scientific reports

Check for updates

OPEN

Deep learning diagnosis of TMJ osteoarthritis using MRI ensembles and CBCT–MRI fusion

Yeon-Hee Lee[1,2,8,9]✉, Seong-Woo Jang[3], Seonggwang Jeon[4], Q.-Schick Auh[2], Fernando P. S. Guastaldi[5], Akhilanand Chaurasia[6] & Yung-Kyun Noh[4,7,9]✉

We compared single-modality models, intra-modality ensembles, and selective cone-beam computed tomography (CBCT)–magnetic resonance imaging (MRI) fusion for diagnosing temporomandibular joint osteoarthritis (TMJ-OA). This single-center retrospective cohort included 1598 patients (1118 females and 480 males; mean age ± SD, 37.88 ± 19.19 years), with patient-level splitting into training, validation, and test sets at a 60:20:20 ratio. Joint-level labels were defined using clinical and imaging criteria. Single models were trained using right/left CBCT views (sagittal and coronal) and MRI sequences (T2-weighted and proton density [PD]). We evaluated intra-modality ensembles (CBCT sagittal + coronal and MRI T2 + PD) and selective CBCT–MRI late fusion. DeLong tests were used to compare model performance, and Grad-CAM was used to assess explainability. In the full dataset, single CBCT models achieved area under the receiver operating characteristic curve (AUROC) values of 0.771 for sagittal views (95% confidence interval [CI], 0.732–0.807) and 0.787 for coronal views (95% CI 0.748–0.821). Single MRI models achieved AUROCs of 0.835 for T2-weighted images (95% CI 0.768–0.893) and 0.827 for PD images (95% CI 0.759–0.888), while the single combined-sequence MRI model achieved an AUROC of 0.881 (95% CI 0.836–0.913). Among the ensemble models, the MRI T2 + PD ensemble showed the highest performance (AUROC, 0.916; 95% CI 0.823–0.977), followed by the CBCT sagittal + coronal ensemble (AUROC, 0.832; 95% CI 0.781–0.881). Selective CBCT–MRI fusion improved performance compared with CBCT alone but did not surpass the best MRI ensemble. Findings were consistent across sex and age strata. Overall, MRI sequence aggregation provides a strong benchmark for TMJ-OA discrimination, whereas selective CBCT–MRI fusion may offer incremental value over CBCT-based models in selected pairings.

Keywords Temporomandibular joint disorders, Osteoarthritis, Magnetic resonance imaging, Cone-beam computed tomography, Deep learning, Artificial intelligence, Neural network

Temporomandibular joint osteoarthritis (TMJ OA) is a prevalent degenerative condition associated with chronic pain, functional limitation, and reduced quality of life[1]. Radiographic signs of TMJ-OA are reported in 18–85% of patients with temporomandibular disorders (TMDs), with the prevalence increasing with age and comorbid musculoskeletal disease[2]. In older adults, imaging studies suggest a high burden, approaching 70% in some cohorts, despite relatively low rates of overt clinical signs[3]. Therefore, accurate and early diagnosis is crucial to prevent progression and optimize outcomes; however, clinical findings often overlap with those of other TMD subgroups, complicating timely and reliable diagnosis[4].

A major challenge in TMJ OA assessment is that no single imaging modality reliably captures the full spectrum of osseous and soft-tissue pathology. Cone-beam computed tomography (CBCT) offers high-resolution

[1]School of Dentistry, Dental Research Institute, Seoul National University, 101, Daehak-ro, Jongro-gu, Seoul 03080, Republic of Korea. [2]Department of Orofacial Pain and Oral Medicine, Kyung Hee University Dental Hospital, #26 Kyunghee-Daero, Dongdaemun-gu, Seoul 02447, Republic of Korea. [3]Equipment Technology Group, Samsung SDI, 146, Gwahaksaneop 4-ro, Ochang-eup, Cheongwon-gu, Cheongju-si, Chungcheongbuk-do 28121, Republic of Korea. [4]Department of Computer Science, Hanyang University, Seoul 04763, Republic of Korea. [5]Division of Oral and Maxillofacial Surgery, Department of Surgery, Massachusetts General Hospital, Harvard School of Dental Medicine, 50 Blossom Street, Thier 513A, Boston, MA 02114, USA. [6]Department of Oral Medicine and Radiology, King George's Medical University, Lucknow, India. [7]School of Computational Sciences, Korea Institute for Advanced Study (KIAS), Seoul 02455, Korea. [8]Department of Orofacial Pain and Oral Medicine, Kyung Hee University Dental Hospital, #613 Hoegi-dong, Dongdaemun-gu, Seoul 02447, Republic of Korea. [9]Yeon-Hee Lee and Yung-Kyun Noh have contributed equally as corresponding authors. ✉email: omod0209@gmail.com; nohyung@hanyang.ac.kr

visualization of osseous changes—such as sclerosis, erosions, condylar flattening, and osteophytes—but provides limited information on the articular disc and inflammatory soft-tissue findings[5]. In contrast, magnetic resonance imaging (MRI) depicts disc position and morphology as well as inflammatory features including synovitis, effusion, and bone marrow edema, yet can be less sensitive to early cortical or subtle subcortical osseous change[6]. Reflecting this complementarity, recent American Academy of Oral and Maxillofacial Radiology–American Academy of Orofacial Pain guidelines recommend CBCT for osseous assessment and MRI for soft-tissue evaluation, with judicious, question-driven use of both modalities when appropriate[7,8]. Notably, the correspondence between CBCT-based osseous findings and MRI-based internal derangements is often weak, which can complicate interpretation from a single modality and supports integrative, question-driven imaging strategies [9].

Deep learning has enabled accurate, scalable interpretation of medical images by learning discriminative features directly from pixel data. Prior work has applied these methods to TMJ OA imaging, but has largely been single-modality (CBCT or MRI) and often lacks robust evidence of generalizability[10,11]. Leveraging the complementary information in CBCT and MRI, a structured multimodal framework with sequence- and view-level aggregation may improve diagnostic performance[12,13]. However, the relative contributions of individual CBCT views and MRI sequences—and the added value of intra-modality ensembling versus cross-modality fusion—remain insufficiently defined[14].

Accordingly, we developed and evaluated convolutional neural network (CNN)-based models for TMJ–OA diagnosis using CBCT views and MRI sequences as (1) single-modality baselines, (2) intra-modality ensembles, and (3) selective CBCT–MRI late fusion. Following contemporary reporting standards (Checklist for Artificial Intelligence in Medical Imaging; Standards for Reporting of Diagnostic Accuracy Studies-AI), we pre-specified patient-wise data partitioning, reference standards, and comparative analyses to ensure transparent and reproducible evaluation[15,16]. We hypothesized that aggregating MRI sequences will yield a strong baseline and that selective fusion with CBCT will improve discrimination over CBCT alone, consistent with modality complementarity, while not necessarily surpassing the best MRI ensemble. This strategy aims to provide a more reliable, acquisition-aware tool for the early detection and management of TMJ–OA.

Methods

Study design, population, and groups

This single-center retrospective diagnostic accuracy study included patients aged 11–90 years who presented with TMD symptoms, including TMJ-localized pain, at the Department of Orofacial Pain and Oral Medicine, Kyung Hee University Dental Hospital, Seoul, Republic of Korea, between January 2023 and September 2024. After applying eligibility criteria, 1598 patients were enrolled (1118 females and 480 males; mean ± SD, 37.88 ± 19.19 years).

The study was organized into two stages. Stage 1 (the baseline cohort) comprised the data of 1214 patients collected from January to December 2023 and was used to develop and compare single-mode and intra-modal ensemble models. Stage 2 (the expanded cohort) integrated the data of an additional 384 patients collected between January and September 2024 to form the expanded dataset (n = 1598), enabling the effect of increased dataset size (data expansion) on model performance to be quantitatively assessed.

Eligibility

The inclusion criteria were as follows: (1) clinical symptoms/signs of TMD (including TMJ-localized pain), (2) availability of CBCT and/or TMJ MRI acquired as part of the diagnostic work-up, and (3) adequate clinical documentation. The exclusion criteria were prior TMJ surgery, systemic arthritis (e.g., juvenile idiopathic arthritis, rheumatoid arthritis), prior fracture or major orofacial surgery, severe degenerative skeletal deformity, severe cognitive impairment precluding questionnaires, poor image quality, and incomplete clinical data. Multimodal fusion analyses were performed only on the subsets that included both CBCT and MRI data.

Ethics

This study was approved by the Institutional Review Board of Kyung Hee University Dental Hospital (KH-DT23052). Given the retrospective design and the use of de-identified clinical data, the requirement for written informed consent was waived. All methods were performed in accordance with the relevant guidelines and regulations.

Reference standard and reader agreement

Two calibrated examiners (Y.H.L. and Q.–S.A.) trained in orofacial pain/TMD independently reviewed the images (CBCT and/or MRI) in conjunction with diagnostic criteria for temporomandibular disorders (DC/TMD) clinical records to generate per-joint labels (0 = non-TMJ-OA, 1 = TMJ-OA)[17]. Disagreements were resolved by consensus. Inter-rater agreement (Cohen's κ), calculated separately for CBCT and MRI, ranged from 0.82 to 0.94.

Imaging acquisition and image assessment

CBCT

CBCT was performed using an Alphard VEGA system (Asahi Roentgen, Kyoto, Japan) with a tube voltage of 80 kVp, tube current of 10 mA, scan time of 17 s, field of view (FOV) 20 × 17.9 cm, and isotropic voxel size of 0.39 mm. The DICOM data were archived using INFINITT PACS software (INFINITT Healthcare, Seoul, Korea). Views were partitioned as right and left sagittal (CBCT views 1 and 3) and right and left coronal (CBCT views 2 and 4). Degenerative osseous changes indicating TMJ-OA included sclerosis, erosion, deformity or flattening, and osteophyte formation; normal condylar morphology criteria followed standard references: round

or oval on axial images; convex, round, or flat on coronal images; and round or S-shaped on sagittal images with an intact, smooth cortical surface). CBCT images were analyzed dichotomously by the experts and defined as either "no evidence of OA" (non-TMJ OA, normal) or "TMJ OA." A normal condyle was defined as round or oval in shape in the axial plane, and convex, round, or flat in the coronal plane[18]. In the sagittal plane, the condyle should be round and S-shaped with an intact smooth cortical outline. Changes in the appearance of the TMJ, indicative of TMJ-OA, were determined based on the presence of sclerosis, erosion, deformity, flattening, or osteophyte formation[19].

MRI

MRI was performed using a 3.0-T system (Signa Genesis; GE Healthcare, Chicago, IL, USA) and a 6×8 cm surface coil. Protocols included T2-weighted sequences (MRI views 1 and 2) and proton density (PD) sequences (MRI views 3 and 4) acquired in the oblique sagittal and coronal planes (slice thickness ≤ 3 mm; FOV 15 cm; matrix 256×224). Detailed sequence parameters (TR/TE, echo train length, bandwidth) have been summarized in our previous studies[20,21]. Images were reviewed using INFINITT PACS. MRI assessment of TMJ-OA focused on marrow edema, joint effusion, synovitis, and subchondral cystic changes; when osseous erosion was suspected on MRI, it was corroborated with CBCT when available.

Data cohorts, labeling, and splits

The analytical unit was the joint (left/right). To prevent information leakage, both the baseline and expanded cohorts were independently split at the patient level into training, validation, and test sets (using a 60:20:20 ratio). For the expanded cohort, the number of joint-view image instances was CBCT 3290 (training), 1100 (validation), and 1100 (test); MRI 1267 (training), 424 (validation), and 427 (test). The participant flow diagram and full per-patient/per-joint counts are provided in Fig. 5.

Preprocessing and augmentation

All models employed an identical preprocessing pipeline. The images were resized to 140×140 pixels and the intensity was normalized to [0,1]. Single-channel grayscale inputs were replicated across three channels for compatibility with ImageNet-initialized backbones. To improve generalization while preserving anatomic plausibility, the following training-only light augmentations were applied: small rotations (± 10°), translations (± 10%), isotropic scaling (± 10%), mild shear, horizontal flips, and mild brightness/contrast jitter bounded by empirical distributions. Validation and test sets were only resized and normalized. The batch size was 32.

Single-mode models

All single-mode models used a VGG16 backbone (ImageNet initialization) with a shallow classification head (flatten → dense 256 → normalization → dropout 0.3 → sigmoid). We optimized the class-weighted binary cross-entropy loss using Adam (maximum of 30 epochs), employing early stopping and learning-rate reduction on plateau. The final decision threshold was fixed at the Youden index on the validation set and was applied to the test set. The configurations included: CBCT views 1 & 3 (sagittal), CBCT views 2 & 4 (coronal), CBCT views 1–4 (all views); MRI sequences 1 & 2 (T2), MRI sequences 3 & 4 (PD), and MRI sequences 1–4 (all sequences).

Multimodal late fusion (subset with both modalities)

To investigate whether integrating the complementary information from osseous (CBCT) and soft-tissue (MRI) imaging can yield superior diagnostic performance, an ensemble analysis was performed. This analysis was conducted on the subset of patients for whom both CBCT and MRI scans were available. Within each data split (training, validation, test), the CBCT and MRI samples were meticulously aligned one-to-one at patient level to ensure valid pairing. The ensemble was implemented as a post-hoc late fusion of the probabilities generated by the pre-trained single-mode models. This decision-level fusion strategy was selected because of its modularity and straightforward implementation, enabling a clear assessment of how the independent predictions from each modality contribute to a combined diagnosis.

The primary fusion rule was an equal-weight mean of the prediction probabilities.

$$P_{ens} = \frac{1}{2}P_{CBCT} + \frac{1}{2}P_{MRI}$$

No additional training or fine-tuning was performed on the ensemble itself. The decision threshold was determined on the validation set using the Youden index and then fixed for testing, consistent with the single-mode models.

To ensure the robustness of our primary findings, we conducted sensitivity analyses to explore alternative fusion strategies. These included: (1) logit averaging (product-of-odds), logit (P ensemble),

$$\text{logit}\ (P_{ens}) = \text{logit}\ (P_{CBCT}) + \text{logit}\ (P_{MRI}), P_{ens} = \sigma\,(\cdot)$$

(2) validation-tuned weighted averaging, (3) reliability-weighted averaging using inverse Brier/ Expected Calibration Error (ECE) weights, (4) rank averaging, and (5) the geometric/product rule. For completeness, a light logistic stacking meta-learner was also considered.

$$\sigma(\beta_0 + \beta_1 logit(P_{CBCT}) + \beta_2 logit(P_{MRI}))$$

These alternative strategies were evaluated solely as sensitivity analyses to confirm the stability of the ensemble approach and were not used to generate separate standalone results. The final performance of the primary ensemble model was assessed on the test set using the area under the receiver operating characteristic curve (AUROC) with 95% bootstrap confidence intervals (CIs; 2000 resamples), along with the accuracy, confusion matrix, precision, recall, and F1-score.

Explainability

Grad-CAM heatmaps were computed from the last convolutional layer using uniform visualization parameters. Spatial focus was summarized as activation sparsity (the proportion of pixels above a fixed normalized threshold) and reported in conjunction with discrimination metrics to aid interpretation. Model performance was evaluated using AUROC, accuracy, and F1-score. In addition, receiver operating characteristic (ROC) curves were generated to visually represent the discriminative ability of the models. Collectively, these analyses provide a comprehensive evaluation of the predictive performance of the deep learning models. The complete source code used for model development and analysis is publicly available at: [https://github.com/seongwoojang1123/Improved-Diagnostic-Accuracy-of-TMJ-osteoarthritis-through-Machine-Learning-Integration-of-CBCT-MRI].

Statistical analysis

Analyses were conducted using SPSS software (version 25.0; IBM), R 4.0.2, and Python 3.9.7. Continuous variables are summarized as mean ± SD and categorical variables as n (%). Group differences in demographic and clinical characteristics were assessed using Student's t test for continuous variables and χ^2 or Fisher's exact tests for categorical variables, as appropriate. Diagnostic performance was evaluated across the pre-specified imaging configurations, including single-modality CBCT view models, single-modality MRI sequence models, intra-modality ensembles, and selective CBCT-MRI late-fusion models. The primary performance metric was AUROC, with 95% CIs estimated using patient-level bootstrapping with 2000 resamples. Secondary performance metrics included accuracy, sensitivity, specificity, precision, F1 score, and confusion matrices. Pairwise AUROC comparisons between models were performed using DeLong's test. AUROC values were interpreted as follows: 0.5 = no discrimination, 0.6–0.7 = poor, 0.7–0.8 = acceptable, 0.8–0.9 = excellent, and ≥ 0.9 = outstanding [22]. Grad-CAM heatmaps and activation sparsity analyses were used to support model interpretability. In all analyses, a p value < 0.05 was considered statistically significant.

Results

Demographic and clinical characteristics

Of the 1598 participants, 944 (59.1%) were classified as TMJ-OA and 654 (40.9%) as non-TMJ-OA. The TMJ-OA group was older than the non-TMJ-OA group (40.35 ± 19.97 vs 34.33 ± 17.42 years; $p < 0.001$), and age-group distributions differed ($p < 0.001$). Females were more frequent in the TMJ-OA group (74.7%) than in the non-TMJ-OA group (63.0%; $p < 0.001$; Table 1), with male-to-female ratios of 239:705 (~ 1:2.95) in the TMJ-OA group and 242:412 (~ 1:1.70) in the non-TMJ-OA group. Age distributions were broadly similar between the sexes, and paired analyses revealed no side-specific prevalence difference (right vs left) for either modality (Fig. 1). These findings support treating the joint as the primary analytical unit.

Diagnostic performance of deep-learning models

Consistent with the study design, we first report the results for the baseline dataset (n = 1214) and then for the expanded dataset (n = 1598) to quantify the effect of data expansion. The AUROC (95% CI) was the primary metric, and pairwise AUROC differences were tested employing DeLong's method (Table 2).

Parameter	Overall (n = 1598) Mean ± SD or n (%)	Non-TMJ-OA (n = 654) Mean ± SD or n (%)	TMJ-OA (n = 944) Mean ± SD or n (%)	P-value
Age (years)[a]	37.88 ± 19.19	34.33 ± 17.42	40.35 ± 19.97	**2.15e-10*** **
Age group[b]				
< 20 years	276 (17.3%)	124 (19.0%)	152 (16.1%)	**1.29e-07*** **
20–39 years	649 (40.6%)	309 (47.2%)	340 (36.0%)	
40–59 years	406 (25.4%)	147 (22.5%)	259 (27.4%)	
60–79 years	241 (15.1%)	69 (10.6%)	172 (18.2%)	
≥ 80 years	26 (1.6%)	5 (0.8%)	21 (2.2%)	
Sex[b]				
Male	481 (30.1%)	242 (37.0%)	239 (25.3%)	**7.35e-07*** **
Female	1117 (69.9%)	412 (63.0%)	705 (74.7%)	

Table 1. Epidemiologic characteristics of the cohort stratified by OA status. SD, standard deviation. [a]Results were obtained using Welch's t-test adjusted by Bonferroni correction to compare the non-TMJ-OA and TMJ-OA groups. [b] Results were analyzed using a χ^2 test between two age groups with Bonferroni adjustment, comparing the non-TMJ-OA and TMJ-OA groups. To obtain significant results, the two-tailed level of statistical significance was set at $p < 0.05$. ***: $p < 0.001$. Significant values are in bold.

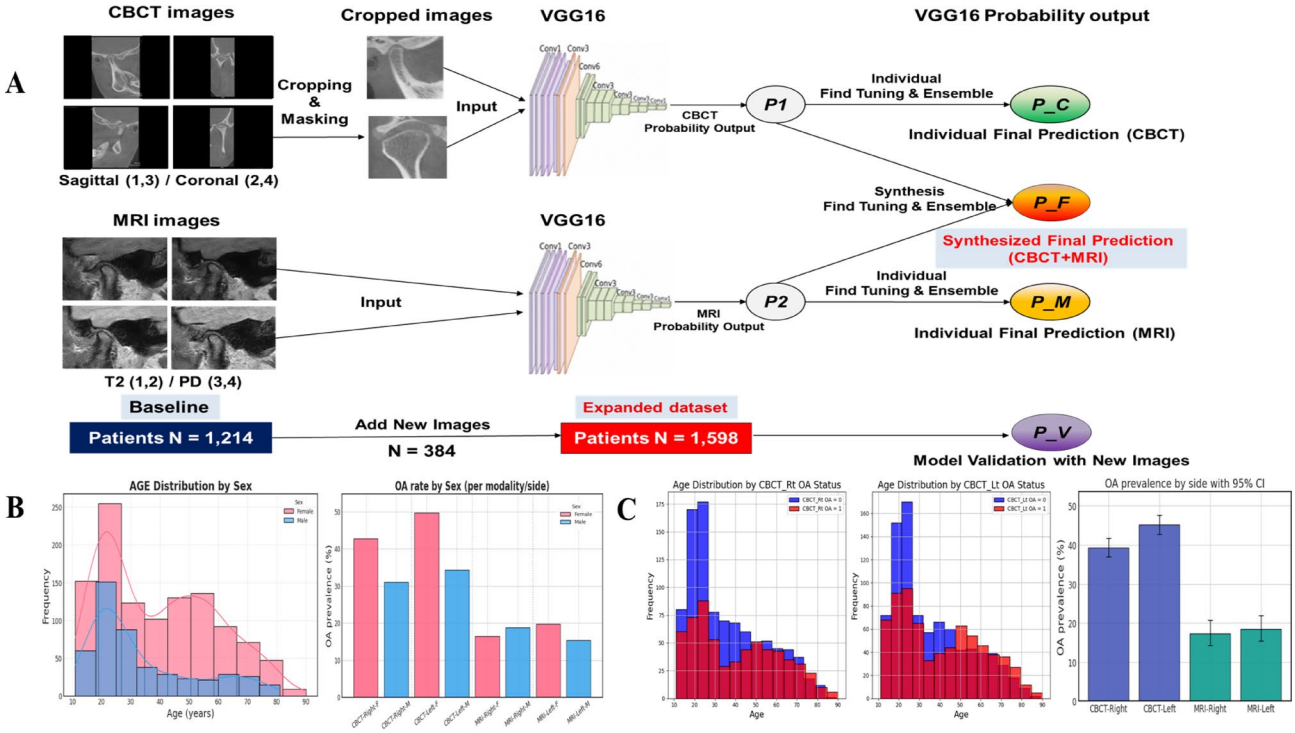


Fig. 1. Study flow and distribution of sex and disease status by age. (**A**) Overall research framework. The CBCT images were pre-processed (cropped and masked) and fed into a VGG16 model to generate an initial probability output (P1). Similarly, the MRI images were input into a separate VGG16 model to produce an initial probability output (P2). Subsequently, three distinct model flows were investigated to derive the final predictions: the CBCT-only model performs an individual fine-tuning and ensemble process using P1 to yield the individual final predictions for CBCT (P_C). The MRI-only model follows a similar individual fine-tuning and ensemble process using P2 to yield the individual final predictions for MRI (P_M). The fusion model combines the initial outputs (P1 and P2) and processes them using a synthesis fine-tuning and ensemble process to generate the synthesized final prediction (P_F). Added step: After model development and comparison in the baseline cohort (N = 1214), the pipeline was further evaluated in the expanded dataset formed by adding 384 patients, yielding a total cohort of 1598 patients. Patient-level splitting was maintained for the training, validation, and test sets to avoid information leakage and to obtain expanded-dataset predictions (P_V in the schematic). (**B**) Sex-specific age distribution and prevalence of OA. Left: Histograms with kernel-density overlays show broadly overlapping age distributions for females and males across the cohort. Right: OA prevalence (%) by sex for each modality or side (CBCT, right/left, MRI, right/left). The bars depict female (pink) and male (blue) rates. The accompanying table lists counts (OA_n), denominators (Total), and percentages (OA_%); "p_value_side" reports the statistical test results for a female–male difference within each modality/side (Pearson's χ^2 or Fisher's exact, as appropriate). No material sex-related imbalances were evident. (**C**) Side-wise OA distribution in CBCT and prevalence with a 95% CI. Left and Center: Histograms of CBCT right views and CBCT left views show age distributions stratified by OA status (OA = 0 vs OA = 1). The shapes were comparable across sides, indicating that age is unlikely to drive a side-specific diagnosis. Right: OA prevalence (%) by side with Wilson 95% CIs for CBCT and MRI. Paired analysis (McNemar) shows no significant right–left difference in diagnostic proportions for either modality ($p \geq 0.05$), supporting side symmetry.

On the baseline dataset, the combined-sequence MRI model (MRI sequences 1–4) achieved the highest performance among single-modality configurations (AUROC, 0.830; 95% CI 0.751–0.891), whereas the best intra-modality ensemble was the CBCT sagittal + coronal ensemble (AUROC, 0.809; 95% CI 0.752–0.865). In the full expanded dataset (n = 1598), the MRI T2 + PD ensemble achieved the highest overall performance (AUROC, 0.916; 95% CI 0.823–0.977). This ranking pattern indicates that complementary MRI sequence aggregation was the dominant contributor to diagnostic performance.

Single-mode models

Diagnostic performance of single-view CBCT models

On the baseline dataset, the coronal-view model (CBCT views 2 & 4) achieved an AUROC of 0.781 (95% CI 0.733–0.832), which was marginally higher than that of the sagittal-view model (CBCT views 1 & 3) at 0.768 (95% CI 0.723–0.812); the difference was not statistically significant ($p > 0.05$; Fig. 2). A naïve combination of all four CBCT views yielded an AUROC of 0.756 (95% CI 0.718–0.788), suggesting that unselective aggregation may introduce feature redundancy. On the full expanded dataset (n = 1598), the AUROC performance improved

Experimental model	AUROC (95% CI)			AUROC (95% CI)		
	Baseline (n = 1214)	*P*-value[a]	New Images (n = 384)	All Images (n = 1598)	*P*-value[b]	*P*-value[c]
Baseline single models						
CBCT Sagittal	0.768 (0.723–0.812)	0.1899	0.703 (0.615–0.788)	0.771 (0.732–0.807)	0.9195	0.0852
CBCT Coronal	0.781 (0.733–0.832)	0.1463	0.698 (0.601–0.802)	0.787 (0.748–0.821)	0.8484	0.1936
MRI T2	0.706 (0.586–0.831)	0.6769	0.747 (0.595–0.879)	0.835 (0.768–0.893)	0.0660	**Top Model**
MRI PD	0.718 (0.593–0.814)	0.7732	0.744 (0.603–0.845)	0.827 (0.759–0.888)	0.0950	0.8614
Combined models or fusion ensembles						
CBCT (Sagittal + Coronal) Combining	0.756 (0.718–0.788)	0.0686	0.696 (0.641–0.760)	0.766 (0.738–0.794)	0.6619	**0.0003*****
MRI (T2 + PD) Combining	0.830 (0.751–0.891)	0.8421	0.818 (0.713–0.903)	0.881 (0.836–0.913)	0.2108	0.4255
CBCT Sagittal + CBCT Coronal Fine Tuning	0.742 (0.692–0.788)	0.6431	0.718 (0.630–0.804)	0.757 (0.714–0.797)	0.6431	**0.0003*****
CBCT Coronal + CBCT Sagittal Fine Tuning	0.737 (0.688–0.784)	0.1652	0.654 (0.564–0.744)	0.772 (0.731–0.809)	0.2673	**0.0011****
CBCT Sagittal + MRI T2 Fine Tuning	0.655 (0.519–0.778)	0.2520	0.761 (0.647–0.864)	0.799 (0.715–0.872)	0.0624	**0.0371***
CBCT Coronal + MRI PD Fine Tuning	0.726 (0.608–0.831)	0.9912	0.727 (0.565–0.874)	0.839 (0.781–0.892)	0.0754	0.1118
MRI T2 + CBCT Sagittal Fine Tuning	0.714 (0.664–0.763)	0.3703	0.644 (0.574–0.757)	0.731 (0.688–0.771)	0.6060	**0.0001*****
MRI PD + CBCT Coronal Fine Tuning	0.717 (0.668–0.765)	0.9168	0.726 (0.636–0.816)	0.742 (0.701–0.784)	0.4427	**0.0001*****
CBCT Sagittal + CBCT Coronal Ensemble	0.809 (0.752–0.865)	**0.0104***	0.640 (0.500–0.777)	0.832 (0.781–0.881)	0.5502	0.0729
MRI T2 + MRI PD Ensemble	0.750 (0.395–1.000)	0.6288	0.807 (0.622–0.954)	0.916 (0.823–0.977)	0.2973	**Top Model**
CBCT Sagittal + MRI T2 Ensemble	0.801 (0.643–0.993)	0.9080	0.700 (0.476–0.884)	0.808 (0.750–0.864)	0.9401	**0.0271***
CBCT Coronal + MRI PD Ensemble	0.667 (0.465–0.865)	0.8672	0.623 (0.457–0.772)	0.859 (0.785–0.925)	0.0757	0.2831
MRI T2 + CBCT Sagittal Ensemble	0.798 (0.637–0.927)	0.3540	0.737 (0.556–0.899)	0.839 (0.781–0.892)	0.6010	0.1118
MRI PD + CBCT Coronal Ensemble	0.709 (0.514–0.879)	0.8252	0.748 (0.557–0.901)	0.851 (0.762–0.926)	0.1642	0.2573
CBCT (Sagittal + Coronal) + MRI (T2 + PD) Ensemble	0.826 (0.729–0.913)	0.2312	0.709 (0.522–0.876)	0.805 (0.748–0.859)	0.7021	**0.0219***

Table 2. Comparison of AUROC performance across all CBCT and MRI image combinations using baseline and new images. These results were obtained using DeLong's test. This table summarizes the classification performance of the diagnostic models before and after inclusion of the new imaging data. Performance is reported as the area under the receiver operating characteristic curve (AUROC), with the corresponding 95% confidence intervals (CIs) in parentheses. Experimental Models: The models evaluated in this study, including the individual, fine-tuned, and ensemble strategies. Baseline: Performance of models evaluated using the baseline cohort collected from January to December 2023. New Images: Performance of the same model configurations evaluated using the additional patient data collected from January to September 2024. All Images: Performance evaluated using the expanded dataset, which combined the baseline cohort and the additional 384 patients, yielding a total of 1598 patients. *p*-value: Derived from a statistical test comparing the AUROC obtained using the baseline and new images. A small *p*-value (< 0.05) indicates a significant change in diagnostic performance after incorporating new data. [a] Comparison of the AUROC values of the baseline models. [b]: Comparison of the AUROC values of the baseline and new image models; [c]: Comparison of the individual models when using all data. CBCT: 1 = right sagittal, 2 = right coronal, 3 = left sagittal, 4 = left coronal, MRI: 1 = right T2, 2 = left T2, 3 = right PD, 4 = left PD, CBCT Sagittal (CBCT 1 + 3), CBCT Coronal (CBCT 2 + 4), MRI T2 (MRI 1 + 2), MRI PD (MRI 3 + 4). Significant values are in bold.

modestly: the result of CBCT views 2 & 4 increased to 0.787 (95% CI 0.748–0.821), that of CBCT views 1 & 3 increased to 0.771 (95% CI 0.732–0.807), and that of CBCT views 1–4 increased to 0.766 (95% CI 0.738–0.794).

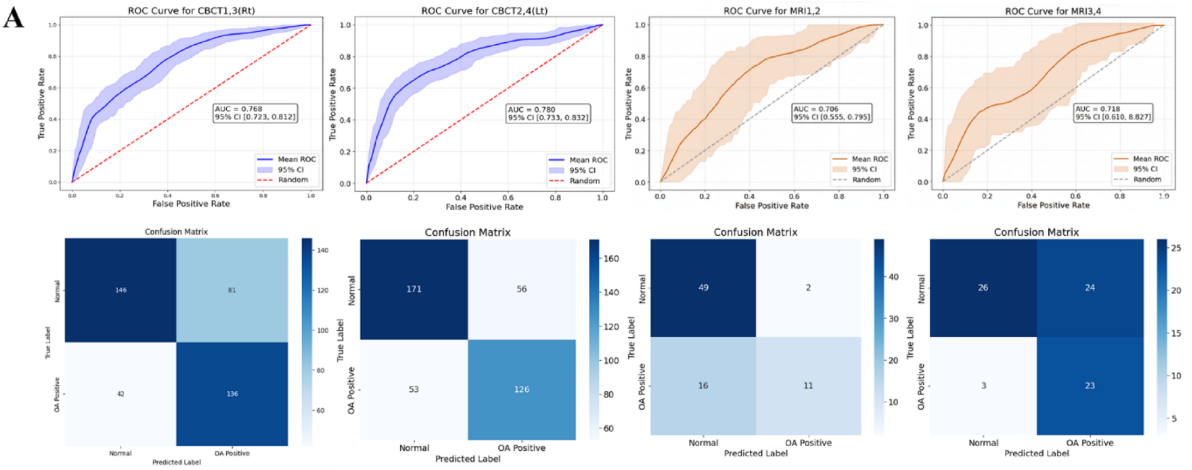
Late-fusion of T2 and PD MRI sequences

On the baseline dataset, the proton-density (PD) model (MRI sequences 3 & 4) achieved an AUROC of 0.718 (95% CI 0.593–0.814), and the T2-weighted model (MRI sequences 1 & 2) achieved an AUROC of 0.706 (95% CI 0.586–0.831; Fig. 2); the difference was not statistically significant ($p > 0.05$). Integrating all four MRI sequences improved AUROC performance substantially to 0.830 (95% CI 0.751–0.891), which highlights sequence complementarity even at smaller sample sizes. On the expanded dataset, all MRI models further increased: MRI sequences 1–4 reached 0.881 (0.836–0.913), PD (MRI sequences 3 & 4) rose to 0.827 (0.759–0.888), and T2 (MRI sequences 1 & 2) increased to 0.835 (0.768–0.893).

Intra-modality ensembles

CBCT ensemble

Late fusion of the sagittal and coronal models (CBCT views 1 & 3 + CBCT views 2 & 4) improved the AUROC from 0.781 (for the best single-view model) to 0.809 (95% CI 0.752–0.865) on the baseline dataset. On the expanded dataset (n = 1598), the AUROC further increased to 0.832 (95% CI 0.781–0.881), confirming the complementary nature of the sagittal and coronal bony features captured by CBCT imaging. This improvement was statistically significant compared with the results of the single-view models ($p = 0.0104$, DeLong's test).



B

Comparison	P-Value
CBCT Sagittal vs MRI T2	0.3245
CBCT Sagittal vs MRI PD	0.3879
CBCT Coronal vs MRI T2	0.2038
CBCT Coronal vs MRI PD	0.2519
CBCT Sagittal vs CBCT Coronal	0.5703
MRI T2 vs MRI PD	0.9319

- **CBCT 1: Right sagittal view**
- **CBCT 2: Right coronal view**
- **CBCT 3: Left sagittal view**
- **CBCT 4: Left coronal view**
- **MRI 1: Right T2-weighted view**
- **MRI 2: Left T2-weighted view**
- **MRI 3: Right PD view**
- **MRI 4: Left PD view**

Fig. 2. Performance evaluation and statistical comparison of CBCT and MRI-based models for TMJ OA diagnosis. (**A**) Diagnostic performance of individual baseline models. The diagnostic performance for OA is shown for four distinct baseline models: CBCT Sagittal (CBCT 1 + 3), CBCT Coronal (CBCT 2 + 4), MRI T2 (MRI 1 + 2), MRI PD (MRI 3 + 4). Receiver operating characteristic (ROC) Curve (top row) illustrates the diagnostic capability of each model. The area under the curve (AUC) represents the overall performance, where a value closer to 1.0 indicates better classification accuracy. The confusion matrix (bottom row) provides a detailed breakdown of the prediction results, comparing the model's predicted labels ("Normal" or "OA Positive") against the true labels. (**B**) Statistical comparison of the performance of the baseline models. The table displays the *p*-values from the pairwise statistical comparisons of the AUCs of the baseline models. A *p*-value greater than 0.05 indicates no statistically significant difference in the performance of the models. Because all p-values in the table were greater than 0.05, the diagnostic performances of the four baseline models are considered statistically similar.

MRI ensemble

Integrating T2-weighted and PD models (MRI sequences 1 & 2 + MRI sequences 3 & 4) yielded an AUROC of 0.750 (95% CI 0.395–1.000) on the baseline dataset, which increased significantly to 0.916 (95% CI 0.823–0.977) on the expanded dataset, representing the highest diagnostic performance of the intra-modal ensembles. This substantial gain underscores the complementary soft-tissue information contained in the T2 and PD sequences.

Multimodal fusion (CBCT + MRI)

On the baseline dataset, the multimodal late-fusion models integrating CBCT and MRI achieved strong discrimination. For instance, combining all inputs (CBCT views 1–4 and MRI sequences 1–4) resulted in an AUROC of 0.826 (95% CI 0.729–0.913). On the expanded dataset, strategic pairings outperformed naïve aggregation: the combination of coronal CBCT (CBCT views 2 & 4) and PD MRI (MRI sequences 3 & 4) achieved an AUROC of 0.859 (95% CI 0.785–0.925), whereas merging all CBCT and MRI views decreased the AUROC to 0.805 (95% CI 0.748–0.859), suggesting that unfiltered aggregation may dilute discriminative signals because of feature redundancy or modality imbalance. Although selective CBCT-MRI fusion improved discrimination relative to CBCT-only models, it did not surpass the MRI sequence ensemble. The MRI T2 + PD ensemble achieved the highest AUROC (0.916; 95% CI 0.823–0.977), whereas the strongest selective CBCT–MRI fusion pairing reached an AUROC of 0.859 (95% CI 0.785–0.925). This pattern suggests that the MRI sequence ensemble provided a strong diagnostic signal, and that adding CBCT through simple probability-level fusion did not translate into a higher AUROC than the best MRI ensemble.

Subgroup analyses

Performance was stable across sex and age strata. For the MRI sequences 1–4 (T2 + PD) ensemble, the AUROCs were similar between females (0.839; 95% CI 0.783–0.890) and males (0.848; 95% CI 0.763–0.919; $p = 0.631$). By age, the AUROC ranking was as follows: ≥ 80 years (1.000; 95% CI 1.000–1.000) > 60–79 years (0.882; 95% CI 0.775–0.967) > < 20 years (0.871; 95% CI 0.759–0.957) > 20–39 years (0.866; 95% CI 0.805–0.922) > 40–59 years (0.801; 95% CI 0.686–0.895), although these differences were not statistically significant ($p = 0.905$). For the CBCT views 2 & 4 (coronal) + MRI sequences 3 & 4 (PD) ensemble, the AUROC scores were 0.823 for females

and 0.915 for males (p = 0.197), with wider CIs reflecting smaller subgroup sizes. By age, the AUROC results were 0.944, 0.905, 0.875, and 0.667 for the < 20, 20–39, 40–59, and 60–79 age groups, respectively (overall $p > 0.05$). Collectively, these findings indicate no systematic demographic bias within the precision limits of the available subgroup sizes (Fig. 3).

Model interpretability

To assess face validity and mitigate shortcut learning, we examined Grad-CAM maps for representative true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) cases across the top configurations: CBCT views 1 & 3 (sagittal), CBCT views 2 & 4 (coronal), MRI sequences 1 & 2 (T2), MRI sequences 3 & 4 (PD), MRI sequences 1–4 (T2 + PD ensemble), and the combined CBCT views 2 & 4 (coronal) + MRI sequences 3 & 4 (PD) fusion model.

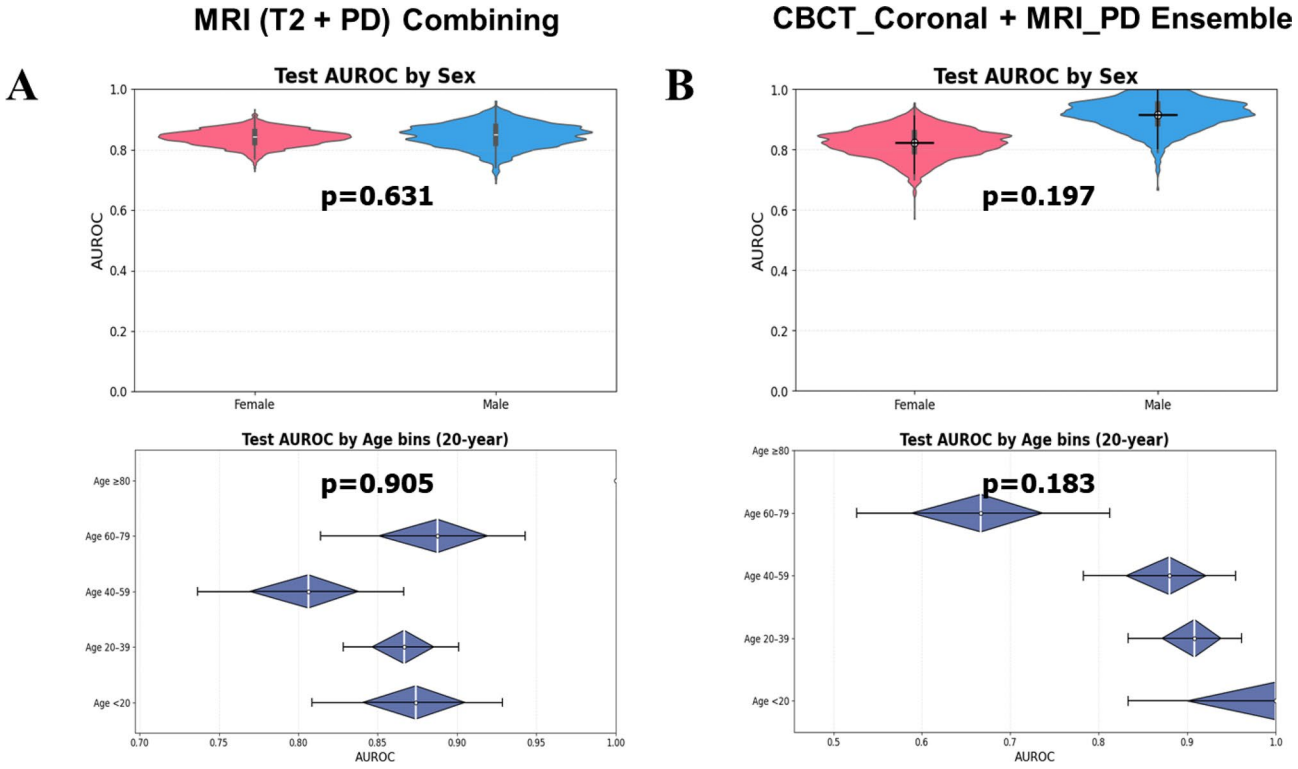


Fig. 3. Sex- and age-stratified test performance using AUROC. (**A**) Sex- and age-stratified AUROC analysis of the combined-sequence MRI model using MRI sequences 1–4 (T2 + PD) on the test set. The overall test AUROC was 0.881 (95% CI 0.836–0.913). In the sex-stratified analysis, the AUROC was 0.839 in females (95% CI 0.783–0.890; n = 306, pos_n = 51) and 0.848 in males (95% CI 0.763–0.919; n = 121, pos_n = 29), with no significant difference between sexes (p = 0.631). In the age-stratified analysis, the AUROC was 0.871 for patients aged < 20 years (95% CI 0.759–0.957; n = 68, pos_n = 13), 0.866 for those aged 20–39 years (95% CI 0.805–0.922; n = 163, pos_n = 30), 0.801 for those aged 40–59 years (95% CI 0.686–0.895; n = 123, pos_n = 26), 0.882 for those aged 60–79 years (95% CI 0.775–0.967; n = 67, pos_n = 10), and 1.000 for those aged ≥ 80 years (95% CI 1.000–1.000; n = 6, pos_n = 1), with no significant difference across age groups (p = 0.905). The perfect AUROC estimate in the ≥ 80-year group should be interpreted with caution because of the very small sample size. Confidence intervals were estimated using bootstrapping with 2000 resamples. (**B**) Sex- and age-stratified AUROC analysis of the selective CBCT–MRI fusion model combining CBCT views 2 and 4 (coronal) with MRI sequences 3 and 4 (PD) on the test set. The overall test AUROC was 0.859 (95% CI 0.785–0.925). In the sex-stratified analysis, the AUROC was 0.823 in females (95% CI 0.719–0.914; n = 73, pos_n = 22) and 0.915 in males (95% CI 0.800–1.000; n = 33, pos_n = 8), with no significant difference between sexes (p = 0.197). In the age-stratified analysis, the AUROC was 0.944 for patients aged < 20 years (95% CI 0.750–1.000; n = 11, pos_n = 2), 0.905 for those aged 20–39 years (95% CI 0.797–0.985; n = 44, pos_n = 14), 0.875 for those aged 40–59 years (95% CI 0.728–0.985; n = 30, pos_n = 12), and 0.667 for those aged 60–79 years (95% CI 0.444–0.882; n = 20, pos_n = 2), with no significant difference across age groups (p = 0.183). The ≥ 80-year group was not estimable because of insufficient cases (n = 1, pos_n = 0). The lower AUROC estimate and wider confidence interval in the 60-79-year group likely reflect the small sample size and class imbalance. Confidence intervals were estimated using bootstrapping with 2000 resamples. AUROC, area under the receiver operating characteristic curve; CBCT, cone-beam computed tomography; CI, confidence interval; MRI, magnetic resonance imaging; PD, proton density; pos_n, number of TMJ-OA-positive cases; T2, T2-weighted; TMJ-OA, temporomandibular joint osteoarthritis.

Modality-specific attention patterns

On the CBCT data, the TPs reveal a high-intensity focus along the superior and anterosuperior condylar cortex, joint space, and subchondral regions, corresponding to sclerosis/erosion or marginal osteophytes. The TNs yield low-intensity and diffuse maps without persistent hotspots along cortical defects. On the MRI data, the TPs emphasize the disc-condyle interface, retrodiscal tissue, joint effusion, and marrow edema signals on T2/PD, whereas the TNs result in low-saliency maps or show attention confined to non-pathologic soft tissues.

Attention compactness (activation sparsity)

Channel-wise activation sparsity for the top-100 channels indicates that 82%–87% of the image area was non-activated, consistent with a compact, anatomically plausible focus rather than global, non-specific highlighting (Fig. 4). Qualitatively, higher model confidence co-occurred with more compact saliency, particularly for the MRI sequence ensemble (MRI sequences 1–4) and the CBCT views 2 & 4 (coronal) + MRI sequences 3 & 4 (PD) fusion.

Error characteristics

In FP CBCT cases, saliency is occasionally localized to sharp cortical bends, dental metallic streaks, or external auditory canal margins, suggesting that edge-based cues or minor artifacts are triggers. FP MRI cases sometimes highlight synovial folds or faint fluid without corroborating osseous changes. FN CBCT cases included significantly early or limited cortical irregularities, or suboptimal head positioning, which reduced the depiction of marginal changes. FN MRI cases often showed minimal effusion or subtle marrow signals that fell below the sensitivity of the single-sequence model but were captured more effectively after sequence aggregation (MRI sequences 1–4).

Cross-modality complementarity

For the best multimodal pair (CBCT views 2 & 4 + MRI sequences 3 & 4), visual inspection suggested co-localized yet complementary attention: CBCT highlighted cortical and subchondral osseous changes, and MRI emphasized inflammatory and soft tissue signals in the same joint. This aligns with the observed AUROC gains over CBCT alone and supports the biological plausibility of late-fusion improvements.

Collectively, the models appeared to rely on clinically meaningful anatomical substrates of TMJ-OA rather than spurious correlations (Fig. 4).

Discussion

Consistent with our hypothesis, selective integration of structural and soft-tissue information improved discrimination relative to CBCT-only models, while the best-performing approach remained the MRI sequence

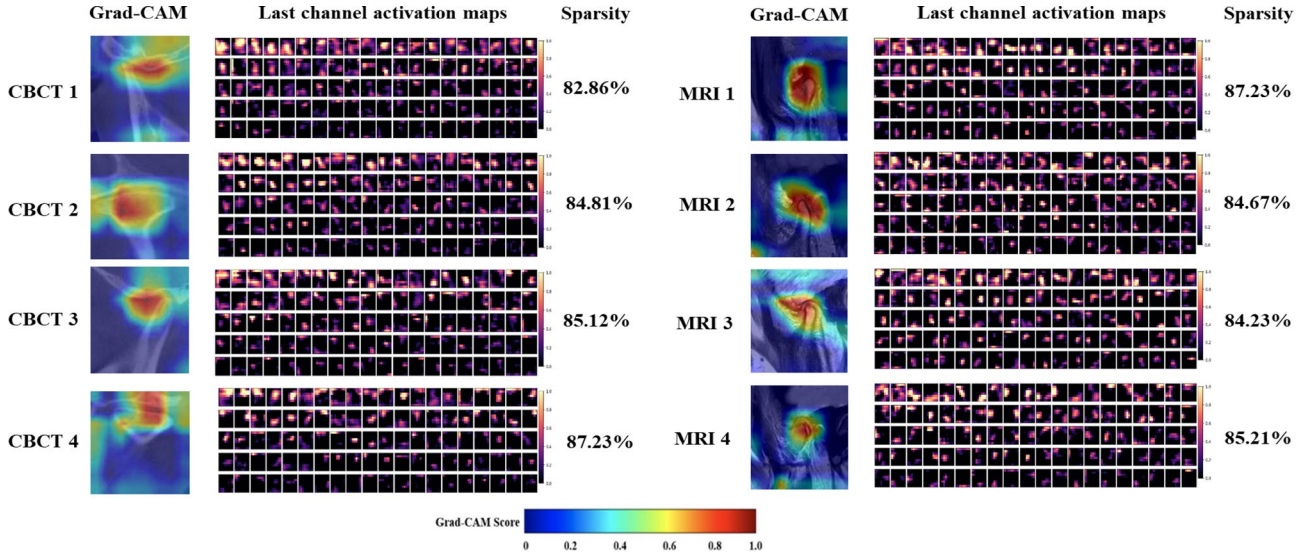


Fig. 4. Heatmaps with Grad-CAM and sparsity. Representative examples of Grad-CAM heatmaps (left) and the corresponding last-layer activation maps from the top-100 most active channels (right) for TMJ-OA cases. Grad-CAM highlights regions that contribute the most to the model's decision, demonstrating that the fine-tuned network selectively attends to disease-related areas. Activation maps are displayed after percentile-based normalization (vmax = p99), and sparsity is defined as the proportion of pixels with normalized activation < 0.3. For the examples shown, the overall sparsity across all 512 channels ranged between 94 and 96% (activation density 4–6%), while the displayed top-100 channels showed a sparsity of 82–87% (activation density 13–18%). These values indicate that only a subset of channels and spatial locations contributed strongly to the prediction, reflecting enhanced feature selectivity after fine-tuning. The color scale bar indicates Grad-CAM scores ranging from 0 (blue) to 1 (red).

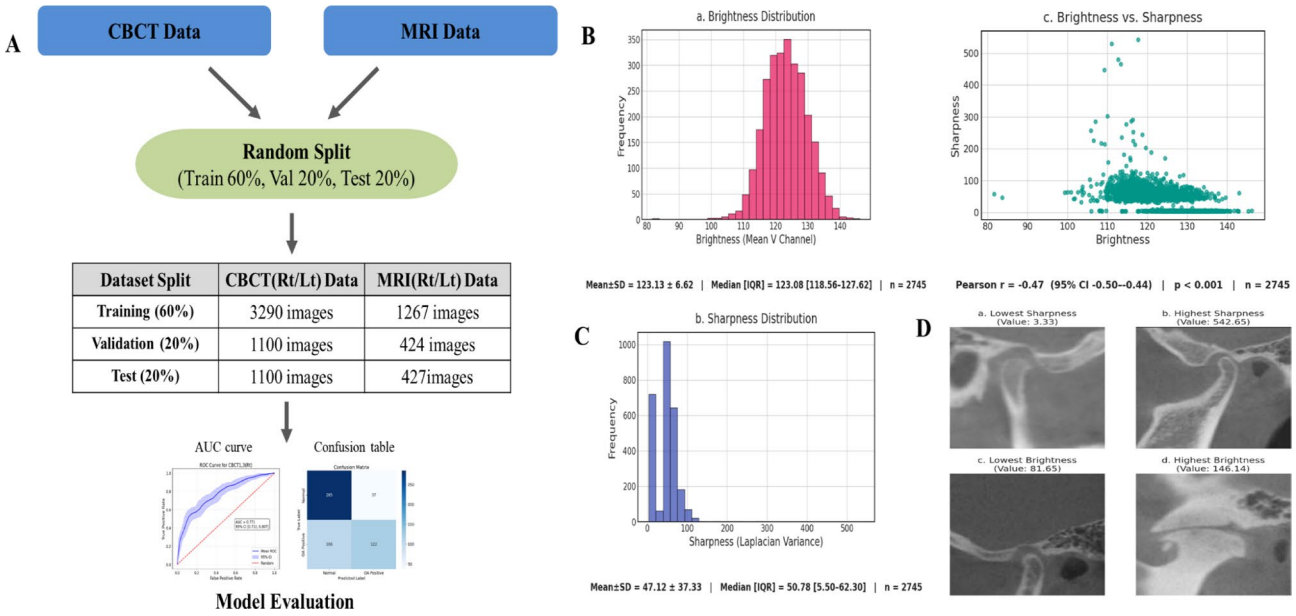


Fig. 5. Overall framework for data preprocessing, model training, and evaluation. (**A**) CBCT and MRI data were randomly split into training (60%), validation (20%), and test (20%) sets. The split was applied before preprocessing to avoid leakage. The validation set was used for hyper-parameter tuning and early -stopping, whereas the held-out test set provided an unbiased estimate of generalization. The final performance was summarized using the AUROC and confusion matrices on the test set, forming the basis for model selection and error analysis. (**B**) Brightness characteristics. Left: Brightness histogram (HSV-V mean) with a mean ± SD of 123.13 ± 6.62 and median (IQR) of 123.08 (118.56–127.62; n = 2,745), indicating near-normal global intensity across images. Right: Relationship between brightness and sharpness, revealing a moderate negative correlation (Pearson r − 0.47; 95% CI − 0.50 to − 0.44; $p < 0.001$; n = 2745). These statistics establish objective quality control (QC) baselines and directly inform preprocessing (brightness normalization) and augmentation ranges, such that the synthetic variation remains within the empirically observed envelope. (**C**) Sharpness exhibits a right-skewed distribution (mean ± SD 47.12 ± 37.33; median [IQR] 50.78 [5.50–62.30]; n = 2745), reflecting heterogeneity in edge clarity due to acquisition differences. This skew justifies the robust statistics (median and IQR) for QC thresholds and motivates augmentation with realistic blur and noise to improve robustness. (**D**) Lowest vs highest sharpness (3.33 vs 542.65); lowest vs. highest brightness (81.65 vs. 146.14) for examples from CBCT views 1 & 3 images. Visualizing the tails validates that anatomical landmarks remain interpretable at extremes and helps set augmentation bounds (brightness, jitter, blur, and noise) without degrading the clinically salient detail.

ensemble. The best single-modality result was the MRI sequence ensemble (T2 + PD; AUROC 0.916), and the most effective multimodal pairing was coronal CBCT + PD MRI (AUROC 0.859), which outperformed CBCT in isolation but did not surpass the MRI ensemble. The non-ensemble combined–sequence MRI model (T2 + PD) also exhibited effective performance (AUROC 0.881), reinforcing sequence complementarity (Fig. 5).

In our cohort, the multi-sequence MRI approach outperformed CBCT, while CBCT retained a superior depiction of cortical detail, underscoring the complementarity of the modalities. CBCT is particularly effective for detecting osseous changes such as cortical irregularity, subchondral sclerosis, and osteophytes[5]. By contrast, MRI provides a superior assessment of disc position and inflammatory changes, including joint effusion and marrow edema[23]. Additional reports have emphasized the value of MRI for soft-tissue derangement in TMDs and its limitations in depicting fine bony contours[24]. Notably, our MRI models outperformed the CBCT models, which is consistent with the evidence that multi-sequence MRI can also interrogate osseous involvement; bone-marrow-level signal changes (edema/osteitis), subchondral cystic changes, and even erosions on fat-suppressed T2/PD can be detected and have clinical correlates[25,26]. CBCT remains more sensitive to cortical detail, but MRI-detected marrow and subchondral alterations have prognostic and diagnostic value, and the performance improves when appropriate sequences are combined. Collectively, these findings support the complementary, rather than interchangeable, roles of CBCT and MRI in TMJ-OA assessment, an interpretation further supported by our selective fusion results.

In contrast to prior TMJ-OA AI work, which predominantly employs a single-modality, our selective multimodal strategy provides a broader diagnostic representation and, in selected pairings, improves discrimination over CBCT-only baselines. Most prior machine-learning studies in this domain rely on a single modality, typically CBCT or MRI alone, which constrains generalizability and leaves unresolved questions regarding the optimal integration of these modalities[27]. Subsequent studies have reinforced this single-modality focus, often with modest datasets and limited fusion analyses[28]. By contrast, our results show that principled multimodal integration provides a more comprehensive representation of TMJ-OA by leveraging the osseous

sensitivity of CBCT with the soft-tissue/inflammatory contrast of MRI, yielding superior accuracy when selectively combined rather than indiscriminately pooled.

The ensemble approach is not a panacea. When applied selectively and in an acquisition-aware manner (e.g., coronal CBCT + PD MRI), it improves performance; when applied indiscriminately, it can add redundancy and diminish accuracy. This interpretation is consistent with multimodal learning literature, which emphasizes that effective fusion requires appropriate representation, alignment, and integration of complementary information rather than simple aggregation of modalities[29]. A natural question arising from our results is why the intra-modality MRI ensemble numerically exceeded the strongest cross-modality CBCT-MRI fusion model. This finding should not be interpreted as evidence against multimodal imaging, but rather as an indication that more input modalities do not automatically guarantee better discrimination. Multimodal models can underperform the best unimodal model when the added modality increases model complexity, contributes partially redundant information, or generalizes differently from the dominant modality[30]. The T2- and PD-weighted MRI sequences likely captured mutually complementary TMJ-OA features, including joint effusion, synovitis, marrow signal alterations, and subchondral changes, which may have generated a strong MRI-based diagnostic signal[24–26]. By contrast, CBCT primarily depicts cortical and subchondral osseous information[5]. Although this information improved performance compared with CBCT-only models when selectively combined with MRI, equal-weight late fusion may have introduced redundancy, modality imbalance, and probability-level calibration differences. Because modern neural networks may produce poorly calibrated probability estimates, simple averaging of output probabilities can suboptimally weight the contribution of each modality[31].

Ensemble strategies are widely employed to reduce variance and improve generalizability by combining models with complementary strengths[32]. However, gains are not universal; when base learners are highly correlated or redundant, ensembles may provide limited added value[33], and the additional computational burden can hinder clinical deployment[34]. Moreover, recent evidence suggests that strong single models such as transformer architectures can sometimes rival or outperform CNN ensembles[35]. Aligning with these nuances, our data support selective, acquisition-aware fusion: targeted pairing (coronal CBCT + PD MRI) outperformed CBCT-only models, whereas naïvely aggregating all views and sequences marginally reduced performance. Taken together, the apparent advantage of the MRI ensemble likely reflects both the diagnostic strength of multi-sequence MRI and the limitations of uniform probability-level fusion. Future work should therefore focus on calibration-aware or attention-based fusion strategies, such as learned weighting or attention bottlenecks[36], so that genuinely complementary CBCT information can be selectively incorporated without diluting the MRI-driven signal.

CNNs provide a suitable foundation for this study because they are particularly adept at capturing complex spatial hierarchies in medical imaging data. Unlike traditional machine-learning approaches that rely on handcrafted features, CNNs automatically learn discriminative features directly from imaging inputs, thereby improving scalability and reducing observer bias[10]. Prior TMJ imaging studies have reported promising CNN performance for tasks such as osteoarthritic change and disc displacement detection, albeit in single-mode settings[37]. By extending CNNs into a multimodal ensemble framework, we demonstrated that jointly leveraging structural (CBCT) and soft-tissue/inflammatory (MRI) cues maximizes diagnostic accuracy, and that advanced multimodal approaches (e.g., attention-based fusion or transformer backbones) may further improve performance and interpretability[38]. This MRI T2 + PD performance corresponds to a 0.129 absolute -point improvement over the best CBCT-only model (AUROC 0.787), quantifying the added value of sequence aggregation.

From a clinical standpoint, these findings underscore the potential of multi-modal deep learning approaches to aid in the early detection and more precise management of TMJ-OA. Age- and sex-stratified analyses further indicated stable diagnostic performance across subgroups, with particularly high discrimination in patients aged 20–39 years, the age group in which TMJ-OA commonly presents. The small drop observed in older patients and the perfect AUROC in patients ≥ 80 years should be interpreted with caution because of the limited sample size and class imbalance. Although TMJ-OA is more frequently reported in females, likely because of differences in craniofacial morphology, bone composition, hormonal influences, and central sensitization[39,40], our analyses did not find any meaningful sex-related differences in diagnostic accuracy. Similarly, a previous systematic review noted that the prevalence of TMJ-OA increases with age and is more common in older subgroups of TMDs[41]. However, in our data, the diagnostic performance did not differ significantly across age strata, suggesting that the proposed models maintain robust discrimination irrespective of patient sex or age.

The strengths of this study include the use of a large, well-characterized cohort, rigorous patient-wise data partitioning to prevent information leakage, and the statistical comparison of AUROCs using DeLong's test. Explainability analyses also confirmed that the models attended to clinically relevant anatomical regions, supporting the biological plausibility of the predictions. In addition to these methodological strengths, this study makes several important contributions to the field of TMD imaging. First, it integrates CBCT and MRI to provide a more comprehensive analysis of TMJ-OA. Previous machine learning studies have typically focused on either CBCT or MRI alone, often with limited datasets[42,43]. By contrast, we combined the complementary strengths of CBCT, which excels in visualizing bone structure, and MRI, which provides detailed soft tissue contrast, to offer a more nuanced understanding of TMJ-OA. This multi-modal approach improves diagnostic assessment by addressing the limitations of single-mode data [44]. In TMJ-OA, multimodal imaging more effectively reflects the complexity of the disease, wherein both bony and soft tissue changes are critical. Second, we systematically explored a variety of imaging techniques, including CBCT sagittal and coronal views as well as MRI sequences such as PD and T2 images. We evaluated all possible combinations of these four distinct image types and identified the optimal combination to enhance diagnostic performance. Our study highlights the value of identifying the most effective imaging combinations and offers insights that can directly inform clinical

practice and future research. Finally, heatmap-based visualization confirmed that the model predictions were guided by clinically interpretable features, thereby enhancing the transparency and explainability of the results[45].

This single-center cohort may limit generalizability. Future work should leverage multi-institutional datasets with balanced demographics, perform external validation with paired comparisons, and report calibration (e.g., Brier/ECE) and decision-curve analyses, ideally incorporating longitudinal imaging for progression modeling. Moreover, multimodal fusion was evaluated only in patients with both CBCT and MRI, which may limit generalizability. Naïve late fusion underperformed for some pairs on the expanded dataset (e.g., CBCT sagittal + MRI T2), possibly reflecting score miscalibration; after fine-tuning the single-mode models and rebuilding the ensemble, the AUROCs improved to 0.808 and 0.839, respectively. The all-views ensemble (CBCT all + MRI all) decreased marginally ($0.826 \rightarrow 0.805$; $\Delta - 0.021$; DeLong $p \approx 0.70$), suggesting redundancy and heterogeneity that could be mitigated by attention-based gating and calibration-aware ensembling. Older age groups and rare subpopulations were under-represented, limiting the precision of the stratified estimates. Finally, we evaluated late fusion only; more sophisticated multi-input or attention-based architectures (including transformer backbones) may further improve performance.

In conclusion, MRI sequence aggregation yielded the best overall discrimination for TMJ-OA, while selective CBCT–MRI fusion provided incremental value over CBCT-only models in selected pairings. Together, these findings highlight the complementary imaging information captured by MRI and CBCT and support further development of acquisition-aware multimodal models. Future external validation and prospective evaluation are needed to clarify their role in imaging-based TMJ-OA assessment.

Data availability

The data used to generate all tables and figures presented in this study are available at [https://github.com/seongwoojang1123/Improved-Diagnostic-Accuracy-of-TMJ-osteoarthritis-through-Machine-Learning-Integration-of-CBCT-MRI].

Code availability

The code for the deep learning framework developed in this study for predicting TMJ-OA diagnosis is available on GitHub at [https://github.com/seongwoojang1123/Improved-Diagnostic-Accuracy-of-TMJ-osteoarthritis-through-Machine-Learning-Integration-of-CBCT-MRI]. The archived version of the code is also available through Zenodo with the following DOI: https://doi.org/10.5281/zenodo.20629318.

Received: 14 April 2026; Accepted: 1 July 2026
Published online: 16 September 2026

References

1. Lee, Y. H., Hong, I. K. & Chun, Y. H. Prediction of painful temporomandibular joint osteoarthritis in juvenile patients using bone scintigraphy. *Clin. Exp. Dent. Res.* **5**, 225–235. https://doi.org/10.1002/cre2.175 (2019).
2. Liu, Q. et al. Initiation and progression of dental-stimulated temporomandibular joints osteoarthritis. *Osteoarthritis Cartil.* **29**, 633–642. https://doi.org/10.1016/j.joca.2020.12.016 (2021).
3. Schmitter, M. et al. Prevalence of clinical and radiographic signs of osteoarthrosis of the temporomandibular joint in an older persons community. *Dentomaxillofac. Radiol.* **39**, 231–234. https://doi.org/10.1259/dmfr/16270943 (2010).
4. Shrivastava, M., Battaglino, R. & Ye, L. A comprehensive review on biomarkers associated with painful temporomandibular disorders. *Int. J. Oral Sci.* **13**, 23. https://doi.org/10.1038/s41368-021-00129-1 (2021).
5. CömertKiliç, S., Kiliç, N. & Sümbüllü, M. A. Temporomandibular joint osteoarthritis: Cone beam computed tomography findings, clinical features, and correlations. *Int. J. Oral Maxillofac. Surg.* **44**, 1268–1274. https://doi.org/10.1016/j.ijom.2015.06.023 (2015).
6. Bag, A. K. et al. Imaging of the temporomandibular joint: An update. *World J. Radiol.* **6**, 567–582. https://doi.org/10.4329/wjr.v6.i8.567 (2014).
7. Mallya, S. M., Ahmad, M., Cohen, J. R., Kaspo, G. & Ramesh, A. Recommendations for imaging of the temporomandibular joint. Position Statement from the American Academy of Oral and Maxillofacial Radiology and the American Academy of Orofacial Pain. *J. Oral Facial Pain Headache* **37**, 7–15. https://doi.org/10.11607/ofph.3268 (2023).
8. Mallya, S. M., Ahmad, M., Cohen, J. R., Kaspo, G. & Ramesh, A. Recommendations for imaging of the temporomandibular joint. Position statement from the American Academy of Oral and Maxillofacial Radiology and the American Academy of Orofacial Pain. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **134**, 639–648. https://doi.org/10.1016/j.oooo.2022.06.007 (2022).
9. ElShennawy, E. M., Hamed, W. M. & Samir, S. M. Diagnostic accuracy of MRI-CBCT fused images in assessment of clinically diagnosed internal derangement of the temporomandibular joint. *Oral Radiol.* **40**, 226–233. https://doi.org/10.1007/s11282-023-00727-1 (2024).
10. Li, M., Jiang, Y., Zhang, Y. & Zhu, H. Medical image analysis using deep learning algorithms. *Front. Public Health* **11**, 1273253. https://doi.org/10.3389/fpubh.2023.1273253 (2023).
11. Kim, S., Lee, Y. H., Noh, Y. K., Park, F. C. & Auh, Q. S. Age-group determination of living individuals using first molar images based on artificial intelligence. *Sci. Rep.* **11**, 1073. https://doi.org/10.1038/s41598-020-80182-8 (2021).
12. Talaat, W. M. et al. An artificial intelligence model for the radiographic diagnosis of osteoarthritis of the temporomandibular joint. *Sci. Rep.* **13**, 15972. https://doi.org/10.1038/s41598-023-43277-6 (2023).
13. Nozawa, M. et al. Can temporomandibular joint osteoarthritis be diagnosed on MRI proton density-weighted images with diagnostic support from the latest deep learning classification models?. *Dentomaxillofac. Radiol.* **54**, 56–63. https://doi.org/10.1093/dmfr/twae040 (2025).
14. Wang, Y. H. et al. Diagnostic efficacy of CBCT, MRI and CBCT-MRI fused images in determining anterior disc displacement and bone changes of temporomandibular joint. *Dentomaxillofac. Radiol.* **51**, 20210286. https://doi.org/10.1259/dmfr.20210286 (2022).
15. Sounderajah, V. et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat. Med.* https://doi.org/10.1038/s41591-025-03953-8 (2025).
16. Tejani, A. S. et al. Updating the checklist for artificial intelligence in medical imaging (CLAIM) for reporting AI research. *Nat. Mach. Intell.* **5**, 950–951. https://doi.org/10.1038/s42256-023-00717-2 (2023).
17. Schiffman, E. et al. Diagnostic criteria for temporomandibular disorders (DC/TMD) for clinical and research applications: Recommendations of the International RDC/TMD Consortium Network* and Orofacial Pain Special Interest Group†. *J. Oral Facial Pain Headache* **28**, 6–27. https://doi.org/10.11607/jop.1151 (2014).

18. Yale, S. H., Allison, B. D. & Hauptfuehrer, J. D. An epidemiological assessment of mandibular condyle morphology. *Oral Surg. Oral Med. Oral Pathol.* **21**, 169–177. https://doi.org/10.1016/0030-4220(66)90238-6 (1966).
19. Brooks, S. L. et al. Imaging of the temporomandibular joint: a position paper of the American Academy of Oral and Maxillofacial Radiology. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* **83**, 609–618. https://doi.org/10.1016/s1079-2104(97)90128-1 (1997).
20. Lee, Y.-H., Jeon, S., Won, J.-H., Auh, Q. S. & Noh, Y.-K. Automatic detection and visualization of temporomandibular joint effusion with deep neural network. *Sci. Rep.* **14**, 18865. https://doi.org/10.1038/s41598-024-69848-9 (2024).
21. Lee, Y.-H. et al. Clinical and MRI markers for acute vs chronic temporomandibular disorders using a machine learning and deep neural networks. *Commun. Med.* **5**, 401. https://doi.org/10.1038/s43856-025-01081-5 (2025).
22. Metz, C. E. Basic principles of ROC analysis. *Semin. Nucl. Med.* **8**, 283–298. https://doi.org/10.1016/s0001-2998(78)80014-2 (1978).
23. Higuchi, K., Chiba, M., Kondo, T. & Echigo, S. The relationship between bone marrow edema and bone changes in the mandibular condyle: A longitudinal study with MR imaging. *Oral Sci. Int.* **10**, 33–39. https://doi.org/10.1016/S1348-8643(12)00053-5 (2013).
24. Li, C. & Zhang, Q. Comparison of magnetic resonance imaging findings in 880 temporomandibular disorder patients of different age groups: A retrospective study. *BMC Oral Health* **22**, 651. https://doi.org/10.1186/s12903-022-02666-5 (2022).
25. Del Grande, F., Farahani, S. J., Carrino, J. A. & Chhabra, A. Bone marrow lesions: A systematic diagnostic approach. *Indian J. Radiol. Imaging* **24**, 279–287. https://doi.org/10.4103/0971-3026.137049 (2014).
26. Xu, L., Hayashi, D., Roemer, F. W., Felson, D. T. & Guermazi, A. Magnetic resonance imaging of subchondral bone marrow lesions in association with osteoarthritis. *Semin. Arthritis Rheum.* **42**, 105–118. https://doi.org/10.1016/j.semarthrit.2012.03.009 (2012).
27. Gopinath, K. et al. Synthetic data in generalizable, learning-based neuroimaging. *Imaging Neurosci. (Camb.)* **2**, 1–22. https://doi.org/10.1162/imag_a_00337 (2024).
28. Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digit. Med.* **3**, 136. https://doi.org/10.1038/s41746-020-00341-z (2020).
29. Baltrusaitis, T., Ahuja, C. & Morency, L. P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443. https://doi.org/10.1109/tpami.2018.2798607 (2019).
30. Wang, W., Tran, D. & Feiszli, M. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 12695–12705 (2020). https://doi.org/10.1109/CVPR42600.2020.01271
31. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, PMLR 70* 1321–1330 (2017).
32. Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M. & Suganthan, P. N. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **115**, 105151. https://doi.org/10.1016/j.engappai.2022.105151 (2022).
33. Naderalvojoud, B. & Hernandez-Boussard, T. Improving machine learning with ensemble learning on observational healthcare data. *AMIA Annu. Symp. Proc.* **2023**, 521–529 (2023).
34. Yang, H. S. et al. Building the model. *Arch. Pathol. Lab. Med.* **147**, 826–836. https://doi.org/10.5858/arpa.2021-0635-RA (2023).
35. Takahashi, S. et al. Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *J. Med. Syst.* **48**, 84. https://doi.org/10.1007/s10916-024-02105-8 (2024).
36. Nagrani, A. et al. Attention bottlenecks for multimodal fusion. *Adv. Neural Inform. Process. Syst.* **34**, 14200–14213 (2021).
37. Xu, L., Chen, J., Qiu, K., Yang, F. & Wu, W. Artificial intelligence for detecting temporomandibular joint osteoarthritis using radiographic image data: A systematic review and meta-analysis of diagnostic test accuracy. *PLoS ONE* **18**, e0288631. https://doi.org/10.1371/journal.pone.0288631 (2023).
38. Li, J. et al. Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. *Med. Image Anal.* **85**, 102762. https://doi.org/10.1016/j.media.2023.102762 (2023).
39. Sannajust, S. et al. Females have greater susceptibility to develop ongoing pain and central sensitization in a rat model of temporomandibular joint pain. *Pain* **160**, 2036–2049. https://doi.org/10.1097/j.pain.0000000000001598 (2019).
40. Kang, S. C. et al. Association between estrogen receptor polymorphism and pain susceptibility in female temporomandibular joint osteoarthritis patients. *Int. J. Oral Maxillofac. Surg.* **36**, 391–394. https://doi.org/10.1016/j.ijom.2006.12.004 (2007).
41. Wu, Z., Lin, Y. & Lin, Y. Impact of age on degenerative joint disease of the temporomandibular joint: A systematic review and meta-analysis. *Medicine* **104**, e41915. https://doi.org/10.1097/md.0000000000041915 (2025).
42. Islam, T., Hafiz, M. S., Jim, J. R., Kabir, M. M. & Mridha, M. F. A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions. *Healthc. Anal.* **5**, 100340. https://doi.org/10.1016/j.health.2024.100340 (2024).
43. Lee, Y.-H., Won, J. H., Kim, S., Auh, Q. S. & Noh, Y.-K. Advantages of deep learning with convolutional neural network in detecting disc displacement of the temporomandibular joint in magnetic resonance imaging. *Sci. Rep.* **12**, 11352. https://doi.org/10.1038/s41598-022-15231-5 (2022).
44. Zubair, M., Hussain, M., Albashrawi, M. A., Bendechache, M. & Owais, M. A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis. *Comput. Methods Programs Biomed.* **272**, 109014. https://doi.org/10.1016/j.cmpb.2025.109014 (2025).
45. Watanabe, A., Ketabi, S., Namdar, K. & Khalvati, F. Improving disease classification performance and explainability of deep learning models in radiology with heatmap generators. *Front. Radiol.* **2**, 991683. https://doi.org/10.3389/fradi.2022.991683 (2022).

Acknowledgements

The authors extend their special thanks to Dr. Kyoungsu Kim of Gangbuk Baeksang Dental Hospital.

Author contributions

Writing and original draft preparation, Y-HL and S-WJ; conceptualization, Y-HL, S-WJ, and SJ; methodology, Y-HL, S-WJ, and SJ; software, Y-HL, S-WJ, and SJ; validation and formal analysis, Y-HL, Q-SA, and S-WJ; investigation, Y-HL, Q-SA, and S-WJ; resources, Y-HL and Q-SA; data curation, Y-HL and S-WJ; writing, review, and editing, Y-HL, S-WJ, SJ, Q-SA, AC, and Y-KN; visualization, Y-HL, S-WJ, and SJ; supervision, Y-HL, FG, AC, and Y-KN; project administration, Y-HL; and funding acquisition, Y-HL and Y-KN. All the authors contributed to and approved the submission of the manuscript.

Funding

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) (RS-2024-00421203); an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by MSIT (IITP-2021-0-02068, RS-2020-II201373, RS-2023-00220628); and a grant from Kyung Hee University in 2025 (KHU-20251299).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Y.-H.L. or Y.-K.N.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

© The Author(s) 2026