting challenges of explainability, robustness, and generalizability across institutions [26–29]. These studies provide methodological context for the present work rather than direct

evidence for periodontal assessment, and they emphasize the importance of external validation when evaluating multimodal medical AI systems.

Automated periodontal severity assessment

Automated assessment of periodontal disease severity has received increasing attention alongside the development of standardized periodontal classification frameworks [30, 31]. Xia et al. [11] introduced the Bone Resorption Age Ratio (BRAR), an age-adjusted measure combining radiographic bone resorption with patient age, together with a publicly available dataset of panoramic radiographs, structured metadata, and BRAR-based severity labels. The BRAR categories therefore represent an age-normalized radiographic classification target and should not be interpreted as equivalent to a comprehensive clinical diagnosis of periodontal severity.

The original study evaluated an image-only EfficientNet-B0 model and reported 63.3% accuracy on a single data split, with Level 3 recall of 0.348 [11]. Although the dataset includes structured metadata, the reported baseline focused on image-based classification and did not evaluate multimodal fusion of the radiographic and structured variables. Moreover, performance was reported for a single split, leaving the variability of the model across alternative data partitions uncharacterized.

Against this background, the present study systematically evaluates image-only, metadata-only, and multimodal approaches for BRAR-based severity classification under a repeated-split evaluation protocol. In addition to quantitative performance assessment, we examine the influence of age as a metadata predictor and use targeted expert re-evaluation and qualitative Grad-CAM analysis to explore model failure patterns and interpretability.

Methodology

This section describes the dataset, preprocessing pipeline, proposed multimodal architecture, training strategy, and evaluation protocol. An overview of the complete framework is provided in Fig. 1.

Dataset

This study uses the publicly available BRAR-anchored multimodal dataset [11], comprising 988 panoramic radiographs with associated structured metadata: age, gender, number of missing teeth, number of implants, number of residual roots, and number of pairs of functional teeth. The original data descriptor reports 1,104 patients, implying one radiograph per patient, whereas the publicly available release contains 988 panoramic radiographs; the release does not include a patient identifier that would allow the one-to-one correspondence to be verified independently. All available cases were included in the present study without further exclusion.

Disease severity is defined using the Bone Resorption Age Ratio (BRAR), calculated as

$$BRAR = \frac{\text{Bone Resorption} \times 100}{\text{Age}}.$$

Cases are categorized into three severity levels: Level 1 (mild, $BRAR < 0.25$, $n = 170$), Level 2 (moderate, $0.25 \leq BRAR < 1.0$, $n = 560$), and Level 3 (severe, $BRAR \geq 1.0$, $n = 258$). These labels are referred to as reference annotations throughout. The dataset is imbalanced toward Level 2 cases; this imbalance is addressed using the class-weighting strategy described in section “Training strategy”.

Table 1 summarizes patient and radiographic characteristics by severity level. All structured variables differed across severity levels. Mean age differed substantially between Level 1 (29.2 ± 10.0 years) and the two higher severity levels, but was similar between Level 2 (43.3 ± 14.5) and Level 3 (42.0 ± 14.7), indicating that age distinguishes mild cases from the remainder but does not separate the moderate and severe categories. Because age is incorporated directly into the BRAR definition and is also available as an input variable, its influence is evaluated

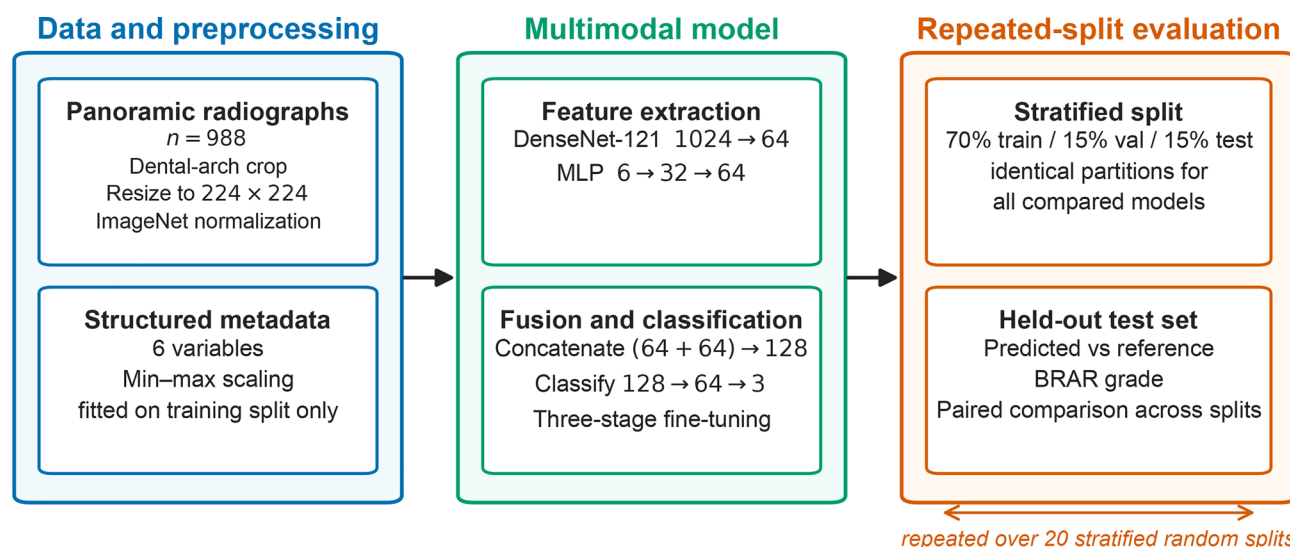


Fig. 1 Study and model framework. Panoramic radiographs are cropped to the dental-arch region, resized, and normalized, while the six structured metadata variables are min–max scaled using statistics fitted on the training split only. Image features from a DenseNet-121 backbone and metadata features from a multilayer perceptron are each projected to 64 dimensions, concatenated, and passed through a classification head producing three severity classes. The model is trained using a three-stage progressive fine-tuning strategy. The entire procedure is repeated across 20 stratified random splits, with identical data partitions used for all compared models to permit paired comparisons. Predictions are evaluated against the reference BRAR grade. The model configuration was fixed before the repeated-split evaluation, which was not used for model development

Table 1 Patient and radiographic characteristics by BRAR severity level for the 988 panoramic radiographs analysed in this study. Values are mean $\pm$ standard deviation unless stated otherwise. *p*-values are from Kruskal–Wallis tests for continuous variables and a chi-squared test for gender, and are reported descriptively. Full distributional summaries for the original dataset are available in [11]

Characteristic	Level 1 (n = 170)	Level 2 (n = 560)	Level 3 (n = 258)	All (n = 988)	<i>p</i>
Age (years)	29.2 $\pm$ 10.0	43.3 $\pm$ 14.5	42.0 $\pm$ 14.7	40.6 $\pm$ 14.8	<0.001
Female, n (%)	48 (28.2)	317 (56.6)	120 (46.5)	485 (49.1)	<0.001
Bone resorption	0.01 $\pm$ 0.04	0.29 $\pm$ 0.11	0.58 $\pm$ 0.23	0.32 $\pm$ 0.23	<0.001
BRAR	0.03 $\pm$ 0.08	0.69 $\pm$ 0.17	1.41 $\pm$ 0.44	0.77 $\pm$ 0.52	<0.001
Missing teeth	0.3 $\pm$ 1.4	1.5 $\pm$ 3.3	2.5 $\pm$ 4.4	1.6 $\pm$ 3.5	<0.001
Implants	0.05 $\pm$ 0.25	0.24 $\pm$ 0.89	0.32 $\pm$ 1.32	0.23 $\pm$ 0.96	0.026
Residual roots	0.02 $\pm$ 0.13	0.28 $\pm$ 0.98	0.39 $\pm$ 1.15	0.26 $\pm$ 0.95	<0.001
Pairs of functional teeth	13.9 $\pm$ 1.1	12.9 $\pm$ 2.4	12.1 $\pm$ 3.1	12.9 $\pm$ 2.5	<0.001

separately in an ablation analysis. Because the data were partitioned using 20 different stratified random splits, the composition of the training, validation, and test subsets varies across splits by construction; stratification preserved the class distribution in every split.

Before model training, each panoramic radiograph was manually cropped by the first author to retain the dental arch and supporting alveolar bone while excluding peripheral craniofacial structures not directly relevant to periodontal assessment. Cropping decisions were based solely on anatomical landmarks, and BRAR severity labels were not consulted when defining the crop boundaries. The same cropping procedure was applied consistently to all images. Because this preprocessing step is manual, the current workflow is not fully end-to-end automated. Automated dental-structure localization or segmentation could be incorporated in future work to replace manual cropping [9].

Data preprocessing

Image preprocessing: Following the manual dental-arch cropping described above, all images were resized to 224×224 pixels to match the input dimensions of the pretrained CNN backbone. Images were then converted to tensors and normalized using the ImageNet channel-wise mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225).

Data augmentation: For the training subset, data augmentation was applied on-the-fly and included random horizontal flipping ($p = 0.5$), random rotation within $\pm 5^\circ$, random affine transformations with translations of up to 5% of the image dimensions and scaling in the range 0.95–1.05, and random brightness and contrast adjustments of up to 0.3. Random histogram equalization ($p = 0.2$) and automatic contrast adjustment ($p = 0.2$) were also applied. Augmentation was restricted to the training images; validation and test images underwent only deterministic resizing and normalization. The same augmentation configuration was used for all 20 repeated stratified random splits.

Metadata preprocessing: Numerical features (age, number of missing teeth, number of implants, number of residual roots, and pairs of functional teeth) were normalized to $[0, 1]$ using min–max scaling. Scaling parameters were estimated exclusively from the training subset of each split and subsequently applied to the corresponding validation and test subsets to avoid information leakage. Gender was encoded as a binary variable (0/1) and was not further normalized.

Experimental setup

Experiments were conducted across 20 repeated stratified random splits to characterize performance variability across different data partitions. For each split, the dataset was partitioned into training (70%), validation (15%), and test (15%) subsets while preserving the class distribution. The same data partition was used for all models within each split, enabling paired comparisons between the image-only, metadata-only, and multimodal approaches.

Baseline models

Metadata-only baseline: A metadata-only multilayer perceptron (MLP) was trained using the six available structured variables and followed the same architecture and training procedure as the metadata branch of the proposed multimodal model. This baseline evaluates the predictive performance achievable using structured metadata without radiographic information.

Image-only baseline: An ImageNet-pretrained DenseNet-121 [32] was trained using radiographic images alone and followed the same three-stage progressive fine-tuning strategy used for the image branch of the multimodal model. The image-only, metadata-only, and multimodal models were evaluated using identical data partitions within each split, enabling direct paired comparisons among these approaches.

Reproduced published-model baseline: For additional context, the publicly available implementation accompanying the original BRAR dataset study [11] was run without modification on the original full panoramic radiographs across 20 repeated random runs. Because this implementation generates its own data partitions, these runs were not used in paired statistical comparisons with the proposed models. Instead, the reproduced results are reported descriptively to characterize the performance and variability of the previously proposed approach under repeated evaluation.

Published baseline: The originally reported single-split accuracy of 63.3% [11] is also provided for historical context. Because this result was obtained under a different evaluation protocol, it is not used for direct statistical comparison with the models evaluated in the present study.

Proposed multimodal architecture

The proposed multimodal framework integrates radiographic features and structured metadata through three components: an image branch, a metadata branch, and a feature-fusion classification module, as illustrated in Fig. 1.

Image branch: An ImageNet-pretrained DenseNet-121 serves as the visual feature extractor, producing a 1024-dimensional feature vector after global average pooling. This representation is projected to a 64-dimensional embedding through a fully connected layer followed by ReLU activation and dropout regularization with stage-dependent dropout rates described in section “[Training strategy](#)”.

Metadata branch: The six structured metadata variables are processed by a two-layer multilayer perceptron. The first fully connected layer maps the six input features to 32 dimensions, followed by ReLU activation and dropout. A second fully connected layer maps the representation from 32 to 64 dimensions, followed again by ReLU activation and dropout.

Fusion and classification: The 64-dimensional image and metadata representations are concatenated to form a 128-dimensional joint feature vector. This representation is passed through a fully connected layer reducing it to 64 dimensions, followed by ReLU activation and dropout, and subsequently through a final linear layer producing three output logits corresponding to Level 1 (mild), Level 2 (moderate), and Level 3 (severe) BRAR categories. Softmax normalization is applied at inference to obtain class probabilities.

Ensemble model: A soft-voting ensemble was additionally evaluated by combining the predicted class probabilities of the multimodal, image-only, and metadata-only models:

$$P_{\text{ensemble}} = w_m P_{\text{multimodal}} + w_i P_{\text{image}} + w_d P_{\text{metadata}}, \\ w_m + w_i + w_d = 1.$$

Ensemble weights were selected using a separate set of 10 tuning runs that were not included in the 20-run primary evaluation. Candidate weight combinations were compared using macro F1-score as the optimization criterion. The selected weights were $w_m = 0.8$, $w_i = 0.1$, and $w_d = 0.1$, and this fixed combination was subsequently evaluated across the 20 main evaluation runs. The ensemble is evaluated as a complementary post-hoc analysis; the single multimodal model is the primary proposed architecture.

Training strategy

The image-only and multimodal models were trained using a three-stage progressive fine-tuning strategy to adapt the ImageNet-pretrained representations to the target task while limiting abrupt changes to the pretrained backbone.

- **Stage 1:** The CNN backbone was fully frozen. Only the image projection layer, metadata branch, and classification head were trained using Adam with a learning rate of 1×10^{-3} , weight decay of 1×10^{-5} , and early-stopping patience of 8 epochs.
- **Stage 2:** DenseBlock3, DenseBlock4, and the intervening transition and final normalization layers were unfrozen and jointly optimized with the remaining trainable layers using a learning rate of 2×10^{-5} , weight decay of 2×10^{-4} , and early-stopping patience of 10 epochs.
- **Stage 3:** The entire CNN backbone was unfrozen and the model was fine-tuned end-to-end using a learning rate of 5×10^{-6} , weight decay of 1×10^{-5} , and early-stopping patience of 10 epochs.

In all stages, learning-rate scheduling was performed using `ReduceLROnPlateau` with a reduction factor of 0.5, and early stopping was based on validation loss. Dropout regularization was increased across the fine-tuning stages: projection dropout was set to $p = 0.15$, 0.45 , and 0.45 for Stages 1–3, respectively, while fusion dropout was set to $p = 0.25$, 0.60 , and 0.60 , respectively. The checkpoint with the lowest validation loss across all stages

was retained for final evaluation. Representative training and validation curves for the three stages are shown in Additional file 1.

Cross-entropy loss was used without label smoothing. To address class imbalance, class-specific weights were incorporated into the loss function using a fourth-root inverse-frequency formulation:

$$w_c = \frac{1}{n_c^{0.25}},$$

where n_c denotes the number of training samples belonging to class c . Class weights were computed from the training subset of each split. The fourth-root weighting scheme was selected during model development based on training and validation performance, providing a milder correction for class imbalance than full inverse-frequency weighting. Stronger weighting schemes were found during model development to reduce majority-class performance without a corresponding improvement in minority-class detection.

The metadata-only model was trained using a single-stage procedure with the same optimizer configuration as Stage 1, as it contains no pretrained CNN backbone requiring progressive unfreezing.

Evaluation protocol

Model performance was evaluated on the held-out test subset of each repeated split. Three metrics were designated as the primary outcomes for model comparison: macro-averaged F1-score, macro-averaged AUC-ROC, and Level 3 recall. Macro F1 and macro AUC provide class-balanced measures of overall classification performance, whereas Level 3 recall was included to specifically assess sensitivity for severe cases. Overall accuracy, macro average precision (macro-AP), and per-class precision, recall, and F1-score were additionally reported as secondary descriptive metrics.

Because the severity labels are ordinal, mean absolute grade error (MAGE) and the proportion of predictions differing from the reference label by more than one severity level were also calculated to characterize the magnitude of classification errors.

Performance was summarized as mean $\pm$ standard deviation across 20 repeated stratified random splits. For the image-only, metadata-only, and multimodal models, identical data partitions were used within each split, allowing paired comparisons. The repeated splits are not intended to provide statistically independent replications but rather to characterize the variability of performance estimates across different data partitions. Pairwise differences in the three primary metrics were evaluated using the Wilcoxon signed-rank test at $\alpha = 0.05$, with zero differences handled by the zsplit method. Holm–Bonferroni correction was applied across the three primary outcomes for each model comparison. Primary comparisons were additionally summarized as the median paired difference with a 95% confidence interval. Confidence intervals for absolute performance were computed using the t -distribution across the 20 splits; confidence intervals for paired differences were obtained by bootstrap resampling (10,000 replicates, percentile method) of the per-split differences. Secondary metrics were treated as descriptive and were not used to define the principal statistical conclusions.

An exploratory probability-level ensemble was also evaluated by combining the multimodal, image-only, and metadata-only model outputs using fixed weights of 0.8, 0.1, and 0.1, respectively. These weights were selected using 10 separate tuning runs that were not included in the 20-run main evaluation, after which the weights were fixed and applied across the 20 repeated stratified test splits. Because the ensemble analysis was exploratory, its results were interpreted separately from the primary model comparisons.

Model interpretability was explored using Gradient-weighted Class Activation Mapping (Grad-CAM) [33]. Activation maps were computed from the final convolutional layer of the DenseNet-121 backbone and visualized on the corresponding radiographs. These maps were used as qualitative illustrations of image regions contributing to model predictions and were not treated as evidence of validated anatomical localization.

A targeted expert error analysis was additionally performed on 88 panoramic radiographs selected according to model prediction outcome and confidence, comprising 33 correctly classified and 55 misclassified cases,

with cases drawn from each reference-predicted combination. An independent licensed dentist graded these images without access to the original dataset labels or model predictions. Agreement between the dentist and the original annotations was quantified separately for the correctly classified and misclassified subsets using both unweighted and weighted Cohen's κ [34]. Because the subset was purposefully selected based on model behavior and involved a single dentist, this analysis was considered exploratory rather than an estimate of population-level inter-annotator agreement.

Results

Baseline comparison and model performance

For descriptive comparison with the previously published image-only approach, we reproduced the publicly available implementation of [11] without modification on the original full panoramic radiographs across 20 repeated random runs. Because that implementation generates its own data partitions, these runs were not paired with the partitions used here and were not included in the paired statistical comparisons.

The originally reported single-split accuracy of 0.633 fell outside the range obtained when the same implementation was evaluated repeatedly (0.54 ± 0.06), illustrating the sensitivity of performance estimates to data partitioning. Our image-only model achieved an accuracy of 0.62 ± 0.04 and a Level 3 recall of 0.56 ± 0.09 , with lower variability across partitions than the reproduced baseline (standard deviation of Level 3 recall 0.09 versus 0.19). These comparisons are descriptive: the two pipelines differ simultaneously in preprocessing, architecture, training strategy, class weighting, and data partitioning, so no difference can be attributed to any single component. Full metric comparisons and an example of the manual dental-arch cropping are provided in Additional file 1.

Multimodal fusion and ablation study

Figure 2 shows the distribution of the three prespecified primary outcomes across the 20 repeated stratified random splits, and Table 2 summarizes performance for each model. The multimodal model exceeded the image-only model on all three primary outcomes. Figure 3 shows this comparison within each split: the multimodal model achieved a higher macro F1-score and macro AUC-ROC on 20 and 19 of the 20 splits respectively, and a

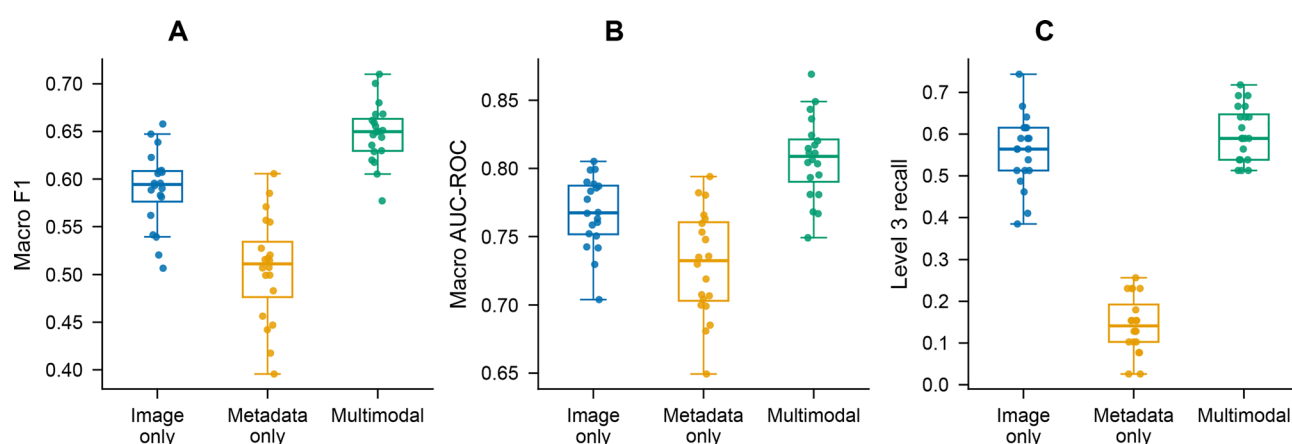


Fig. 2 Distribution of the three prespecified primary outcomes across the 20 repeated stratified random splits for the image-only, metadata-only, and multimodal models. **(A)** Macro F1-score. **(B)** Macro AUC-ROC. **(C)** Level 3 recall. Each point is one split; boxes show the median and interquartile range, and whiskers extend to 1.5 times the interquartile range. All models were trained and evaluated on identical data partitions within each split. The within-split comparison between the image-only and multimodal models is shown in Fig. 3, with paired effect estimates and 95% confidence intervals in Fig. 4

Table 2 Performance across the 20 repeated stratified random splits (mean $\pm$ standard deviation). Macro F1, macro AUC-ROC, and Level 3 recall are the three prespecified primary outcomes; accuracy and mean absolute grade error (MAGE) are reported descriptively. 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 maximizing macro F1 on 10 separate tuning seeds not included among the 20 evaluation seeds. within-split comparisons are shown in Fig. 3, and paired effect estimates with 95% confidence intervals are shown in Fig. 4. Per-class metrics are provided in Additional file 1

Metric	Image-only	Metadata-only	Multimodal	Ensemble
Macro F1	0.59 $\pm$ 0.04	0.51 $\pm$ 0.05	0.65 $\pm$ 0.03	0.65 $\pm$ 0.03
Macro AUC-ROC	0.77 $\pm$ 0.03	0.73 $\pm$ 0.04	0.81 $\pm$ 0.03	0.81 $\pm$ 0.03
Level 3 recall	0.56 $\pm$ 0.09	0.14 $\pm$ 0.07	0.60 $\pm$ 0.07	0.57 $\pm$ 0.07
Accuracy	0.62 $\pm$ 0.04	0.62 $\pm$ 0.04	0.67 $\pm$ 0.03	0.68 $\pm$ 0.02
MAGE	0.39 $\pm$ 0.04	0.41 $\pm$ 0.03	0.35 $\pm$ 0.03	0.34 $\pm$ 0.03

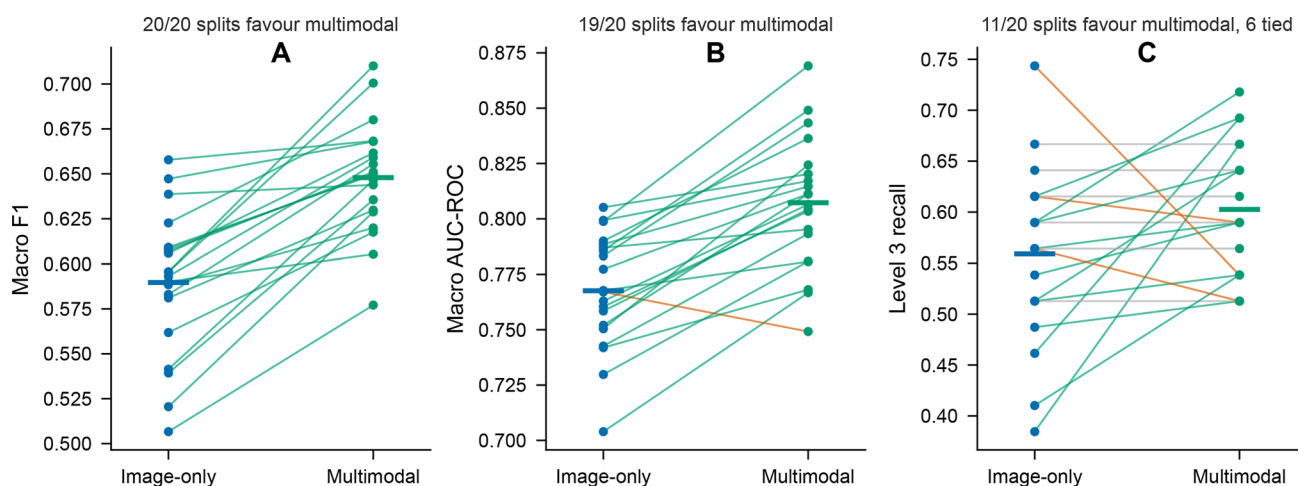


Fig. 3 Within-split comparison of the multimodal and image-only models across the 20 repeated stratified random splits. (A) Macro F1-score. (B) Macro AUC-ROC. (C) Level 3 recall. Each line joins the two models evaluated on the same data partition; green indicates a higher value for the multimodal model on that split, vermillion a higher value for the image-only model, and grey an identical value. Horizontal bars show group means, and panel titles give the number of splits in each direction. Level 3 recall is quantized in steps of $1/39$ by the number of Level 3 cases in each test partition, so identical values occur on some splits. Paired differences with confidence intervals and corrected significance tests are reported in Fig. 4, and the corresponding comparison against the metadata-only model is provided in Additional file 1

higher Level 3 recall on 11, with 6 splits tied. Figure 4 reports the corresponding paired differences with bootstrap 95% confidence intervals. After Holm–Bonferroni correction, the multimodal model significantly outperformed both unimodal baselines on all three outcomes; the smallest margin was against the image-only model in Level 3 recall ($p = 0.036$).

The metadata-only model achieved an accuracy comparable to the image-only model but a markedly lower macro F1-score. This discrepancy reflects a tendency to assign the majority class: Level 2 recall reached 0.85 ± 0.05 while Level 3 recall was 0.14 ± 0.07 . Structured metadata alone therefore carried predictive information, but was insufficient to identify severe bone resorption without radiographic features. Per-class metrics for all models are provided in Additional file 1.

An exploratory probability-level ensemble combining the multimodal, image-only, and metadata-only models with weights of 0.8, 0.1, and 0.1 was also evaluated. The weights were selected by grid search maximizing macro F1 on 10 separate tuning seeds not included among the 20 evaluation seeds; performance was insensitive to the exact values, with the five best weight combinations differing by less than 0.005 in macro F1. Compared with the single multimodal model, the ensemble did not significantly differ in macro F1 or macro AUC-ROC but significantly reduced Level 3 recall. Small improvements in micro-averaged ranking metrics and in mean absolute

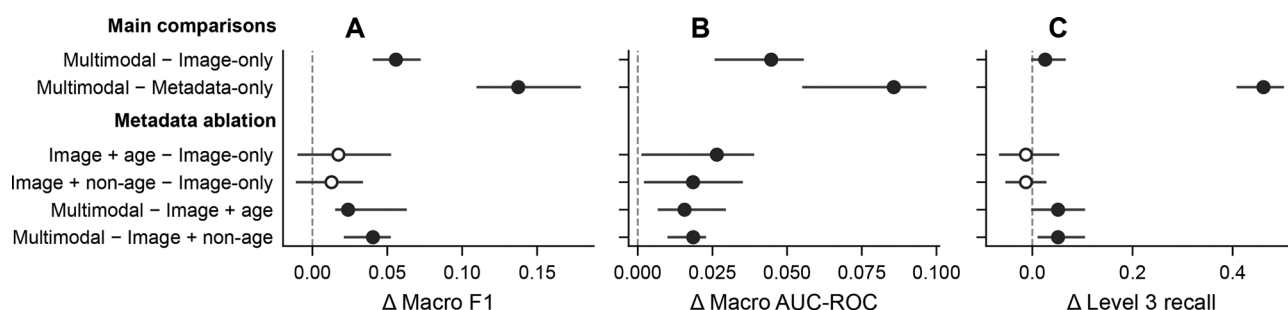


Fig. 4 Paired differences between models across the 20 repeated stratified random splits, shown as the median difference with a bootstrap 95% confidence interval (horizontal bars). Positive values favour the first-named model. Filled markers indicate Holm–Bonferroni-adjusted $p < 0.05$ across the three primary outcomes for that comparison; open markers indicate no significant difference. The metadata ablation group shows two routes from the image-only model to the full multimodal model, adding age or the remaining five structured variables first; note that age is mathematically incorporated into the BRAR label definition. Level 3 recall is quantized in steps of $1/39$, so interval bounds can fall on multiples of this value. Absolute performance is shown in Table 2 and Fig. 2 for the principal models, and in Additional file 1 for the ablation configurations. Comparisons with late fusion and with the exploratory ensemble are reported in Additional file 1

grade error were accompanied by this reduction; because missed Level 3 cases were reclassified predominantly as Level 2, both are consistent with a shift in the decision boundary rather than independent effects.

A late-fusion variant was additionally evaluated to examine whether the observed benefit arises from learning a joint representation or from access to both information sources. The separately trained image-only and metadata-only models were combined by weighted averaging of their predicted class probabilities, with weights selected on the same 10 tuning seeds. Both configurations therefore use identical input information and differ only in the stage at which the modalities are combined. Late fusion performed significantly worse than joint fusion on all three primary outcomes after Holm–Bonferroni correction, with the largest difference in Level 3 recall (0.40 ± 0.11 versus 0.60 ± 0.07); as with the ensemble, it achieved higher Level 2 recall and Level 3 precision but substantially lower Level 3 recall. Full results for both comparisons are provided in Additional file 1.

Given the clinical importance of Level 3 detection, the single multimodal model was retained as the principal model for subsequent analyses.

Metadata ablation and the contribution of age

Because age is incorporated directly into the BRAR definition and is also included among the structured metadata inputs, we evaluated the contribution of age and the remaining metadata using four configurations: image-only, image plus age, image plus the five non-age metadata variables, and the full multimodal model. Paired differences between these configurations are shown in the metadata ablation group of Fig. 4.

Adding either metadata subset alone to the image-only model produced only modest changes. Neither differed significantly from the image-only model on macro F1 or Level 3 recall after Holm–Bonferroni correction (image plus age: $p = 0.307$ and $p = 0.627$; image plus non-age metadata: $p = 0.330$ and $p = 0.432$); both improved only macro AUC-ROC ($p = 0.011$ and $p = 0.005$, respectively). Notably, supplying age alongside the radiograph did not improve Level 3 recall relative to the image-only model. In contrast, the full multimodal model significantly outperformed both partial configurations on all three primary outcomes after Holm–Bonferroni correction (versus image plus age: $p < 0.001$, $p = 0.003$, and $p = 0.015$; versus image plus non-age metadata: $p < 0.001$, $p < 0.001$, and $p = 0.009$).

These results indicate that age contributes to BRAR classification but does not independently account for the multimodal performance advantage. The two partial configurations performed similarly and neither improved class-balanced classification or sensitivity to severe cases, whereas the full model improved all three primary outcomes. This pattern is consistent with complementary contributions from age and the remaining structured

metadata. In particular, the absence of any Level 3 recall improvement when age was supplied alongside the radiograph is not the result expected if age provided a direct route to the age-normalized label. Absolute performance for each configuration and complete pairwise comparisons are provided in Additional file 1.

Error patterns

Because the severity labels are ordinal, the magnitude of classification errors was characterized using mean absolute grade error (MAGE) and the proportion of predictions differing from the reference label by more than one grade, computed across all 20 splits (Table 2). The ensemble achieved the lowest MAGE (0.34 ± 0.03), followed by the multimodal model (0.35 ± 0.03), indicating an average deviation of less than half a severity grade. Two-grade errors were uncommon in all models, ranging from 1.2% to 2.3% of predictions; for the multimodal model this proportion was $1.5\% \pm 1.2\%$. Errors were therefore overwhelmingly between adjacent severity grades, with confusion between Levels 2 and 3 the predominant pattern.

Per-class discrimination followed the same structure. Across the 20 splits the multimodal model achieved a macro-averaged AUC-ROC of 0.81 ± 0.03 and a macro-averaged average precision of 0.69 ± 0.04 . Level 2 showed the lowest per-class AUC-ROC in all 20 splits, consistent with its position as the intermediate grade bordering both adjacent severity levels. Receiver operating characteristic curves, precision–recall curves, and the confusion matrix for a representative split are provided in Additional file 1.

Model interpretability (GradCAM)

Figure 5 shows GradCAM activation maps for three cases with a Level 3 reference grade: a correct classification, a one-grade error, and a two-grade error. Each case is shown as the cropped panoramic radiograph alongside the corresponding activation overlay. Activations were computed from the final convolutional layer of the DenseNet-121 backbone, and cases were selected as the highest-confidence example in each reference–prediction category from predictions pooled across the 20 repeated stratified random splits. Maps for all nine reference–prediction combinations are provided in Additional file 1.

Across the illustrated examples, GradCAM activation overlaps dentoalveolar regions, including areas surrounding the teeth and supporting alveolar bone. Activation was spatially heterogeneous, ranging from relatively localized regions to broader patterns across the dental arches. In the two-grade error, the most severe bone loss appears in the upper left region, whereas GradCAM activation is concentrated predominantly in the mandibular arch, consistent with the localization challenge inherent in a reference grade determined by the single most severely affected tooth.

These examples illustrate that misclassification can occur even when model activation overlaps anatomically plausible regions related to periodontal bone support. However, no tooth-level localization annotations were available to determine whether the activated regions correspond to the specific tooth or region responsible for the BRAR reference label. These GradCAM examples should therefore be interpreted as qualitative visualizations of model activation rather than evidence that the model correctly localized the bone-resorption features determining the severity grade. Because the examples were selected according to prediction confidence rather than sampled randomly, they are illustrative and are not intended to characterize the distribution of activation patterns across the complete dataset.

Targeted expert error analysis

To further characterize model errors, an independent licensed dentist performed visual severity grading of 88 panoramic radiographs. The subset was not randomly sampled; it was purposively constructed from pooled model predictions and confidence scores and included 33 correctly classified and 55 misclassified cases. This subset was

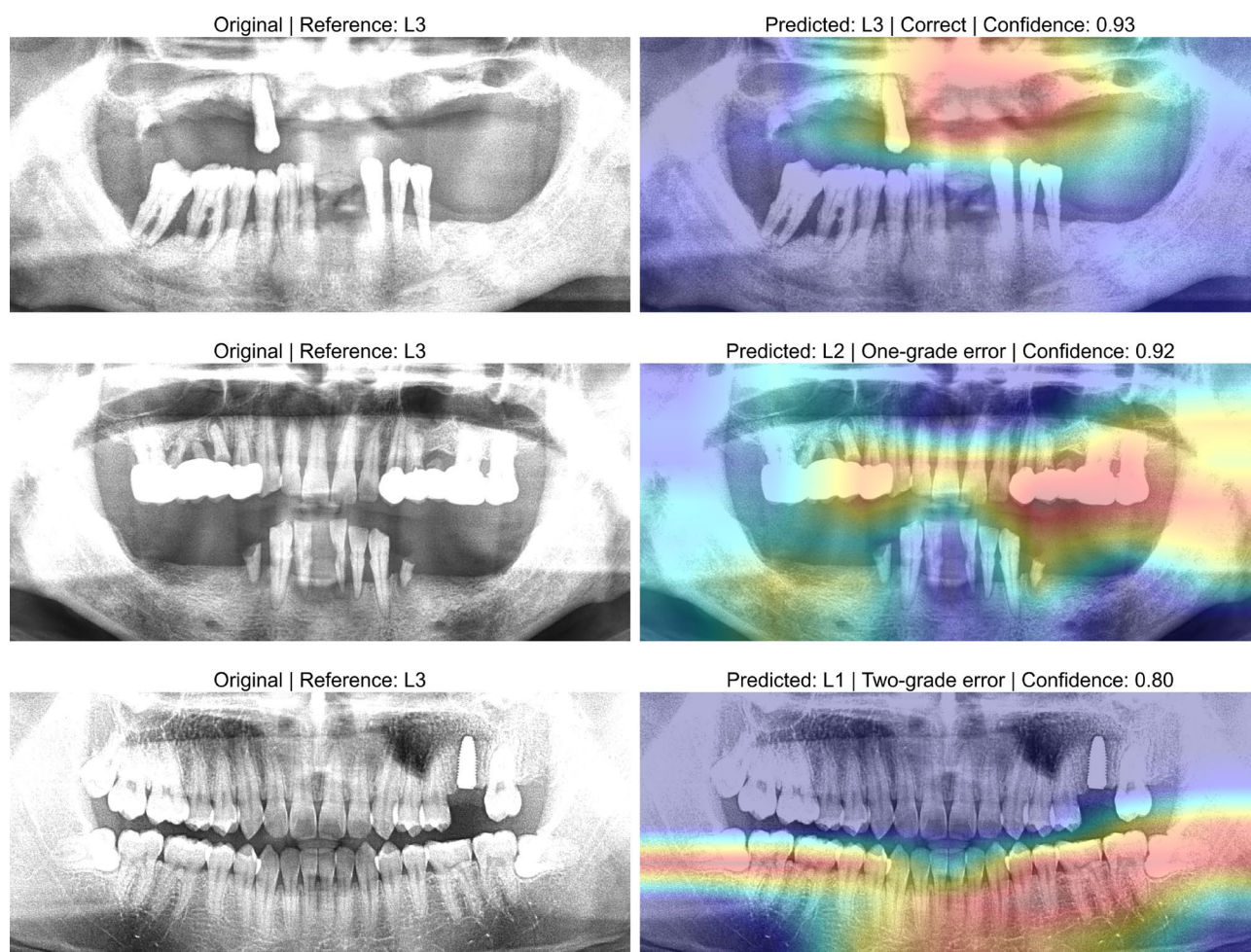


Fig. 5 GradCAM activation maps for three cases with a Level 3 reference grade. Each row shows the cropped panoramic radiograph (left) alongside the corresponding activation overlay (right), with the predicted grade, error category, and model confidence in the predicted class indicated above each panel. Cases were selected as the highest-confidence example in each reference–prediction category from predictions pooled across the 20 repeated stratified random splits; the selected cases therefore represent the model’s most confident predictions within each category and are not a random sample. Activation maps were computed from the final convolutional layer of the DenseNet-121 backbone and are normalized independently for each image using an identical procedure and visual scale, so colour intensity is comparable within but not across panels. These maps indicate image regions that influenced the prediction and do not establish that the model localized the bone-resorption features determining the BRAR severity grade. Activation maps for all nine reference–prediction combinations are provided in Additional file 1

therefore enriched for model errors and was not intended to estimate overall model performance or population-level inter-annotator agreement.

Figure 6 summarizes the analysis. Agreement between the dentist and the reference annotations was fair ($\kappa = 0.353$; weighted $\kappa = 0.409$). Agreement between the model and the reference annotations was lower, as expected given the deliberate enrichment of the subset with model misclassifications. Dentist agreement varied across severity levels: it was highest for Level 1 despite the lowest mean confidence (6.2 ± 2.3 out of 10), and lowest for Level 2 despite the highest mean confidence (7.1 ± 0.9). Within this subset, higher self-reported confidence therefore did not correspond to greater agreement with the reference annotations.

When stratified by whether the model prediction matched the reference annotation, dentist–reference agreement was higher among correctly classified cases ($\kappa = 0.585$; weighted $\kappa = 0.650$) than among misclassified cases ($\kappa = 0.208$; weighted $\kappa = 0.248$). Cases misclassified by the model therefore also showed greater disagreement

Purposively selected, error-enriched subset (n = 88); not representative of overall performance

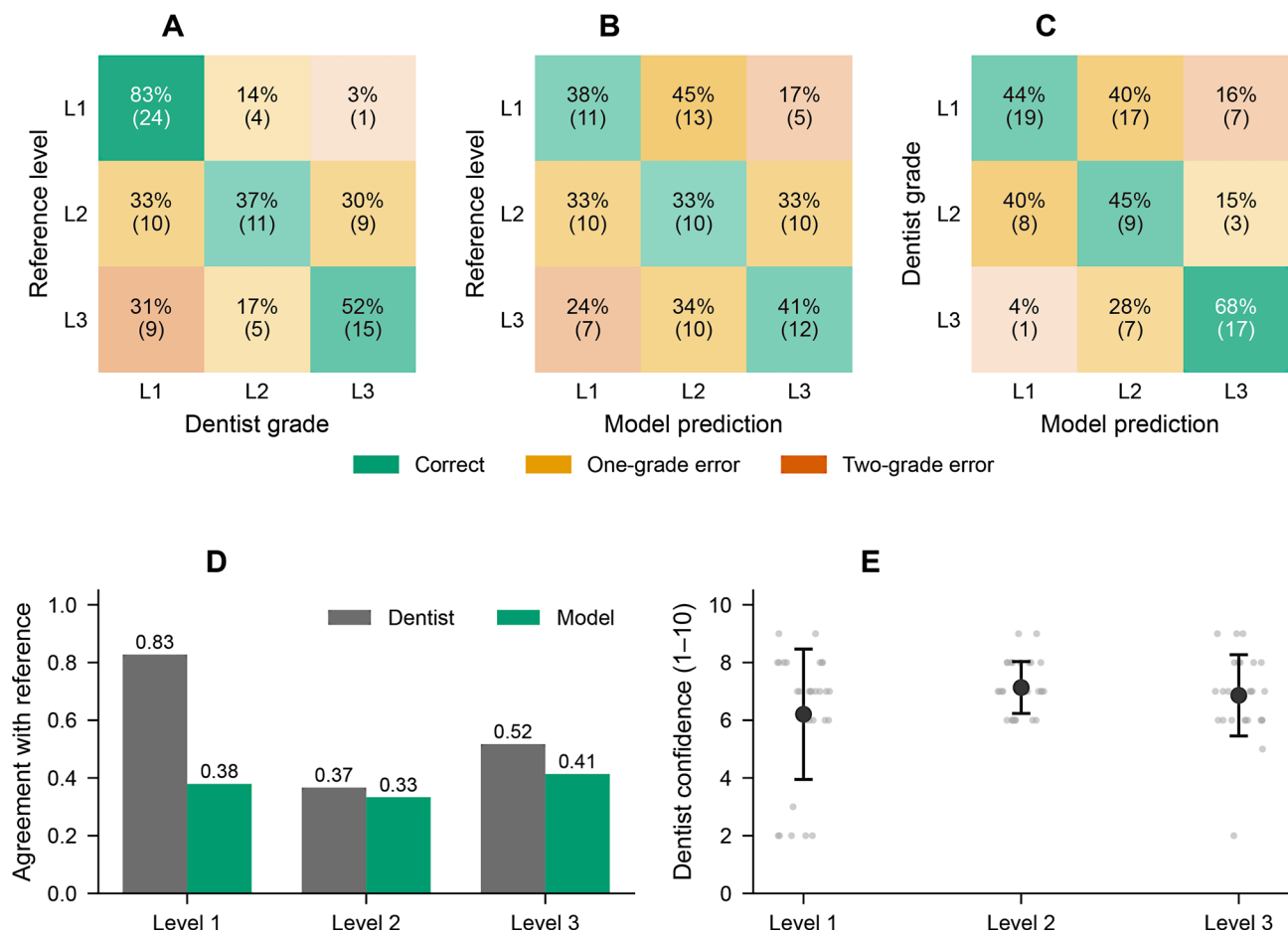


Fig. 6 Targeted expert error analysis on the purposively selected 88-image subset. (A–C) pairwise agreement matrices on identical scales, showing the percentage of cases in each row category assigned to each column category, with counts in parentheses; rows sum to 100%. Cell colour indicates the ordinal distance between the two gradings correct, one-grade, or two-grade difference) rather than frequency magnitude; colour intensity within each category scales with the row percentage. (A) Dentist grading versus the reference annotations. (B) Model predictions versus the reference annotations. (C) Model predictions versus dentist grading. (D) Agreement with the reference annotations by severity level, for the dentist and the model. (E) Dentist confidence by severity level; points are individual cases and markers show the mean with standard deviation. The subset was deliberately enriched for model misclassifications, so panel B and the model values in panel D reflect that selection and are not estimates of overall model performance

between the independent dentist and the reference annotations. Among the misclassified cases, however, agreement between the dentist and the model was also low ($\kappa = 0.074$; weighted $\kappa = 0.167$), indicating that model errors did not simply reproduce an alternative dentist assessment. A modest positive correlation was observed between model and dentist confidence (Spearman $\rho = 0.261$, $p = 0.014$).

These findings are exploratory. Because the subset was purposively selected and evaluated by a single dentist, they do not establish that model errors are caused by diagnostic ambiguity or annotation uncertainty, nor do they provide a population-level estimate of inter-annotator agreement.

Discussion

This study evaluated multimodal deep learning for periodontal bone resorption severity classification from panoramic radiographs and structured metadata. Across 20 repeated stratified random splits, the multimodal model achieved higher performance than both unimodal baselines on the three primary outcomes of macro F1, macro AUC-ROC, and Level 3 recall. Paired comparisons across the same data partitions remained statistically significant after Holm–Bonferroni correction. The ablation analysis showed that age contributed to BRAR classification but did not independently account for the multimodal performance improvement. The full model significantly outperformed the image-plus-age model on all three primary outcomes, consistent with complementary contributions from age and the remaining structured metadata.

Contribution of structured metadata

The metadata-only model performed poorly on the primary outcomes, achieving a macro F1-score of 0.51 ± 0.05 and identifying only 14% of severe cases. Its accuracy was nonetheless comparable to that of the image-only model, reflecting a tendency to assign the majority class. Structured metadata alone was therefore insufficient to identify severe bone resorption, and the discrepancy between accuracy and class-balanced measures illustrates why accuracy alone is an inadequate summary in an imbalanced setting.

The ablation analysis showed that age and the remaining structured metadata provide complementary information when combined with radiographic features. Neither partial configuration differed significantly from the image-only model in macro F1 or Level 3 recall, although both significantly improved macro AUC-ROC, whereas the full multimodal model significantly outperformed both on all three primary outcomes.

These findings indicate that age alone does not account for the observed multimodal improvement. Together with the similar performance of the image-plus-age and image-plus-non-age configurations, the results are consistent with complementary contributions from age and the remaining structured metadata. Nevertheless, because age is mathematically incorporated into the BRAR definition, the relationship between age and the prediction target remains inherent to the formulation and cannot be eliminated by this ablation.

The present study retained BRAR-derived severity as the prediction target to maintain consistency with the original study and enable direct comparison, as the primary objective was to classify the established BRAR severity levels. An alternative formulation would be to estimate bone resorption directly from the radiograph and subsequently normalize the predicted resorption by patient age to derive BRAR severity. This two-stage approach would separate radiographic estimation of bone resorption from the age normalization used to define BRAR and represents an important direction for future work.

Fusion strategy and model combination

The results support joint integration of radiographic and structured information for prediction of the BRAR-derived severity labels. The multimodal model achieved higher macro F1, macro AUC-ROC, and Level 3 recall than either unimodal model, and also outperformed both partial metadata configurations.

The point at which the two modalities were combined was also associated with performance differences. A late-fusion variant, formed by weighted probability averaging of the separately trained image-only and metadata-only models with weights tuned on separate tuning seeds, performed significantly worse than joint fusion on all three primary outcomes, with the largest difference in Level 3 recall (0.40 ± 0.11 versus 0.60 ± 0.07 ; Holm-adjusted $p < 0.001$). Because both approaches used the same input modalities, these findings are consistent with an advantage of learning a joint multimodal representation over combining the predictions of separately trained unimodal models.

The exploratory probability-level ensemble assigned the greatest weight (0.8) to the multimodal model, with weights of 0.1 each for the image-only and metadata-only models. Compared with the single multimodal model,

the ensemble did not significantly improve macro F1 or macro AUC-ROC and showed significantly lower Level 3 recall. Small improvements were observed in some secondary metrics, including mean absolute grade error, but these did not translate into improvement in the primary outcomes. Overall, probability-level combination provided no clear advantage over the single joint multimodal model in this dataset.

Targeted expert error analysis

Within the purposively selected 88-image subset, agreement between the independent dentist and the reference annotations was fair. Dentist agreement varied across severity levels, with the lowest agreement for Level 2 despite the highest mean confidence, indicating that self-reported confidence did not track agreement with the reference annotations.

Dentist–reference agreement was substantially higher among cases correctly classified by the model than among model-misclassified cases. Within this targeted subset, disagreement between the dentist and the reference annotations was therefore greater among cases that the model also misclassified. However, the dentist did not simply reproduce the model’s alternative classifications: among model-misclassified cases, dentist–model agreement was also low. Model confidence showed a modest positive association with dentist confidence.

Taken together, these findings provide additional context for the model errors but should be interpreted descriptively. The subset was deliberately enriched for model misclassifications and was evaluated by a single independent dentist; the analysis does not establish that model errors were caused by diagnostic ambiguity or annotation uncertainty, nor does it provide a population-level estimate of inter-annotator agreement.

Limitations and future work

Several limitations should be considered. First, all experiments used a single publicly available dataset from one institution, limiting assessment of generalizability across institutions, imaging equipment, and patient populations. Performance on single-institution data does not necessarily generalize to independent cohorts, and external validation remains an important direction for future work, alongside prospective data collection planned at our institution. The class imbalance, particularly the smaller numbers of Level 1 and Level 3 cases (26 and 39 in each test partition, respectively), also limits the precision of class-specific performance estimates.

Second, the present study predicts BRAR-derived severity, in which age is incorporated into the definition of the target while also being available to the multimodal model. The metadata ablation showed that age alone did not account for the performance of the full multimodal model; nevertheless, the mathematical relationship between age and the BRAR target remains inherent to the task and cannot be eliminated by this analysis. An alternative formulation for future work would be to estimate bone resorption directly from the radiograph and subsequently normalize the predicted resorption by age to derive BRAR severity.

Third, the reported repeated-split analysis represents internal evaluation within the same public dataset rather than validation on a completely independent cohort. Model development involved preliminary experiments on this dataset before the final configuration was fixed and evaluated across the repeated stratified splits. Although checkpoint selection and early stopping within each evaluation split were based on validation performance, the reported results should therefore not be interpreted as independent external validation.

Fourth, dental-arch cropping was performed manually, and the current workflow is therefore not fully automated. Because the proposed pipeline and the reproduced published implementation differed in architecture, preprocessing, class weighting, augmentation, and training strategy, their comparison does not isolate the effect of cropping, and no independent performance benefit is attributed specifically to this preprocessing step.

Fifth, the targeted expert error analysis was based on a single independent dentist and a purposively selected 88-image subset enriched for model misclassifications. Its agreement statistics therefore should not be interpreted as population-level estimates of inter-annotator agreement. Furthermore, the dentist’s visual grading did not reproduce the original measurement-based BRAR protocol, limiting direct comparison with the original labels.

Similarly, GradCAM analysis was qualitative and lacked tooth-level localization ground truth, preventing assessment of whether model activation corresponded to the specific region determining the BRAR severity label.

Future work should evaluate the approach on external and prospectively collected cohorts and incorporate multiple calibrated dental experts using a standardized BRAR assessment protocol. Explicit tooth-level detection and per-tooth bone-resorption estimation could provide a more localized assessment of disease severity, building on established methods for dental structure segmentation in panoramic radiographs [9]. Automated dental-arch localization could replace the current manual cropping step to support a fully automated pipeline. Finally, richer clinical metadata, including smoking history, systemic conditions, and longitudinal follow-up, could support broader periodontal assessment and progression modelling beyond BRAR severity classification.

Conclusion

This study systematically evaluated multimodal deep learning for BRAR-based periodontal bone resorption severity classification using panoramic radiographs and structured metadata. Across 20 repeated stratified random splits, the multimodal model outperformed the image-only and metadata-only models on the primary outcomes of macro F1, macro AUC-ROC, and Level 3 recall, supporting the value of combining radiographic and structured information for this task. A late-fusion variant combining separately trained unimodal models performed significantly worse than joint fusion, consistent with an advantage of learning a joint multimodal representation for this dataset.

The metadata ablation showed that age contributed to BRAR classification but did not independently account for the performance of the full multimodal model. Image-plus-age and image-plus-non-age configurations showed similar performance, whereas the full multimodal model significantly outperformed the image-plus-age model on all three primary outcomes. These findings are consistent with complementary contributions from age and the remaining structured metadata. Nevertheless, because age is incorporated directly into the BRAR target definition, its mathematical relationship with the prediction target remains an inherent limitation of the current formulation.

The targeted expert error analysis showed greater dentist–annotation disagreement among selected model-misclassified cases than among model-correct cases; however, the purposively selected subset and use of a single independent dentist preclude attributing model errors to annotation uncertainty or diagnostic ambiguity. Grad-CAM provided qualitative evidence of activation in dentoalveolar regions, but anatomical localization was not independently validated.

These findings support further investigation of multimodal approaches for BRAR-based severity classification while emphasizing the need for independent external validation. Future work should evaluate automated dental-arch and tooth-level localization, direct radiographic estimation of bone resorption followed by age normalization, richer clinical information, and validation on independent and prospectively collected cohorts.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1186/s12880-026-02928-6>.

Author contributions A.S. conceived the study, designed and implemented the deep learning framework, conducted all experiments, and wrote the manuscript. I.B. performed the clinical inter-annotator evaluation and contributed to clinical interpretation of results. All authors reviewed and approved the final manuscript.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability The preprocessed (manually cropped) dataset and all experimental data will be made publicly available upon acceptance of the manuscript. The original dataset is publicly available from Xia et al. [11] under Creative Commons Attribution 4.0 (CC BY 4.0) license.

Materials availability Not applicable.

Code availability The source code will be made publicly available upon acceptance at <https://github.com/adnanalshaikh/b-rar-classification>.

Declarations

Ethics approval and consent to participate This study used the publicly available BRAR-anchored multimodal dataset [11], collected and de-identified under ethical approval from the Institutional Ethics Committee of Shanghai Stomatological Hospital (approval number: SHDC22024247). No new patient data were collected in this study.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Hashim NT, Babiker R, Padmanabhan V, Ahmed AT, Chaitanya NCSK, Mohammed R, et al. The global burden of periodontal disease: a narrative review on unveiling socioeconomic and health challenges. *Int J Environ Res Public Health*. 2025;22(4):624. <https://doi.org/10.3390/ijerph22040624>.
2. Pihlstrom BL, Michalowicz BS, Johnson NW. Periodontal diseases. *Lancet*. 2005;366(9499):1809–20. [https://doi.org/10.1016/S0140-6736\(05\)67728-8](https://doi.org/10.1016/S0140-6736(05)67728-8).
3. Tsuchida S, Nakayama T. Recent clinical treatment and basic research on the alveolar bone. *Biomedicines*. 2023;11(3):843. <https://doi.org/10.3390/biomedicines11030843>.
4. Fu D, Shu X, Zhou G, Ji M, Liao G, Zou L. Connection between oral health and chronic diseases. *MedComm*. 2025;6(1):e70052. <https://doi.org/10.1002/mco2.70052>.
5. Krois J, Graetz C, Holtfreter B, Brinkmann P, Kocher T, Schwendicke F. Evaluating modeling and validation strategies for tooth loss. *J Dent Res*. 2019;98(10):1088–95. <https://doi.org/10.1177/0022034519864889>.
6. World Health Organization. Global oral health status report: towards universal health coverage for oral health by 2030. 2022. World Health Organization, Geneva.
7. Lim JI, Regillo CD, Sadda SR, Ipp E, Bhaskaranand M, Ramachandra C, et al. Artificial intelligence detection of diabetic retinopathy. *Ophthalmol Sci*. 2023;3(1):100228. <https://doi.org/10.1016/j.xops.2022.100228>.
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–18. <https://doi.org/10.1038/nature21056>.
9. Rubiu G, Bologna M, Cellina M, Cè M, Sala D, Pagani R, et al. Teeth segmentation in panoramic dental X-ray using mask regional convolutional neural network. *Appl Sci*. 2023;13(13):7947. <https://doi.org/10.3390/app13137947>.
10. Schwendicke F, Golla T, Dreher M, Krois J. Convolutional neural networks for dental image diagnostics: a scoping review. *J Dent*. 2019;91:103226. <https://doi.org/10.1016/j.jdent.2019.103226>.
11. Xia Y, Li Z, Lin Z, Wang S, Wang Y, Xie Y. BRAR-anchored multimodal dataset of panoramic radiographs for periodontal bone resorption grading. *Sci Data*. 2025;13(1):89. <https://doi.org/10.1038/s41597-025-06400-y>.
12. Ghorbani Z, Mirebeigi-Jamasbi SS, Hassannia Dargah M, Nahvi M, Hosseinikhah Manshadi SA, Akbarzadeh Fathabadi Z. A novel deep learning-based model for automated tooth detection and numbering in mixed and permanent dentition in occlusal photographs. *BMC Oral Health*. 2025;25(1):455. <https://doi.org/10.1186/s12903-025-05803-y>.
13. Marwaha J, Nath A, Singla M, Arya A. An end-to-end deep-learning system for segmentation and classification of dental caries from radiovisiography images. *J Conservative Dentist Endodontic*. 2025;28(11):1133–38.
14. Vachmanus S, Pornprasertsuk-Damrongsri S, Mongkolwat P, Phinklao N, Papasratorn D, Kitisubkanchana J, et al. Dual-stage deep neural network for tooth localization and caries segmentation in panoramic dental imaging. *Neural Comput Appl*. 2026;38(6):147. <https://doi.org/10.1007/s00521-026-11865-y>.

15. Asif S, Wang VY, Xu D. OralTransNet: a novel hybrid model integrating transformer attention and CNN features for accurate diagnosis of mouth and oral diseases. *Eng Appl Artif Intel*. 2025;159:111609. <https://doi.org/10.1016/j.engappai.2025.111609>.
16. Jundaeng J, Chamchong R, Nithikathkul C. Advanced AI-assisted panoramic radiograph analysis for periodontal prognostication and alveolar bone loss detection. *Front Dent Med*. 2025;5:1509361. <https://doi.org/10.3389/fdmed.2024.1509361>.
17. Putra RH, Astuti ER, Wijaksana IKE, Fariza A, Sitalaksmi RM, Yoda N. Automated periodontal bone loss detection on panoramic radiographs using you only look Once v8 (YOLOv8): a retrospective AI approach. *J Int Oral Health*. 2025;17(3):203. https://doi.org/10.4103/jioh.jioh_20_25.
18. Wu X, Li J, Wang J, Qian Q. Multimodal contrastive learning for radiology report generation. *J Ambient Intell Hum Comput*. 2023;14(8):11185–94. <https://doi.org/10.1007/s12652-022-04398-4>.
19. Lu MY, Chen B, Williamson F, Chen RJ, Zhao M, Chow AK, et al. A multimodal generative AI copilot for human pathology. *Nature*. 2024;634(8033):466–73. <https://doi.org/10.1038/s41586-024-07618-3>.
20. Jin K, Yu T, Grzybowski A. Multimodal artificial intelligence in ophthalmology: applications, challenges, and future directions. *Surv Ophthalmol*. 2026;71(1):158–67. <https://doi.org/10.1016/j.survophthal.2025.07.003>.
21. Hangloo S, Arora B. Multimodal fusion techniques: review, data representation, information fusion, and application areas. *Neurocomputing*. 2025;649:130827. <https://doi.org/10.1016/j.neucom.2025.130827>.
22. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: from unimodal analysis to multimodal fusion. *Inf Fusion*. 2017;37:98–125. <https://doi.org/10.1016/j.inffus.2017.02.003>.
23. Galanakis I, Soldatos RF, Karanikolas N, Voulodimos A, Voyiatzis I, Samarakou M. Early and late fusion for multimodal aggression prediction in dementia patients: a comparative analysis. *Appl Sci*. 2025;15(11):5823. <https://doi.org/10.3390/app15115823>.
24. Zhou Z, Shin J, Zhang L, Gurudu S, Gotway M, Liang J. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In: *Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2017, vol. 2017, pp. 4761–72. <https://doi.org/10.1109/CVPR.2017.506>.
25. Salama ES, Amer ZM, Girgis P. Bayesian optimization and progressive fine-tuning pipeline for kidney image detection with TTA. *Egypt Inf J*. 2026;34:100964. <https://doi.org/10.1016/j.eij.2026.100964>.
26. Jiang C, Zhang Z, Jiang Z, Ding C, Teng Y, Gao L, et al. Interpretable multimodal PET/CT-EHR fusion via mixture-of-experts for prognostic stratification in mantle cell lymphoma: a multicenter study. *BMC Med*. 2026;24(1):330. <https://doi.org/10.1186/s12916-026-04865-1>.
27. Jiang C, Jiang Z, Zhang Z, Huang H, Zhou H, Jiang Q, et al. An explainable transformer model integrating PET and tabular data for histologic grading and prognosis of follicular lymphoma: a multi-institutional digital biopsy study. *Eur J Nucl Med Mol Imag*. 2025;52(7):2384–96. <https://doi.org/10.1007/s00259-025-07090-9>.
28. Jiang C, Jiang Z, Zhang X, Qu L, Fu K, Teng Y, et al. Robust and interpretable deep learning system for prognostic stratification of extranodal natural killer/T-cell lymphoma. *Eur J Nucl Med Mol Imag*. 2025;52(5):1739–50. <https://doi.org/10.1007/s00259-024-07024-x>.
29. Jiang C, Qian C, Jiang Z, Teng Y, Lai R, Sun Y, et al. Robust deep learning-based PET prognostic imaging biomarker for DLBCL patients: a multicenter study. *Eur J Nucl Med Mol Imag*. 2023;50(13):3949–60. <https://doi.org/10.1007/s00259-023-06405-y>.
30. Messina BM, Polizzi A, Angelova A, Jovanova E, Tartaglia G, Isola G. The impact of artificial intelligence on periodontal disease detection and treatment. *Front Dent Med*. 2026;7:1784123. <https://doi.org/10.3389/fdmed.2026.1784123>.
31. Patil S, Joda T, Soffe B, Awan KH, Fageeh HN, Tovani-Palone MR, et al. Efficacy of artificial intelligence in the detection of periodontal bone loss and classification of periodontal diseases: a systematic review. *J Am Dent Assoc*. 2023;154(9):795–804.e1. <https://doi.org/10.1016/j.adaj.2023.05.010>.
32. Huang G, Liu Z, Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4700–08.
33. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–26.
34. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychologic Meas*. 1960;20(1):37–46. <https://doi.org/10.1177/001316446002000104>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Adnan Salman¹ · Ibtehal Raghieb Bishtawi²

✉ Adnan Salman
aalshaikh@najah.edu

Ibtehal Raghieb Bishtawi
Ibtehal.bishtawi@najah.edu

¹ Computer Science Department, An-Najah National University, P.O. Box 7, Nablus, Palestine

² Faculty of Dentistry and Dental Surgery, An-Najah National University, P.O. Box 7, Nablus, Palestine