Journal of Dentistry 166 (2026) 106322

Contents lists available at ScienceDirect

Journal of Dentistry

journal homepage: www.elsevier.com/locate/jdent

M⁴4TMD: A multimodal, multi-task deep learning framework for comprehensive assessment of TMD-related abnormalities

Xinrui Lang [a], Rundong Zhang [b], Zhouhang Yuan [c,d], Zheqi Lyu [c,e,*], Jiawei Wang [f], Bo Qiao [b], Yanzhen Zhang [b,**], Zhengxing Huang [c], Fan Yang [a,***]

[a] Center for Plastic & Reconstructive Surgery, Department of Stomatology, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, Zhejiang, China
[b] Department of General Dentistry, The Second Affiliated Hospital of Zhejiang University, Hangzhou, Zhejiang, China
[c] College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang, China
[d] Sir Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China
[e] Weill Cornell Medicine, Cornell University, NY, USA
[f] School of Computing, National University of Singapore, Singapore, Singapore

ARTICLE INFO

Keywords:
Deep learning
Temporomandibular disorders
Magnetic resonance imaging
Multi-task
Multimodal

ABSTRACT

Objective: Existing deep learning (DL) approaches for assessing temporomandibular disorders (TMD) are limited by underutilization of magnetic resonance imaging (MRI) in some tasks, a narrow focus on single-task detection, and predominant reliance on unimodal data. This study proposes a multimodal DL framework to address these issues.

Methods: We collected 12,690 MRI slices and clinical data from 765 participants (1410 temporomandibular joints), with each joint annotated for degenerative joint disease (DJD), anterior disc displacement (ADD), and effusion. We developed M⁴4TMD, utilizing multimodal data including multi-sequence and multi-slice MRI with clinical data, for concurrent assessment of DJD, ADD, and effusion. Performance was benchmarked against three recent DL methods and four clinicians with varying expertise across internal, temporal, and external test sets; assessments included generalization and visual interpretability experiments.

Results: Built upon ResNet50, M⁴4TMD exhibited superior internal test performance (ROC-AUC: 0.831, 0.913, and 0.961), surpassing prior methods. The accuracy of M⁴4TMD for three abnormalities was superior to that of junior dentists and comparable to that of two senior dentists (10 and 20 years experience): DJD (74.9% vs. 74.9%/72.5%; $P > 0.05$, > 0.05), ADD (78.2% vs. 71.1%/75.8%; $P > 0.05$, > 0.05), and effusion (90.5% vs. 88.6%/79.6%; $P > 0.05$, < 0.01). Strong robustness and interpretability were validated through generalization and visual interpretability experiments.

Conclusion: The M⁴4TMD framework enables concurrent assessment of TMD-related abnormalities by integrating multimodal MRI and clinical data, exhibiting assessment performance comparable to senior dentists and demonstrating excellent robustness.

Clinical Significance: The M⁴4TMD framework represents a critical step toward advancing DL-based TMD diagnosis in clinical practice.

* Corresponding author at: Weill Cornell Medicine, Cornell University, 575 Lexington Ave., New York, NY 10022, USA.
** Corresponding author at: Department of General Dentistry, The Second Affiliated Hospital of Zhejiang University, 88 Jiefang Road, Shangcheng District, Hangzhou, Zhejiang 310009, China.
*** Corresponding author at: Center for Plastic & Reconstructive Surgery, Department of Stomatology, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, 158 Shangtang Road, Gongshu District, Hangzhou, Zhejiang 310014, China.
E-mail addresses: zheqilyu@gmail.com, zhl4020@med.cornell.edu, zheqilv@zju.edu.cn (Z. Lyu), 2191004@zju.edu.cn (Y. Zhang), yangfan@hmc.edu.cn (F. Yang).

https://doi.org/10.1016/j.jdent.2025.106322
Received 11 September 2025; Received in revised form 22 December 2025; Accepted 27 December 2025
Available online 27 December 2025

1. Introduction

Temporomandibular disorders (TMD) are a group of conditions affecting the temporomandibular joint (TMJ) complex that cause oral and maxillofacial pain and dysfunction [1], and are present in up to one-third of adults [2]. TMD is the second-most common reason for dental visits after dental pain [3], and assessment of the TMJ is essential before performing dental treatments such as prosthodontic or orthodontic procedures [4]. Early and accurate diagnosis is critical to prevent progression from acute to chronic TMD [3,5].

The Diagnostic Criteria for TMD (DC/TMD) cover 12 common classifications, among which degenerative joint disease (DJD) and anterior disc displacement (ADD) are the most prevalent [1]. ADD is further divided into two types: with reduction (ADDWR) and without reduction (ADDWoR) [6]. Additionally, TMJ effusion, which indicates potential inflammation, is a key imaging feature for assessment of TMD [7,8].

The diagnosis of TMD primarily relies on clinical evaluation, including assessment of medical history and physical examination [1,9]. However, accurate diagnosis of TMD without imaging assessments, a critical adjunct in confirming the condition, is challenging [1,10,11]. Computed tomography (CT) and cone-beam computed tomography (CBCT) provide excellent visualization of bony structures [10,12], while magnetic resonance imaging (MRI) is preferred for evaluating soft tissues [11]. MRI is noninvasive, free of ionizing radiation, and suitable for repeated examination and follow-up assessments [13,14]. It is the gold standard for assessing disc position and joint effusion and can provide some information regarding the TMJ bony structures [15]. However, its diagnostic value for bony abnormalities remains controversial [11,16], and in clinical practice, additional CT or CBCT exams are often required to further diagnose DJD [15].

Recent advances in deep learning (DL) have shown promise in automating the diagnosis of TMD from imaging assessments [17–21]. However, DL research based on TMJ MRI faces several challenges: (1) Underutilization of MRI for certain diagnostic tasks: Although CT and CBCT are widely used inputs in DL studies on TMD [19,20,22], the potential of MRI remains largely untapped. This is especially true for the detection of DJD, where CT or CBCT is the conventional modality. Thus, attempts to design DL models for DJD detection using only MRI data are virtually nonexistent, representing a significant and underexplored research gap. Similarly, the use of DL for detecting joint effusion using MRI is also infrequent. (2) Narrow focus on single-task detection: The majority of existing DL models are trained to identify only a single pathological form of TMD [22], such as disc displacement or osseous changes. This single-task paradigm, however, fails to capture the complex, multifactorial nature of TMD, where multiple conditions frequently coexist and interact [1]. Consequently, these models provide an incomplete diagnostic assessment, which significantly limits their real-world clinical utility. (3) Predominant reliance on single-modality data: The existing research on the use of AI for assessing TMD has almost exclusively focused on unimodal inputs [22], treating radiological images and clinical data as separate, isolated sources. In addition to failing to reflect clinical practice, this siloed approach also neglects the substantial performance gains achievable when AI models are trained on multimodal data [1,23].

To address these limitations, this study aimed to develop a comprehensive DL framework for multimodal, multi-task assessment of TMD-related abnormalities on the basis of TMJ MRI and clinical data (Fig. 1). Therefore, we propose M^{4}4TMD, a multi-task framework that concurrently assesses multiple abnormalities in TMD by fusing MRI and clinical data (in this context, "framework" refers to a generalized DL method applicable to multiple models). To enable the development of this framework and facilitate its systematic evaluation, the model was trained and tested on a comprehensive dataset from 765 participants (1410 TMJs), which included 12,690 annotated MRI slices and paired clinical data. The M^{4}4TMD framework utilizes a unified backbone to capture shared knowledge from multimodal data sources (multi-sequence and multi-slice MRI data and clinical information) and separate classifiers for task-specific features. This multi-task architecture

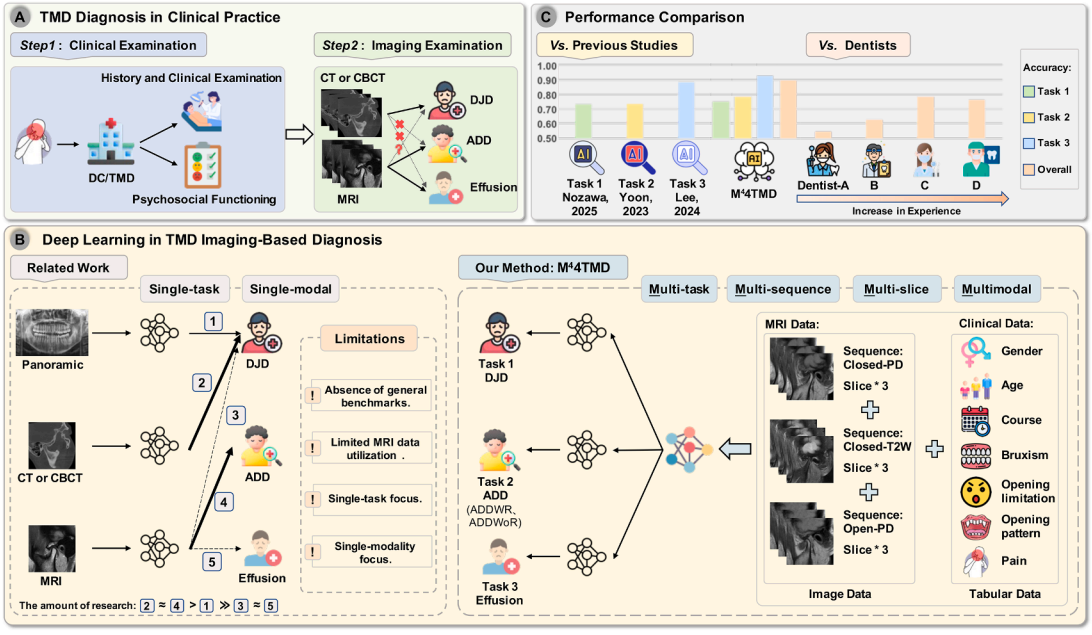


Fig. 1. Overview of TMD assessment and the proposed M^{4}4TMD framework. A: TMD diagnostic process in clinical: Typically involves two steps—step 1: clinical examination; step 2: imaging examination (CT, CBCT, and / or MRI) used for definitive diagnosis. B: Related work and our method: Related work primarily focus on using single image modality (such as panoramic, CT /CBCT, or MRI) to assess specific TMD conditions. Numbers 1–5 correspond to related work, with thicker arrows (above or below) indicating more existing DL research in this field. In contrast, M^{4}4TMD integrates multi-slice, multi-sequence MRI data (closed-PD, closed-T2W, open-PD) with multimodal inputs—comprising both imaging and tabular clinical data (e.g., gender, age, disease course, bruxism)—to perform multi-task prediction of DJD, ADD, and TMJ effusion. C: M^{4}4TMD performance: Assessment accuracy of M^{4}4TMD compared with previous studies and dental practitioners of varying experience levels, demonstrating its superior performance in TMD diagnosis.

allowed concurrent assessment of DJD, ADD (distinguishing ADDWR and ADDWoR), and joint effusion, yielding performance comparable to that of experienced clinicians and surpassing existing MRI-based DL studies [17,24,25]. Notably, our study developed a method for detecting DJD from MRI, advancing beyond the simple application of a generic convolutional neural network (CNN) in the sole previous study [25]. In addition, interpretability analysis using Grad-CAM and Permutation Importance (PI) were employed to visualize the framework's regions of attention and quantify the contribution of clinical features during prediction. The proposed M^{4}4TMD framework represents a critical step toward advancing DL-based comprehensive assessment of TMD in clinical practice.

2. Materials and methods

This retrospective diagnostic study was approved by the Institutional Review Board of the Second Affiliated Hospital of Zhejiang University School of Medicine (I20241032) and conducted in accordance with the Declaration of Helsinki. The study is reported in compliance with the Standards for Reporting Diagnostic Accuracy guidelines (STARD) and the Checklist for Artificial Intelligence in Dental Research [26,27].

2.1. Data collection and annotation

This study collected clinical information and MRI scans from a total of 803 individuals who visited the TMJ Clinic of the Department of General Dentistry at the Second Affiliated Hospital of Zhejiang University School of Medicine. The data were collected from the Binjiang campus between January 1, 2019, and September 30, 2024, and from the Jiefang Road campus between January 1, 2019, and October 31, 2025. These individuals primarily sought care for suspected or confirmed symptoms related to TMD, such as joint or muscular pain, joint noises, or limited mouth opening, as well as pre-orthodontic evaluation of the TMJ in some cases.

The inclusion criteria were as follows:

(1) Aged $\geq$ 16 years.

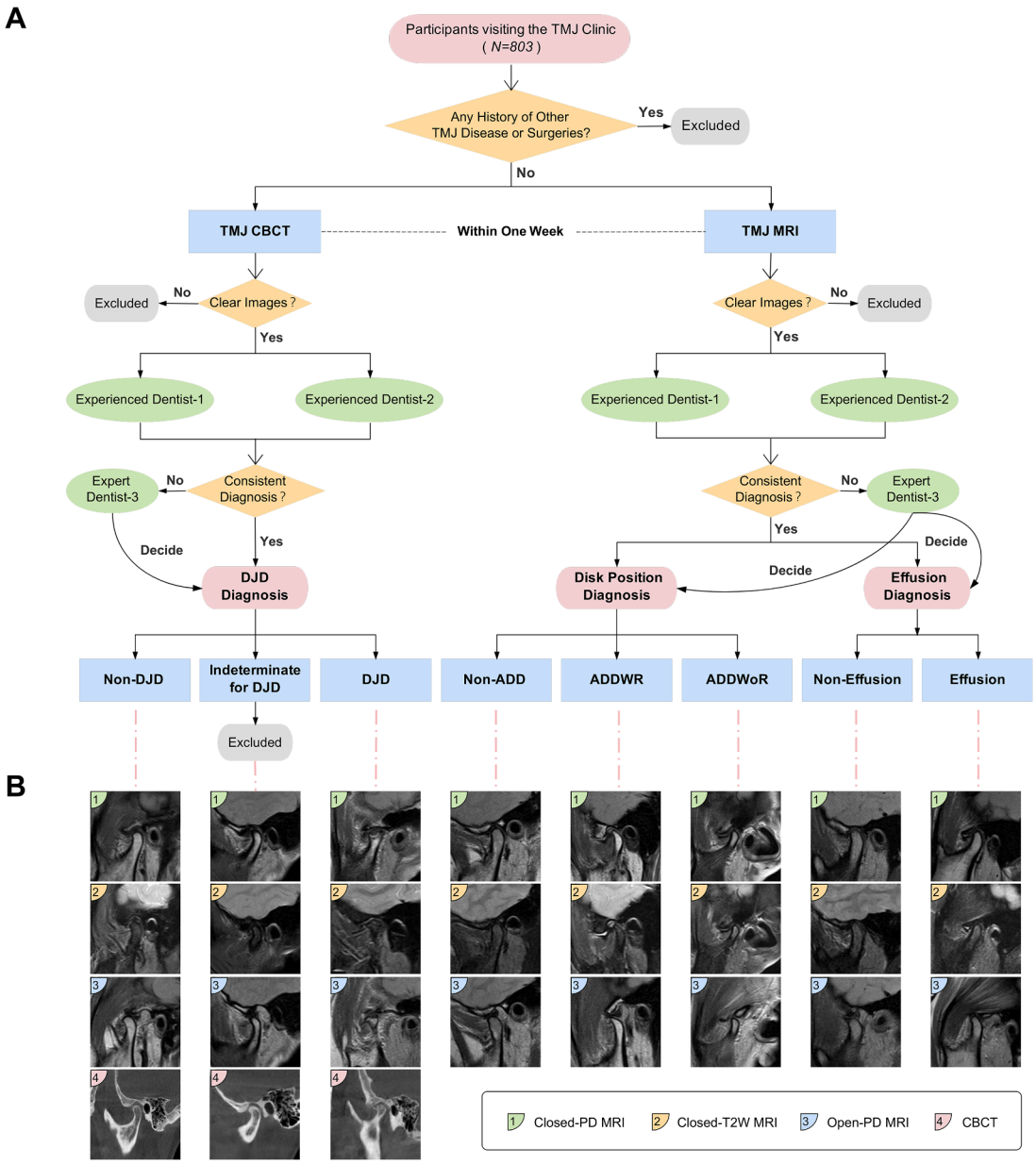


Fig. 2. Overview of data collection and annotation. **A:** Process of data inclusion and exclusion. **B:** Representative imaging examples for each target abnormality.

(2) Able to provide a clear description of their medical history and cooperate with clinical examinations.
(3) Underwent both TMJ MRI and CBCT scans at the hospital within one week.

The exclusion criteria were as follows:

(1) MRI or CBCT scans of low quality, including blurred, artifact-laden, underexposed, or very low-contrast resolution images.
(2) Cases involving TMJ trauma, rheumatism, tumors, dislocation, or ankylosis.
(3) Cases classified as "Indeterminate for DJD" according to the DC/TMD criteria, specifically those showing flattening and/or sclerosis features on CBCT [1,11].

After applying the inclusion and exclusion criteria (Fig. 2A), 765 participants were retained, with a total of 1410 TMJs. For each joint, three closed-mouth proton density-weighted imaging slices (closed-PD), three closed-mouth T2-weighted imaging slices (closed-T2W), and three open-mouth PD-weighted imaging slices (open-PD) were selected, totaling 12,690 MR images. This selection was performed by two experienced dentists with 4–6 years of clinical experience in the management and research of TMD. For each sequence, the slice that showed the most complete visualization of the articular disc and condyle was selected, typically the central oblique sagittal slice through the condylar head and one slice anterior and one posterior to it. This three-slice strategy was designed to comprehensively capture key features such as disc position, joint effusion, and bone contours. To ensure consistency, both annotators received standardized training based on established anatomical landmarks and were supervised by a senior TMJ expert dentist (a chief dentist with over 30 years of experience in TMD diagnosis and management). Disagreements in slice selection were resolved through discussion and consensus. Additionally, the clinical data collected at the time of the first visit for each patient included information regarding "gender", "age", "course", "bruxism", "opening limitation", "opening pattern", and "pain".

Before commencement of the annotation work, the two experienced dentists underwent a standardized consistency training session under the guidance of the TMJ expert. The two annotators strictly adhered to the DC/TMD and utilized each patient's clinical records and imaging data during the assessment. After undergoing the training, the annotators independently assessed each TMJ and assigned diagnostic labels. The assessment for ADD and effusion were based on MR images, while the detection for DJD was determined using the corresponding CBCT images. The results of inter-rater reliability analysis confirmed high agreement across all tasks (overall Cohen's Kappa = 0.86; DJD: 0.87, ADD: 0.82, effusion: 0.91). When the diagnostic labels assigned by both annotators were in agreement, that label was adopted as the final ground truth. For cases where the annotators disagreed, final adjudication was performed by the expert, and the expert's final decision was adopted as the definitive ground truth for model training and testing. This process is illustrated in Fig. 2A.

2.2. Data preprocessing and splitting

Due to differences in the acquisition matrix size of the MRI scans, each image was first resized to 512 × 512 pixels, and then centrally cropped to 256 × 256 pixels to obtain the region of interest (ROI) for image input. Subsequently, the seven clinical data points corresponding to each side of the joint were converted into tabular data. Among these, "gender", "bruxism", "opening limitation", and "pain" were binary variables, "opening pattern" was a four-class variable, while "age" and "course" were continuous variables. All personally identifiable information was removed to ensure patient confidentiality.

Data were partitioned on the basis of the acquisition site (campus) and the MRI scanner used. For individuals who first visited the Binjiang campus, MRI scans were acquired using a 1.5T Magnetom Avanto MRI scanner (Siemens AG, Erlangen, Germany) with an 8 × 8 cm surface coil (repetition time [TR] / echo time [TE] for PD: 2000–2020 / 29 ms, TR / TE for T2W: 2730 / 58 ms; field of view [FOV] for PD: 140 × 140 mm, FOV for T2W: 150 × 150 mm; slice thickness: 2.5 mm), which, along with their clinical data, formed the BJ Dataset (578 participants, 1059 joints, 9531 MR images). In parallel, for individuals who first visited the Jiefang Road campus, MRI scans were acquired using a 1.5T Signa Voyager scanner (General Electric, Boston, USA) with a 7 × 7 cm surface coil (PD: TR / TE: 2000 / 21.52 ms; T2W: 3900 / 84.03 ms; FOV: 120 × 120 mm; slice thickness: 2.0 mm), which, together with their clinical information, constituted the JF Dataset (187 participants, 351 joints, 3159 MR images). Detailed parameters of the MRI scanners used at each campus are provided in Supplementary Tables 13–14.

The BJ Dataset was utilized for training and evaluating the models derived from our framework, and for temporal generalization assessment. As illustrated in Supplementary Fig. 1, it was first randomly split into a training set (80 %) and an internal test set (20 %) for model evaluation. To simulate real-world distributional shifts and assess the model's temporal generalization, the BJ Dataset was also divided chronologically: the earliest 80 % of MRI scans constituted the temporal training set, while the most recent 20 % formed the temporal test set. The JF Dataset served exclusively as the external test set. Importantly, no participant overlap was noted between any training and test sets, ensuring independent and unbiased evaluation.

2.3. $M^4$4TMD framework design

2.3.1. Framework architecture and model implementation

We developed a multi-task learning framework, $M^4$4TMD, to jointly assess three key abnormalities in TMD: DJD (Task 1), ADD with/without reduction (Task 2), and joint effusion (Task 3). This architecture accommodated multi-sequence, multi-slice MRI inputs and clinical data while modeling inter-task dependencies. For the implementation in this study, the $M^4$4TMD framework was instantiated using a ResNet50 backbone pretrained on ImageNet to leverage transfer learning, which was chosen for its robust feature extraction capabilities [28]. In this model, independent ResNet50 branches with shared weights processed closed-PD, closed-T2W, and open-PD sequences, and the extracted features were pooled, flattened, and integrated into a unified representation for multi-task classification (Fig. 3).

2.3.2. Symbol definitions

We present the general symbol definitions here. Model: We used Ω as the backbone, also known as the feature extractor, and $\mathscr{T}$ as the classifier. The DL model was defined as $\mathscr{M} = \{\Omega, \mathscr{T}\}$. Data: The image and tabular modalities were represented by I and T, respectively. We defined the dataset as $\{X, Y\}$, where $X = \{I, T\} = \{x_j\}_{j=1}^{n} = \left\{x_j^I, x_j^T\right\}_{j=1}^{n}$, and n denoted the number of samples in the dataset, $Y = \left\{y_j\right\}_{j=1}^{n}$ and y_j means the annotation for the x_j. $\hat{y}_j$ represents the predicted value for the x_j by the model. For simplicity, we further defined $I = \left\{x_j^I\right\}_{j=1}^{n}$ and $T = \left\{x_j^T\right\}_{j=1}^{n}$, and in the following discussion, we only used I and T. In other words, I represents all the image modality data contained in a sample, and T represents all the tabular modality data contained in a sample. "Concat(·)" represents concatenating two or more features together.

2.3.3. Multi-sequence and multi-slice modeling

Each image was processed by a preprocessing function P, which resized the image to 512 × 512 pixels and then performed center-cropping to 256 × 256 pixels to focus on the ROI. Slice-level features were extracted and averaged within each sequence, resulting in three 2048-dimensional vectors concatenated into a 6144-dimensional rep-

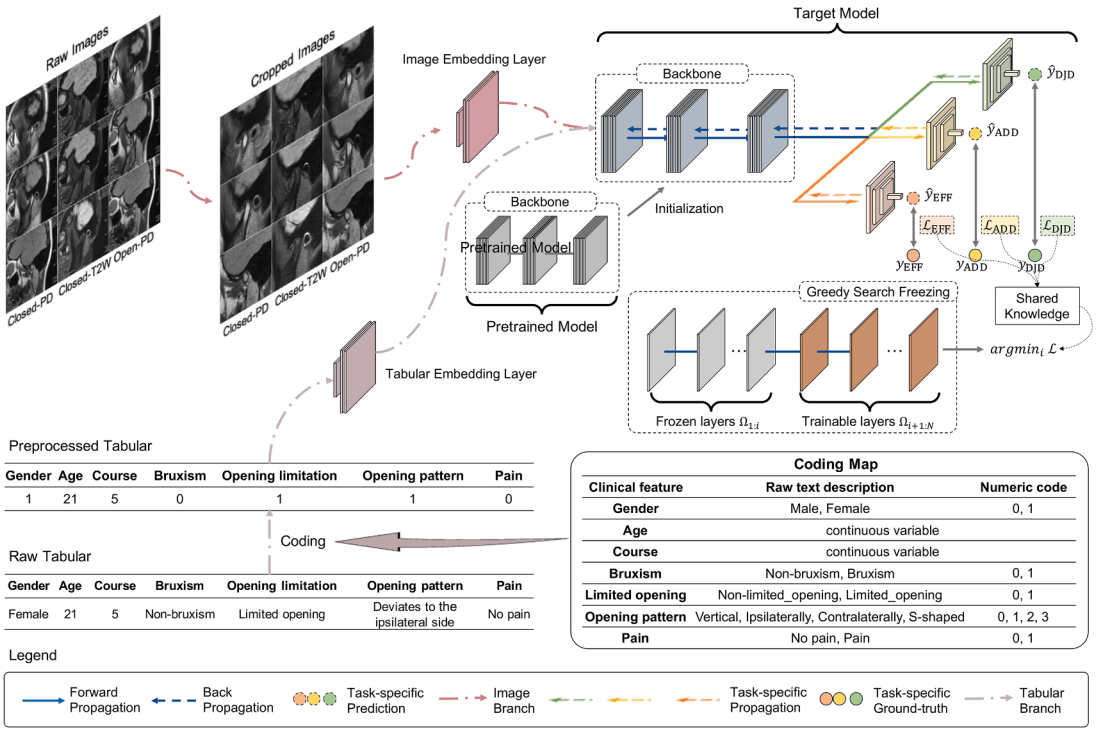


Fig. 3. Overview of our proposed M[4]4TMD framework. The M[4]4TMD framework is a multimodal, multi-task learning approach designed to predict multiple TMD-related abnormalities. The input consisted of multi-sequence TMJ MR multi-slices, including three closed-PD slices, three closed-T2W slices, and three open-PD slices, which were combined into a multi-parametric MRI series, along with corresponding clinical tabular data.

resentation. The fused feature, serving as an image feature, can be passed to the classifier for direct prediction or further fused with features from other modalities.

$$\begin{cases} \underbrace{\mathscr{S} = \{\text{closed} - \text{PD}, \text{closed} - \text{T2W}, \text{open} - \text{PD}\}}_{\textbf{set of all MRI features}}, \\ \underbrace{\boldsymbol{e}_s = \frac{1}{3}\sum_{k=1}^{3} \boldsymbol{\Omega}^{\boldsymbol{I}}\left(\mathscr{P}^{\boldsymbol{I}}\left(\boldsymbol{I}_s^{\boldsymbol{k}}\right)\right), \boldsymbol{s} \in \mathscr{S}}_{\textbf{slice-level multi-sequence and multi-slice modeling for each s}}, \\ \underbrace{\boldsymbol{e}^{\boldsymbol{I}} = \textbf{Concat}\left(\{\boldsymbol{e}_s | \boldsymbol{s} \in \boldsymbol{S}\}\right)}_{\text{obtain image modality features}}. \end{cases} \tag{1}$$

2.3.4. Multimodal integration

To incorporate clinical data, tabular features e^T were embedded by a multilayer perceptron and combined with MRI-derived features. Tabular features are extracted from the seven clinical features ("gender", "age", "course", "bruxism", "opening limitation", "opening pattern", and "pain") by $e^T = \Omega^T(\mathscr{P}^T(T))$, where $\mathscr{P}^T$ is the preprocessing function. The feature extractor Ω^T was designed as a 3-layer MLP with a dimension mapping of 7→64→128→256, yielding a 256-dimensional clinical embedding. The modality-specific classifiers $\mathscr{F}^I$ and $\mathscr{F}^T$ were designed to make predictions based on e^I and e^T, respectively. Subsequently, two fusion strategies were explored: (1) feature fusion, which concatenated image and clinical embeddings to form a unified 6400-dimensional vector (comprising 6144 imaging features and 256 clinical features) for classification, and (2) decision fusion, which used modality-specific classifiers whose outputs were combined by element-wise addition. The proposed model is summarized as follows:

$$\widehat{\boldsymbol{y}} = \begin{cases} \mathscr{F}\left(\textbf{Concat}\left(\boldsymbol{e}^{\boldsymbol{I}}, \boldsymbol{e}^{\boldsymbol{T}}\right)\right), & (\textbf{Feature Fusion}), \\ \mathscr{F}^{\boldsymbol{I}}\left(\boldsymbol{e}^{\boldsymbol{I}}\right) + \mathscr{F}^{\boldsymbol{T}}\left(\boldsymbol{e}^{\boldsymbol{T}}\right), & (\textbf{Decision Fusion}). \end{cases} \tag{2}$$

2.3.5. Multi-task optimization

A shared feature encoder was applied with task-specific output layers. Focal loss was used for the imbalanced DJD classification task [29], while cross-entropy was employed for the remaining tasks. The joint loss function included regularization through L2 wt decay. Training was performed with the Adam optimizer and Dropout for generalization.

$$\begin{cases} \underbrace{\mathscr{T} = \{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3\} = \{\text{DJD}, \text{ADD}, \text{Effusion}\}}_{\text{set of tasks}}, \\ \mathscr{L}_{\text{task}} = \underbrace{\boldsymbol{\lambda} \parallel \boldsymbol{\theta} \parallel_2^2}_{\text{regularization}} + \underbrace{\sum_{\boldsymbol{\tau} \in \mathscr{T}} \begin{cases} \boldsymbol{l}_{\text{Focal}}(\widehat{\boldsymbol{y}}_{\boldsymbol{\tau}}, \boldsymbol{y}_{\boldsymbol{\tau}}), & \boldsymbol{\tau} = \text{DJD} \\ \boldsymbol{l}_{\text{CE}}(\widehat{\boldsymbol{y}}_{\boldsymbol{\tau}}, \boldsymbol{y}_{\boldsymbol{\tau}}), & \boldsymbol{\tau} \in \{\text{ADD}, \text{Effusion}\} \end{cases}}_{\text{task-specific loss functions}}. \end{cases} \tag{3}$$

2.3.6. Greedy search freezing

To optimize transfer learning, a greedy search freezing strategy was introduced for the ImageNet-pretrained ResNet50 backbone. This approach utilizes a ResNet50 pretrained on ImageNet as the base feature extractor and defines the frozen layers $\Omega_{1:i}$ as the first i layers of Ω, the parameters of $\Omega_{1:i}$ as fixed, and the subsequent layers $\Omega_{i+1:N}$ as trainable. For each task, a unified classifier $\mathscr{T}$ was employed for prediction, and the cross-entropy loss function ℓ was used to compute the task-specific loss. Finally, by iterating over different numbers of frozen layers i, the optimal number i^* is selected as the one that maximized the overall ROC-AUC across all tasks. This process was expressed by the following system of equations:

$$\begin{cases} \widehat{\boldsymbol{y}}_{\boldsymbol{\tau}} = \mathscr{T}\left(\boldsymbol{\Omega}_{\boldsymbol{i}+1:\boldsymbol{N}}\left(\boldsymbol{\Omega}_{1:\boldsymbol{i}}(\boldsymbol{x})\right)\right), \boldsymbol{\tau} \in \{\textbf{DJD}, \textbf{ADD}, \textbf{Effusion}\}, \\ \mathscr{L}_{\boldsymbol{\tau}} = \ell\left(\widehat{\boldsymbol{y}}_{\boldsymbol{\tau}}, \boldsymbol{y}_{\boldsymbol{\tau}}\right), \\ \boldsymbol{i}^* = \underset{\boldsymbol{i}}{\textbf{argmax}} \sum\nolimits_{\boldsymbol{\tau} \in \{\text{DJD}, \text{ADD}, \text{Effusion}\}} \textbf{ROC} - \textbf{AUC}_{\boldsymbol{\tau}}\left(\widehat{\boldsymbol{y}}_{\boldsymbol{\tau}}, \boldsymbol{y}_{\boldsymbol{\tau}}\right). \end{cases} \tag{4}$$

2.4. Evaluation

2.4.1. Comparison with previous studies

Given the absence of multi-task deep learning studies for TMD, M^{4}4TMD was benchmarked individually against representative studies for each of the framework's three tasks; the studies were chosen to reflect the current state of the art. For DJD (Task 1), the sole existing MRI-based DL study by Nozawa et al. [25] was selected, which applied a ResNet18 model for detection without designing a specialized architecture. Similarly, for joint effusion (Task 3), the only available method by Lee et al. [24] was implemented, which utilized a fine-tuning VGG16 model. For ADD diagnosis (Task 2), which has been studied more extensively, the advanced ResNet50-based fusion strategy by Yoon et al. [17] was replicated. This selection highlights the limited research on the detection of DJD and effusion using MRI. Since the original codes and datasets were not publicly available, these three methods were replicated to the best of our ability on our BJ Dataset for a fair evaluation against our M^{4}4TMD framework.

2.4.2. Comparison with clinicians

To compare the performance of M^{4}4TMD with clinicians of varying years of experience and expertise in the management of TMD, four dentists—two juniors (dentist-A and dentist-B) and two seniors (dentist-C and dentist-D, with over 10 and 20 years of TMD experience, respectively)—independently evaluated all three test sets. These dentists were not involved in the model development or data annotation. Under blinded conditions, they assessed DJD, ADD, and effusion for each TMJ and had access to the same MR images and clinical data provided to the model. The dentists were unaware their performance was being benchmarked against an AI framework, ensuring a fair and unbiased comparison.

2.4.3. Generalization and adaptability evaluation

The M^{4}4TMD framework's robustness to distributional shifts was evaluated by assessing its generalization performance on temporal and external test sets. To assess the adaptability of the framework, it was implemented using three distinct and representative CNN backbones in addition to ResNet50 [30,31]. VGG16 was selected for its simple sequential architecture [32], DenseNet121 for its feature reuse through dense connections [33], and Inception-V3 for its multi-scale feature extraction capabilities [34]. This selection allowed validation of the framework's robustness and versatility across different architectural designs.

2.4.4. Quantitative interpretability evaluation

To enhance the transparency of the framework, gradient-weighted class activation mapping (Grad-CAM) was employed to visualize the discriminative regions influencing the model's predictions [35]. This technique was applied to generate heatmaps that highlight the anatomical areas most salient for multi-task assessment of DJD, ADD, and joint effusion.

To quantitatively assess the reliance of the framework on clinically relevant anatomical features during diagnosis, a ROI alignment analysis was conducted on a randomly selected 25 % subset of the internal test set (53 TMJs, 477 MR images). Pixel-level masks for three key anatomical structures: the mandibular condylar head (Task 1: DJD), articular disc (Task 2: ADD), and joint space / effusion (Task 3: effusion), were manually annotated by an experienced dentist and subsequently reviewed and calibrated by the TMJ expert. Following the generation of task-specific Grad-CAM heatmaps, an Adaptive Percentile Thresholding Strategy was adopted to mitigate inherent noise; specifically, the 96th percentile of activation intensity within each heatmap served as the threshold, converting pixels exceeding this value (the top 4 % salient regions) into binary "Model Attention Masks". Finally, the Intersection over Union (IoU) between these attention masks and the expert-annotated ROIs was calculated, alongside the Pointing Game accuracy, which defined a "Hit" if the pixel with the maximum activation intensity fell within the target anatomical ROI. Furthermore, to rule out the possibility of the model leveraging background noise for shortcuts, an occlusion sensitivity experiment based on counterfactual reasoning was performed, the details of which are provided in the Supplementary Materials.

2.4.5. Fine-grained annotation and stratified error analysis

To facilitate a fine-grained error analysis and characterize the diagnostic failure modes of M^{4}4TMD, the cases within the internal test set underwent an additional layer of detailed annotation. An experienced dentist reviewed these cases to classify specific pathological features, with the results subsequently verified and confirmed by the senior TMJ expert. Specifically, for the DJD task, osseous changes were stratified into specific radiographic signs (erosion, osteophyte, sclerosis, and subchondral cyst) and graded for severity, utilizing the corresponding CBCT scans as the reference standard and adhering to the imaging criteria proposed by Ahmad et al. [11]. For the ADD task, articular disc morphology was categorized based on MRI following RDC/TMD guidelines to distinguish between normal biconcave shapes and deformed variants (e.g., lengthened, or contracture), alongside a classification based on effusion severity [11]. These detailed sub-classifications served as the ground truth for the stratified performance evaluation.

2.4.6. Permutation importance analysis of clinical features

To elucidate the contribution of clinical variables to the model's decision-making, a Permutation Importance analysis on both internal and external test sets was performed. This method measures feature reliance by calculating the degradation in ROC-AUC after randomly shuffling a specific clinical feature while keeping image embeddings and other variables fixed. The importance score I_j for feature j is defined as:

$$I_j = \mathrm{ROC-AUC}_{\mathrm{base}} - \mathrm{ROC-AUC}_{\mathrm{perm,j}}. \quad (5)$$

where $\mathrm{ROC-AUC}_{\mathrm{base}}$ denotes the baseline performance and $\mathrm{ROC-AUC}_{\mathrm{perm,j}}$ represents the performance after permuting feature j. We repeated this procedure 5 times for each of the seven clinical variables across all three tasks to ensure statistical robustness.

2.5. Statistical analysis

Performance was evaluated using accuracy, sensitivity, specificity, precision, recall, F1-score, Brier score, the precision-recall area under the curve (PR-AUC), and the area under the receiver operating characteristic curve (ROC-AUC) with 95 % confidence intervals (CIs). The CIs for accuracy, sensitivity, and specificity were computed using the Wilson Score method, while the CIs for ROC-AUC were determined by bootstrapping. To compare the performance between the framework and clinicians, the McNemar test was used for binary classification tasks (Task 1, Task 3), whereas the Stuart-Maxwell test was employed for the three-class ADD task (Task 2). To control for the family-wise error rate across multiple comparisons (M^{4}4TMD vs. four dentists), the Holm-Bonferroni correction was applied. All statistical analyses were conducted using Python 3.11.3 (Wilmington, Delaware, USA) with the Statsmodels (version 0.13.5) and Scikit-learn (version 1.2.2) packages. Statistical significance was determined based on an adjusted two-sided P value of less than 0.05.

3. Results

After applying the inclusion and exclusion criteria, a total of 765 participants were included in the study. The primary cohort from the Binjiang campus consisted of 578 participants (452 female, 126 male; mean age, 29.4 ± 11.7 years), yielding 1059 TMJs and 9531 MR images and forming the basis for the training, internal, and temporal test sets.

The external test cohort from the Jiefang Road campus included 187 participants (160 female, 27 male; mean age, 28.4 ± 11.1 years), totaling 351 TMJs and 3159 images. Table 1 summarizes the demographic and clinical characteristics of the joints across all training and test sets. Further details on dataset construction and diagnostic distributions are provided in Supplementary Fig. 1–3.

3.1. Internal performance evaluation

3.1.1. Comparison with prior studies

When benchmarked against three replicated single-task models from recent MRI-based studies [17,24,25], the multi-task M^{4}4TMD framework demonstrated superior performance across all assessment tasks (Table 2). For DJD (Task 1), M^{4}4TMD achieved a ROC-AUC of 0.831 (vs. 0.793) and an accuracy of 74.9 %. For ADD (Task 2), it reached a ROC-AUC of 0.913 (vs. 0.896), with class-specific sensitivities ranging from 68.9 % to 91.5 % and specificities ranging from 82.1 % to 92.3 %. For joint effusion (Task 3), the framework achieved a ROC-AUC of 0.961 (vs. 0.937) and an accuracy of 90.5 %. Detailed performance metrics for all models are presented in Supplementary Table 1.

3.1.2. Comparison with clinicians

For brevity, unless otherwise specified, M^{4}4TMD refers to the ResNet50-based framework with multimodal feature fusion in the following results. Statistical significance was determined based on an adjusted two-sided P value of less than 0.05; hereafter, all reported P values refer to the adjusted values. When benchmarked against four dentists on the internal test set, M^{4}4TMD consistently achieved the highest numerical accuracy across all three tasks (Table 3). Specifically, it demonstrated significantly higher accuracy than dentist-B across all tasks ($P < 0.05$) and dentist-A in Task 1 and Task 3 ($P < 0.01$), though the difference with dentist-A in Task 2 was not statistically significant (78.2 % vs. 44.6 %, $P = 0.193$). Regarding senior clinicians, in Task 1, M^{4}4TMD performed comparably to both senior dentist-C (over 10 years of experience in TMD assessment) and senior dentist-D (over 20 years' experience) (74.9 % vs. 74.9 % vs. 72.5 %, $P > 0.05$). In Task 2, its performance was statistically comparable to both dentist-C and dentist-D (78.2 % vs. 71.1 % vs. 75.8 %, $P > 0.05$). However, in Task 3, M^{4}4TMD performed comparably to senior dentist-C (90.5 % vs. 88.6 %, $P > 0.05$) but significantly better than senior dentist-D (90.5 % vs. 79.6 %, $P < 0.01$). Fig. 4A illustrates the confusion matrices comparing M^{4}4TMD and the dentists. Furthermore, the framework demonstrated robust calibration and discrimination capabilities, evidenced by the calibration curves (Fig. 4B) with low Brier scores (ranging from 0.059 to 0.173) and precision-recall curves (Fig. 4C) with high PR-AUC values (0.897 for DJD, 0.658–0.948 for ADD, and 0.961 for effusion). Detailed metrics are provided in Supplementary Tables 2–3.

3.2. Framework robustness and performance evaluation

3.2.1. Temporal generalization

On the temporal test set, M^{4}4TMD demonstrated strong performance with ROC-AUCs of 0.812, 0.921, and 0.967 for Task 1, Task 2, and Task 3, respectively. Across all three tasks, M^{4}4TMD achieved significantly higher accuracy than dentist-B ($P < 0.05$). Compared to dentist-A, the model showed significantly higher accuracy in Task 2 and Task 3 ($P <$

Table 1
Characteristics of the participants and TMJ joints in training sets and test sets.

Characteristic (Participants)	BJ Dataset (N=578)		BJ Dataset (N=578)		JF Dataset (N=187)
	Internal training (N=462)	Internal test (N=116)	Temporal training (N=465)	Temporal test (N=113)	External test (N=187)
Age (years), $\bar{x}$±SD	29.7±12.1	28.3±10.1	29.4±11.8	29.6±11.6	28.4±11.1
Gender, N (%)					
Female	359 (77.7 %)	93 (80.2 %)	366 (78.7 %)	86 (76.1 %)	160 (85.6 %)
Male	103 (22.3 %)	23 (19.8 %)	99 (21.3 %)	27 (23.9 %)	27 (14.4 %)
Course (years), $\bar{x}$±SD	1.8±2.7	1.6±3.0	1.8±2.8	1.7±2.6	2.8±3.7
Characteristic (Joints)	BJ Dataset (n=1059)		BJ Dataset (n=1059)		JF Dataset (n=351)
	Internal training (n=848)	Internal test (n=211)	Temporal training (n=848)	Temporal test (n=211)	External test (n=351)
Joint side, n (%)					
Right	432 (50.9 %)	106 (50.2 %)	429 (50.6 %)	109 (51.7 %)	169 (48.1 %)
Left	416 (49.1 %)	105 (49.8 %)	419 (49.4 %)	102 (48.3 %)	182 (51.9 %)
DJD diagnosis, n (%)					
Non-DJD	298 (35.1 %)	76 (36.0 %)	284 (33.5 %)	90 (42.6 %)	119 (33.9 %)
DJD	550 (64.9 %)	135 (64.0 %)	564 (66.5 %)	121 (57.4 %)	232 (66.1 %)
ADD diagnosis, n (%)					
Non-ADD	274 (32.3 %)	74 (35.1 %)	268 (31.6 %)	80 (37.9 %)	134 (38.2 %)
ADDWR	228 (26.9 %)	55 (26.1 %)	230 (27.1 %)	53 (25.1 %)	68 (19.4 %)
ADDWoR	346 (40.8 %)	82 (38.9 %)	350 (41.3 %)	78 (37.0 %)	149 (42.5 %)
Effusion diagnosis, n (%)					
Non-effusion	461 (54.4 %)	117 (55.5 %)	467 (55.1 %)	111 (52.6 %)	181 (51.6 %)
Effusion	387 (45.6 %)	94 (44.5 %)	381 (44.9 %)	100 (47.4 %)	170 (48.4 %)
Bruxism, n (%)					
Non-bruxism	658 (77.6 %)	166 (78.7 %)	696 (82.1 %)	128 (60.7 %)	251 (71.5 %)
Bruxism	190 (22.4 %)	45 (21.3 %)	152 (17.9 %)	83 (39.3 %)	100 (28.5 %)
Opening_limitation, n (%)					
Unrestricted	589 (69.5 %)	148 (70.1 %)	572 (67.5 %)	165 (78.2 %)	237 (67.5 %)
Limited	259 (30.5 %)	63 (29.9 %)	276 (32.5 %)	46 (21.8 %)	114 (32.5 %)
Opening_pattern, n (%)					
Vertically	361 (42.6 %)	115 (54.5 %)	371 (43.8 %)	105 (49.8 %)	182 (51.9 %)
Ipsilaterally	151 (17.8 %)	32 (15.2 %)	148 (17.5 %)	35 (16.6 %)	59 (16.8 %)
Contralaterally	142 (16.7 %)	32 (15.2 %)	143 (16.9 %)	31 (14.7 %)	60 (17.1 %)
S-shaped	194 (22.9 %)	32 (15.2 %)	186 (21.9 %)	40 (19.0 %)	50 (14.2 %)
Pain, n (%)					
No pain	527 (62.1 %)	139 (65.9 %)	524 (61.8 %)	142 (67.3 %)	217 (61.8 %)
Pain	321 (37.9 %)	72 (34.1 %)	324 (38.2 %)	69 (32.7 %)	134 (38.2 %)

$\bar{x}$: mean, SD: standard deviation.

Table 2
Comparison of M⁴4TMD and previous studies on TMD assessment.

Method		Nozawa et al. [25] (ResNet18)	Yoon et al. [17] (ResNet50)	Lee et al. [24] (VGG16)	M⁴4TMD (Ours)
ROC-AUC (95 % CI)	Task 1	0.793 (0.729–0.850)	/	/	**0.831** (0.774–0.881)
	Task 2	/	0.896 (0.862–0.927)	/	**0.913** (0.880–0.945)
	Task 3	/	/	0.937 (0.831–0.918)	**0.961** (0.929–0.982)
F1 Score	Task 1	0.735	/	/	**0.754**
	Task 2	/	0.742	/	**0.785**
	Task 3	/	/	0.881	**0.905**
Precision	Task 1	0.751	/	/	**0.773**
	Task 2	/	0.763	/	**0.798**
	Task 3	/	/	0.883	**0.907**
Recall	Task 1	0.730	/	/	**0.749**
	Task 2	/	0.735	/	**0.782**
	Task 3	/	/	0.882	**0.905**
Brier score	Task 1	0.183	/	/	**0.173**
	Task 2	/	0.373	/	**0.351**
	Task 3	/	/	0.125	**0.085**
PR-AUC	Task 1	0.884	/	/	**0.897**
	Task 2	/	0.793	/	**0.829**
	Task 3	/	/	0.936	**0.961**

Since the original codes and datasets were not publicly available, the comparison methods (Nozawa et al. [25], Yoon et al. [17], and Lee et al. [24]) were re-implemented and trained on our BJ Dataset to ensure a fair comparison.
Brier scores range from 0 to 1 for the binary tasks (Task 1, Task 3) and from 0 to 2 for the multiclass task (Task 2).

Table 3
Head-to-head comparison of M⁴TMD and dentists on TMD assessment: evaluation on the internal test set.

Our Method vs. Dentists		Dentist-A	Dentist-B	Dentist-C	Dentist-D	M⁴4TMD
Accuracy (95 % CI)	Task 1	55.0 % (48.2 %-61.5 %)	51.7 % (45.0 %-58.3 %)	**74.9 %** (68.6 %-80.3 %)	72.5 % (66.1 %-78.1 %)	**74.9 %** (68.6 %-80.3 %)
	Task 2	44.6 % (38.0 %-51.3 %)	56.4 % (49.7 %-62.9 %)	71.1 % (64.6 %-76.8 %)	75.8 % (69.6 %-81.1 %)	**78.2 %** (72.2 %-83.2 %)
	Task 3	64.0 % (57.3 %-70.2 %)	79.2 % (73.2 %-84.1 %)	88.6 % (83.6 %-92.2 %)	79.6 % (73.7 %-84.5 %)	**90.5 %** (85.8 %-93.8 %)
Sensitivity (95 % CI)	Task 1	43.7 % (35.6 %-52.1 %)	34.1 % (26.6 %-42.4 %)	68.9 % (60.7 %-76.1 %)	**74.8 %** (66.9 %-81.4 %)	72.6 % (64.5 %-79.4 %)
	Task 2	44.6 % (38.0 %-51.3 %)	56.4 % (49.7 %-62.9 %)	71.1 % (64.6 %-76.8 %)	75.8 % (69.6 %-81.1 %)	**78.2 %** (67.7 %-85.9 %)
	Task 3	58.5 % (48.4 %-67.9 %)	71.3 % (61.4 %-79.5 %)	**94.7 %** (88.2 %-97.7 %)	58.5 % (48.4 %-67.9 %)	92.3 % (86.3 %-95.7 %)
Specificity (95 % CI)	Task 1	75.0 % (64.2 %-83.4 %)	82.9 % (72.9 %-89.7 %)	**85.5 %** (75.9 %-91.7 %)	68.4 % (57.3 %-77.8 %)	79.0 % (68.5 %-86.6 %)
	Task 2	72.0 % (65.1 %-77.2 %)	80.7 % (74.7 %-85.3 %)	87.2 % (82.0 %-91.1 %)	89.6 % (84.2 %-92.6 %)	**90.3 %** (81.3 %-95.0 %)
	Task 3	68.4 % (59.5 %-76.1 %)	85.5 % (78.0 %-90.7 %)	83.8 % (76.0 %-89.4 %)	**96.6 %** (91.5 %-98.7 %)	88.9 % (81.9 %-93.4 %)
Raw *P* value (Adjusted *P* value)	Task 1	<0.001 (<0.001) *	<0.001 (<0.001) *	1.000 (1.000)	0.653 (1.000)	/
	Task 2	0.177 (0.193)	0.005 (0.020) *	0.064 (0.192)	0.076 (0.193)	/
	Task 3	<0.001 (<0.001) *	0.002 (0.007) *	0.557 (0.557)	0.004 (0.008)*	/

The *P* values were calculated for M⁴4TMD in comparison to dentists using the McNemar test on Task 1 and Task 3, using the Stuart-Maxwell test on Task 2. The Holm-Bonferroni correction was applied for multiple comparisons (N=4);
* indicates adjusted $P < 0.05$.

0.01) but performed comparably in Task 1 (70.6 % vs. 61.6 %, $P > 0.05$). In comparison with senior dentists, M⁴4TMD showed more nuanced differences in performance: in Task 1, it was comparable to the top-performing senior dentist (senior dentist-D) for DJD (70.6 % vs. 73.0 %, $P > 0.05$). For Task 2, M⁴4TMD significantly outperformed dentist-C (79.2 % vs. 68.7 %, $P < 0.05$). Notably, while M⁴4TMD achieved the exact same accuracy as dentist-D in Task 2 (79.2 % vs. 79.2 %), the statistical analysis indicated a significant difference in their prediction distributions ($P = 0.024$). In Task 3, it significantly outperformed both senior dentists (90.5 % vs. 82.5 % vs. 81.5 %, $P < 0.01$) (Fig. 5A, C; detailed metrics are provided in Supplementary Tables 4–5).

3.2.2. External generalization

On the external test set, M⁴4TMD maintained strong generalization with ROC-AUCs of 0.782, 0.926, and 0.930 for DJD, ADD, and effusion, respectively. Regarding accuracy, the model performed comparably to the top-performing senior dentists for DJD and effusion: specifically, it was comparable to dentist-D for DJD (72.1 % vs. 73.2 %, $P > 0.05$) and dentist-C for effusion (86.3 % vs. 86.3 %, $P > 0.05$). For ADD (Task 2), while the model achieved significantly higher accuracy than dentist-C (74.6 % vs. 71.2 %, $P < 0.01$), it fell short of dentist-D (74.6 % vs. 80.1 %, $P < 0.01$). Notably, the framework outperformed dentist-B across all tasks ($P < 0.01$) and dentist-A in Task 1 and Task 3 ($P < 0.01$), with a borderline non-significant difference against dentist-A in Task 2 ($P = 0.053$) (Fig. 5B, D). Detailed metrics are provided in Supplementary Tables 6–7.

3.2.3. Framework adaptability

The M⁴4TMD framework demonstrated strong adaptability, with all implemented backbones (ResNet50, VGG16, DenseNet121, Inception-V3) generally outperforming the clinicians. Among the different architectures, the ResNet50-based model achieved the best overall performance, particularly for DJD detection, while the VGG16-based model showed the weakest performance (Table 4). Nevertheless, all framework variations surpassed both junior dentists in all tasks and outperformed all four dentists in the detection of effusion (Task 3). The performance was more nuanced for other tasks: for DJD (Task 1), only the ResNet50-based model exceeded all dentists, while for ADD (Task 2), even the weaker VGG16-based model performed better than three of the four clinicians (73.9 % vs. 44.6 % vs. 56.4 % vs. 71.1 % vs. 75.8 %) (Fig. 6; detailed metrics in Supplementary Tables 8–9).

3.2.4. Stratified analysis of M⁴4TMD performance

To further assess the robustness and clinical generalizability of the framework, we conducted a comprehensive stratified analysis on the internal test set based on seven key demographic and clinical characteristics. As presented in Table 5, M⁴4TMD demonstrated consistent and stable performance across all subgroups. For instance, in the gender-

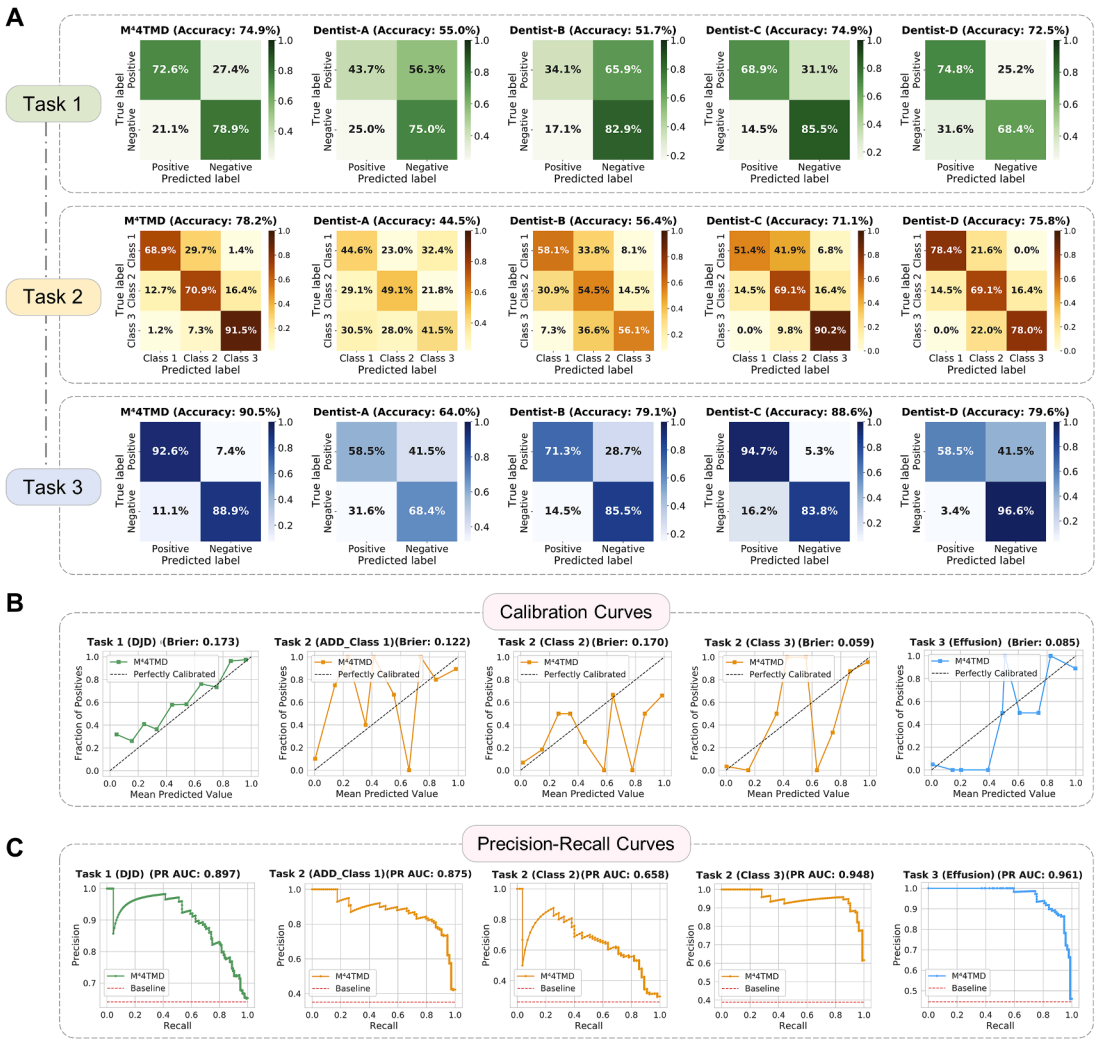


Fig. 4. Performance evaluation of M[4]4TMD on the internal test set. A: Confusion matrices of M[4]4TMD and four dentists across three tasks. B: Calibration curves showing the agreement between predicted probabilities and observed frequencies for each task. C: Precision-Recall curves illustrating the trade-off between precision and recall for M[4]4TMD across all tasks.

stratified analysis, the model achieved comparable ROC-AUCs for females and males across all tasks (e.g., 0.835 vs. 0.821 for DJD). Similarly, the model maintained high efficacy regardless of specific clinical presentations; notably, for joint effusion (Task 3), the ROC-AUC consistently exceeded 0.940 across all stratified categories, including patients with different opening patterns and pain levels. These results confirm that the assessment capability of M[4]4TMD remains robust irrespective of variations in patient demographics or clinical symptoms.

3.3. Visual interpretability and error analysis

3.3.1. Assessment of anatomical consistency and spatial alignment

The quantitative evaluation demonstrated a clear spatial consistency between the attention regions of M[4]4TMD and clinically relevant anatomical structures (Table 6). Specifically, Task 1 exhibited the highest degree of spatial overlap, achieving a mean IoU of 0.126 ± 0.039 and a Pointing Game accuracy of 30.4 %. Task 2 displayed a definitive anatomical alignment, characterized by a mean IoU of 0.089 ± 0.028 and a Pointing Game accuracy of 22.6 %. Notably, Task 3 attained the highest Pointing Game accuracy across all tasks (33.8 %), alongside a mean IoU of 0.101 ± 0.043. These quantitative findings are visually corroborated by the qualitative visualizations presented in Fig. 7. As illustrated, the task-specific Grad-CAM heatmaps and their corresponding binarized attention masks, representing the top 4 % salient regions (green overlays), exhibit distinct spatial alignment with the annotated ground truths (pink for condyle, orange for articular disc, and blue for joint space / effusion). The anatomical rationality of the model's decision-making process was further corroborated by the results of the occlusion sensitivity analysis, the details of which are provided in the Supplementary Materials.

3.3.2. Quantitative analysis of classification error patterns

To further elucidate the patterns of classification errors, a detailed quantitative error analysis was conducted based on the confusion matrices for each task on the internal test set (Fig. 8A-D). A stratified analysis of false negative (FN) cases provided insight into the "hard cases" for DJD detection. As shown in Fig. 8B1-B2, the detection capability of M[4]4TMD was poorest when pathological features appeared in isolation; notably, 78.4 % of missed diagnoses occurred in joints presenting with only a single radiographic sign. Crucially, regarding the severity of these signs, the "missed" cases were predominantly characterized by subtle pathology. Among FN cases involving erosion, 68.2 % were classified as "subtle erosion" (Fig. 8B3); similarly, 57.1 % of missed osteophytes were graded as "slight (< 1 mm)" (Fig. 8B4). Fig. 8E (a-c) presents representative examples of these FN cases, where faint cortical irregularities or minute cysts on MRI were likely insufficient to trigger the detection threshold of the model, despite being identified on the reference CBCT images. For Task 2 (ADD), utilizing the RDC/TMD

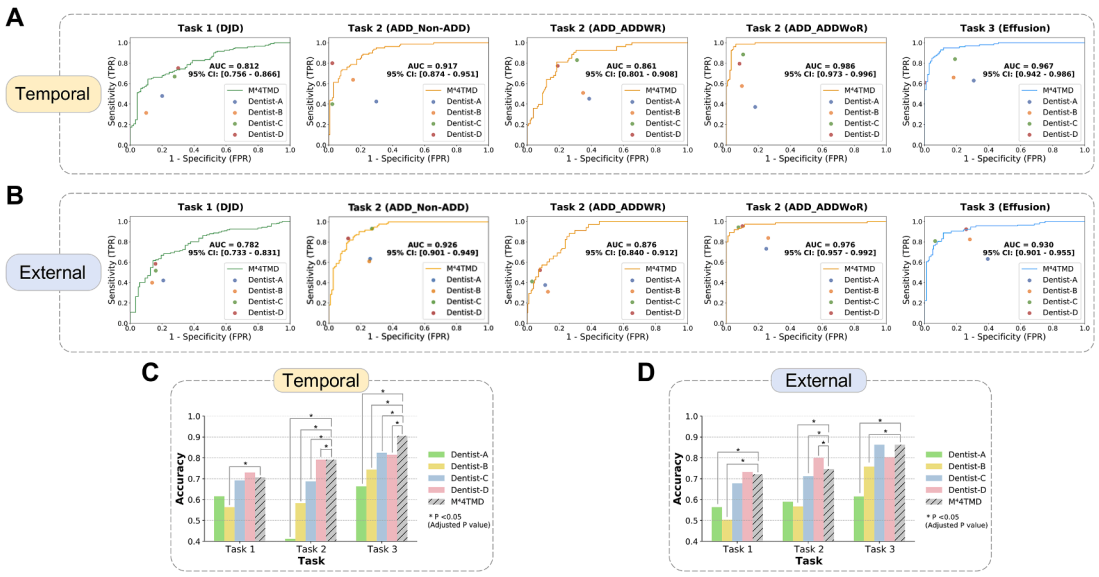


Fig. 5. Generalization performance of M^{4}4TMD against dentists on temporal and external test sets. A-B: Receiver operating characteristic curves of M^{4}4TMD across three tasks on the temporal (A) and external (B) test sets. The performance of individual dentists is plotted for comparison. C-D: Classification accuracy comparisons between M^{4}4TMD and four dentists on the temporal (C) and external (D) test sets. *P* values were calculated to compare M^{4}4TMD with dentists using the McNemar test for binary classification tasks (Task 1 and Task 3) and the Stuart-Maxwell test for the multi-class task (Task 2). The Holm-Bonferroni correction was applied for multiple comparisons (N=4); * indicates an adjusted $P < 0.05$.

Table 4
Adaptability Analysis: ROC-AUC Performance of the M^{4}4TMD Framework across CNN Backbones.

ROC-AUC (95 %CI)	M^{4}4TMD _VGG-16	M^{4}4TMD _DenseNet-121	M^{4}4TMD _Inception-V3	M^{4}4TMD_ResNet50
Task 1	0.741 (0.665–0.811)	0.788 (0.724–0.846)	0.795 (0.734–0.852)	**0.831** (0.774–0.881)
Task 2	0.879 (0.842–0.914)	**0.918** (0.886–0.949)	0.908 (0.876–0.940)	0.913 (0.880–0.945)
Task 3	0.957 (0.930–0.978)	**0.970** (0.943–0.988)	**0.970** (0.949–0.987)	0.961 (0.929–0.982)
Overall	0.844	0.880	0.885	**0.895**

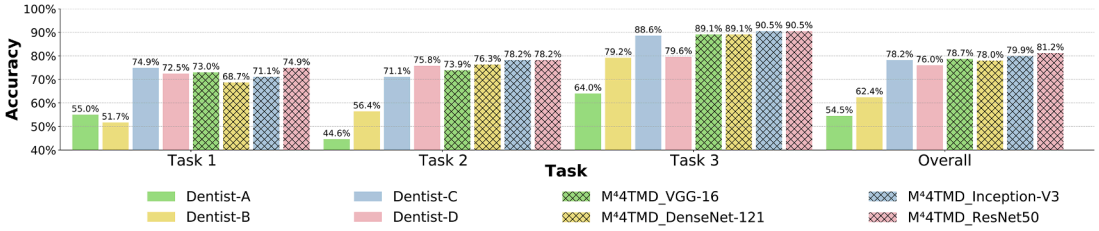


Fig. 6. M^{4}4TMD framework adaptability and comparison with dentists' performance.

classification for articular disc morphology, the error analysis highlighted the confounding influence of disc shape (Fig. 8C). Compared to correctly diagnosed Non-ADD and ADDWR cases, the proportion of discs exhibiting morphological changes was noticeably higher in misdiagnosed cases. Finally, for Task 3 (effusion), given the model's high diagnostic accuracy, the distribution of slight versus frank effusion in FN cases was comparable to that observed in correctly diagnosed cases (Fig. 8D).

3.4. Component evaluation and clinical feature interpretability

3.4.1. Stepwise ablation study of framework components

A stepwise ablation study was conducted to systematically evaluate the contribution of each component of the M^{4}4TMD framework (multi-slice, multi-sequence, multi-task, and multimodal). On the internal test set, performance improved incrementally: incorporating multiple slices (M^{1}4TMD) and multi-sequence MRI (M^{2}4TMD) progressively enhanced accuracy. However, the benefits of multi-task learning (M^{3}4TMD) and multimodal fusion (M^{4}4TMD) were not apparent on this test set, with all three models yielding comparable results (Fig. 9A).

Nevertheless, the advantages of multimodal fusion became evident during external evaluation. On this more challenging test set, M^{4}4TMD demonstrated superior generalization than single-modal frameworks (M^{2}4TMD and M^{3}4TMD) across all tasks compared (Table 7). Specifically, feature fusion strategy proved optimal across all three tasks, achieving ROC-AUCs of 0.782 for DJD, 0.926 for ADD, and 0.930 for effusion. These results confirmed the distinct value of each "M" component for achieving robust and generalizable performance (Supplementary Tables 10–12).

3.4.2. Quantitative assessment of clinical feature contribution

On the internal test set (Fig. 9B, yellow bars), PI scores for the majority of clinical features were generally low. Specifically, for Task 1 (DJD), "age" contributed the most, with an ROC-AUC drop of 0.008. For Task 2 (ADD), "opening limitation" showed the highest importance (0.003), followed by "pain" (0.002). For Task 3 (Effusion), "pain" was

Table 5
Performance of M[4]4TMD stratified by participant demographics and clinical features.

Clinical feature	subgroup	n (%)	Task 1 (DJD)		Task 2 (ADD)		Task 3 (Effusion)	
			ROC-AUC	Accuracy	ROC-AUC	Accuracy	ROC-AUC	Accuracy
All		211 (100 %)	0.831	74.9 %	0.913	78.2 %	0.961	90.5 %
Gender	Female	169 (80.1 %)	0.835	75.2 %	0.909	77.5 %	0.961	90.5 %
	Male	42 (19.9 %)	0.821	73.8 %	0.939	81.0 %	0.967	90.5 %
Age	≤ 35 years	149 (70.6 %)	0.814	73.8 %	0.897	77.2 %	0.963	91.3 %
	> 35 years	62 (29.4 %)	0.876	77.4 %	0.953	80.7 %	0.958	88.7 %
Course	≤ 1 year	111 (52.6 %)	0.867	76.6 %	0.950	82.0 %	0.942	89.2 %
	> 1 year	100 (47.4 %)	0.785	73.0 %	0.876	74.0 %	0.980	92.0 %
Bruxism	Non-bruxism	166 (78.7 %)	0.845	75.3 %	0.916	77.7 %	0.958	91.0 %
	Bruxism	45 (21.3 %)	0.782	73.4 %	0.906	80.0 %	0.973	88.9 %
Opening limitation	Unrestricted	148 (70.1 %)	0.823	75.0 %	0.900	75.7 %	0.970	91.9 %
	Limited	63 (29.9 %)	0.845	74.6 %	0.946	84.1 %	0.940	87.3 %
Opening pattern	Vertically	115 (54.5 %)	0.809	70.4 %	0.920	80.0 %	0.963	89.6 %
	Ipsilaterally	32 (15.2 %)	0.827	71.9 %	0.918	79.6 %	0.941	90.6 %
	Contralaterally	32 (15.2 %)	0.866	84.4 %	0.871	74.0 %	0.981	93.8 %
	S-shaped	32 (15.2 %)	0.813	84.4 %	0.869	74.5 %	0.941	90.6 %
Pain	No pain	139 (65.9 %)	0.803	71.9 %	0.906	77.7 %	0.950	90.7 %
	Pain	72 (34.1 %)	0.863	80.6 %	0.931	79.2 %	0.974	90.3 %

Table 6
Quantitative evaluation of task-specific ROI alignment on the internal test subset.

	IoU ($\bar{x} \pm$ SD)	Pointing Game (Accuracy)
Task 1	0.126 ± 0.039	30.4 %
Task 2	0.089 ± 0.028	22.6 %
Task 3	0.101 ± 0.043	33.8 %

The analysis was performed on a randomly selected subset of 53 TMJs, comprising a total of 477 MR images.
Target Anatomical ROIs: Task 1: mandibular condylar head; Task 2: articular disc; Task 3: joint cavity / effusion.
$\bar{x}$: mean, SD: standard deviation.

the only prominent feature with an importance score of 0.005, while "opening limitation" contributed 0.003.

Conversely, on the external test set (Fig. 9B, blue bars), the importance of key clinical features either increased or remained robust, although the distribution varied across tasks: Task 1: "Age" remained the primary predictor, with its importance score rising to 0.011. "Course" also demonstrated a positive contribution of 0.005. Task 2: Distinct from the internal evaluation, "pain" emerged as the most critical feature (0.007), followed by "opening limitation" (0.005) and "opening pattern" (0.002). Task 3: "Pain" further solidified its role as a key assessment marker, with its importance score increasing to 0.011, significantly surpassing other variables.

Collectively, these results highlight a pivotal shift in feature reliance. While the M[4]4TMD framework predominantly leveraged high-quality MRI features within the internal test set, an elevation in the importance of specific clinical variables was observed in the external evaluation. This suggests that when MRI features become relatively less reliable due to domain shifts (e.g., variations in scanner protocols), the model exhibited increased reliance on domain-invariant clinical signs to maintain performance robustness.

4. Discussion

In this study, the M[4]4TMD framework was developed and evaluated, which, to our knowledge, is the first multimodal, multi-task DL framework for comprehensive assessment of TMD-related abnormalities. Our method was built upon a comprehensive dataset of 765 participants

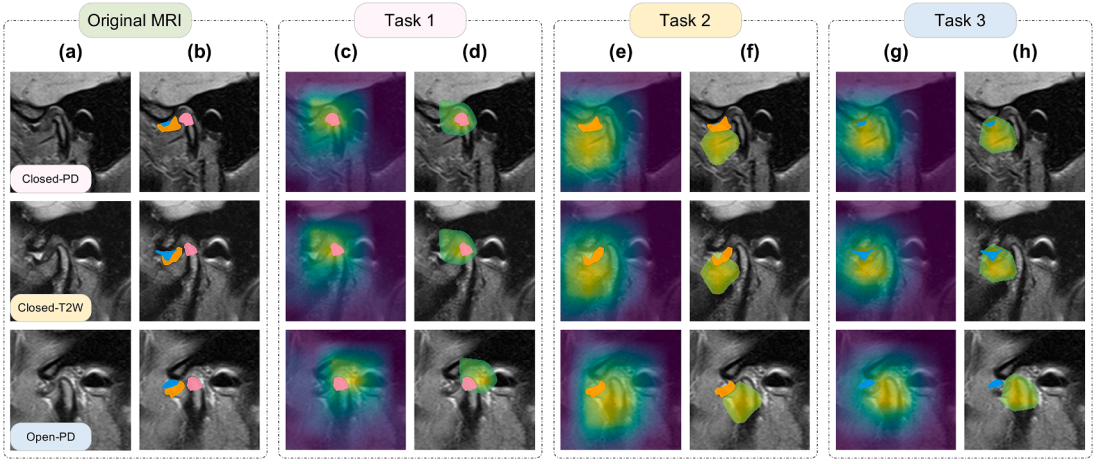
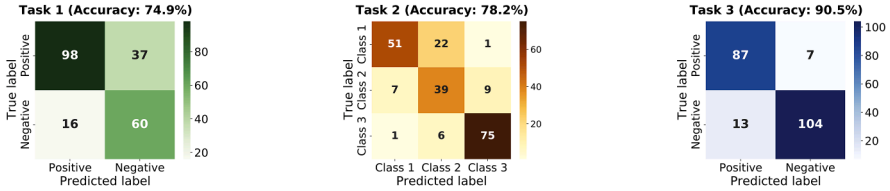


Fig. 7. Qualitative visualization of task-specific model interpretability and ROI alignment. Columns (a)-(b) display the original MRI sequences and the ground truth anatomical annotations by the experienced dentist (Pink: Condyle; Orange: Articular Disc; Blue: Joint space / effusion). Columns (c)-(h) illustrate the task-specific Grad-CAM heatmaps and their corresponding binarized attention masks overlaying the MRI. Rows correspond to the three input MRI sequences: Closed-PD, Closed-T2W, and Open-PD. (c-d) Task 1 (DJD): M[4]4TMD focuses on the condylar head region. (e-f) Task 2 (ADD): Attention is concentrated on the articular disc. (g-h) Task 3 (effusion): The heatmap highlights the joint space, showing high consistency with the effusion area. The green overlays in (d), (f), and (h) represent the top 4 % salient regions (96th percentile threshold) used for quantitative IoU calculation, demonstrating the alignment between model focus and anatomical structures.

A

Confusion Matrices



B

Task 1

B1: FN Analysis by Radiographic Sign

Radiographic Signs of DJD	True Positive Cases (n=98 [%])	False Negative Cases (n=37 [%])
Erosion	59 (60.2%)	22 (59.5%)
Osteophytes	54 (55.1%)	7 (18.9%)
Sclerosis	44 (44.9%)	12 (32.4%)
Subchondral Cysts	32 (32.7%)	5 (13.5%)

B2: FN Analysis by Sign Count

Number of Radiographic Signs of DJD	True Positive Cases (n=98 [%])	False Negative Cases (n=37 [%])
1	36 (36.7%)	29 (78.4%)
2	38 (38.8%)	7 (18.9)
3	19 (19.4%)	1 (2.7%)
4	5 (5.1%)	0 (0%)

B3: FN Analysis: Erosion Severity

Erosion Grade	True Positive Cases (n=59 [%])	False Negative Cases (n=22 [%])
Subtle Erosion	16 (27.1%)	15 (68.2%)
Top of Condyle	35 (59.3%)	6 (27.3%)
Half of the Condyle	8 (13.6%)	1 (4.5%)
Complete Condyles	0 (0%)	0 (0%)

B4: FN Analysis: Osteophyte Size

Osteophytes Grade	True Positive Cases (n=54 [%])	False Negative Cases (n=7 [%])
Slight (< 1mm)	16 (29.6%)	4 (57.1%)
Moderate (1-2mm)	23 (42.6%)	3 (42.9%)
Extensive (> 2mm)	15 (27.8%)	0 (0%)

C

Task 2

C: Error Analysis by Disc Morphology

Disc Morphology	Class 1 (Non-ADD)		Class 2 (ADDWR)		Class 3 (ADDWoR)	
	Correct Predictions (n=51 [%])	Incorrect Predictions (n=23 [%])	Correct Predictions (n=39 [%])	Incorrect Predictions (n=16 [%])	Correct Predictions (n=75 [%])	Incorrect Predictions (n=7 [%])
Biconcave	41 (80.4%)	16 (69.6%)	21 (53.8%)	5 (31.3%)	11 (14.7%)	1 (14.3%)
Deformed	10 (19.6%)	7 (30.4%)	18 (46.2%)	11 (68.8%)	64 (85.3%)	6 (85.7%)

D

Task 3

D: FN Analysis by Effusion Grade

Effusion Grade	True Positive Cases (n=87 [%])	False Negative Cases (n=7 [%])
Slight Effusion	45 (51.7%)	4 (57.1%)
Frank Effusion	42 (48.3%)	3 (42.9%)

E

Cases

Fig. 8. Detailed quantitative error analysis and visualization of representative hard cases on the internal test set. **A:** Confusion matrices illustrating the overall classification performance of M[4]4TMD across Task 1 (DJD), Task 2 (ADD), and Task 3 (effusion). **B:** Stratified analysis of False Negative (FN) versus True Positive (TP) cases for DJD (Task 1). The tables quantify the impact of specific pathological features on detection sensitivity: (B1) by radiographic sign type; (B2) by the count of concurrent signs; (B3) by erosion severit; and (B4) by osteophyte size. **C:** Analysis of the relationship between disc morphology and prediction errors in ADD (Task 2). **D:** Sensitivity analysis for Effusion (Task 3) stratified by effusion severity. **E:** Representative "hard cases" visualizing specific error patterns: (a-c) FN DJD cases involving subtle osseous changes: (a) slight osteophyte (closed-PD MRI), (b) subtle erosion (closed-PD MRI), and (c) small subchondral cyst (closed-T2W MRI). (d-e) Misclassification in ADD involving atypical disc morphology: (d) a Non-ADD case with an elongated / thinned disc misclassified as ADDWR (closed-PD MRI), and (e) an ADDWR case with a contracted disc misclassified as ADDWoR (closed-T2W MRI). (f) A FN case of slight joint effusion missed by the model (closed-T2W). Red arrows indicate the subtle pathological features or the specific anatomical structure.

(1410 TMJs), integrating multi-sequence, multi-slice MRI with clinical data annotated for DJD, ADD, and effusion. The results compellingly demonstrate that the performance of M[4]4TMD surpasses that of junior dentists and is comparable to that of senior dentists. This finding remained consistent across internal, temporal, and external test sets, and the framework's adaptability was further confirmed by its robust performance across various CNN backbone architectures. Furthermore, Grad-CAM analysis confirmed the framework's focus on clinically relevant anatomy, aligning its diagnostic focus with that of clinicians.

The robust performance of M[4]4TMD highlights its considerable potential for assisting clinical practice and broadening access to expert-level assessment. Primarily, this framework could serve as a decision-support tool for general dentists and other specialists, such as prosthodontists and orthodontists, who may have limited experience in TMD evaluation. This is particularly crucial when a thorough TMJ assessment is a prerequisite for complex specialized treatments—for example, before restoration in patients with long-term edentulism or prior to orthodontics for complex malocclusion—to prevent misdiagnoses that

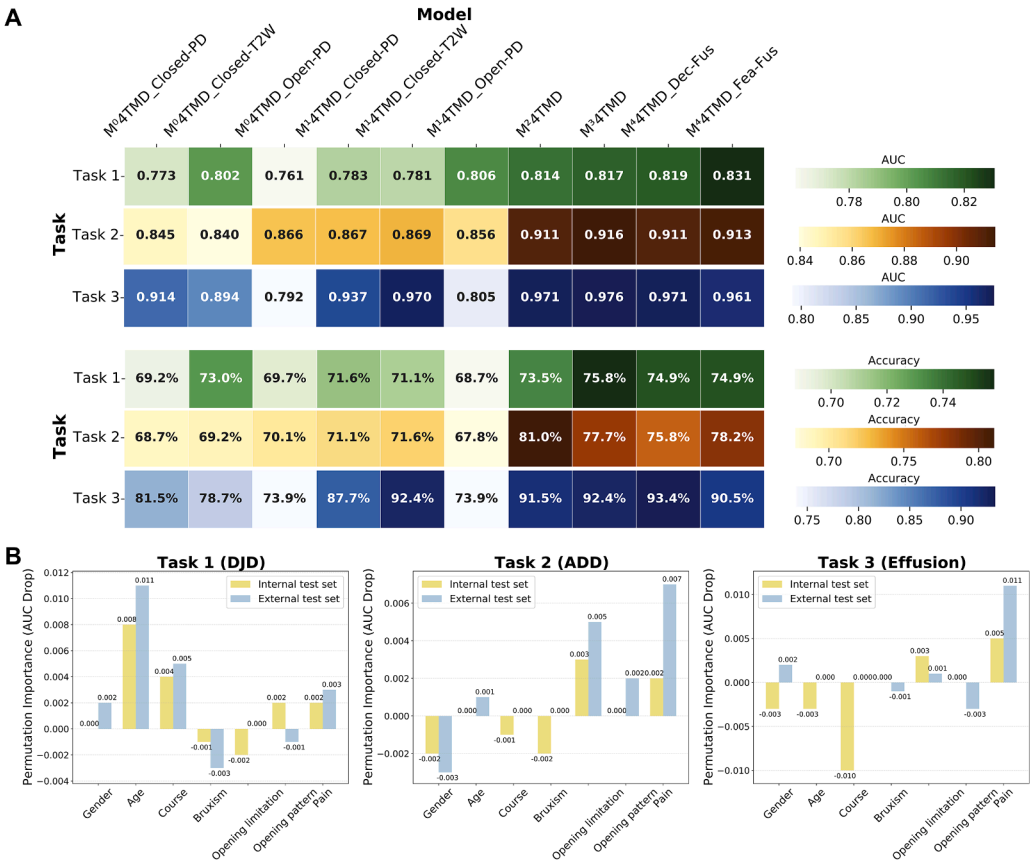


Fig. 9. Stepwise ablation analysis of framework components and permutation importance of clinical features. **A:** Heatmaps illustrating the ROC-AUC and accuracy of different framework variants (ablation study) across three tasks on the internal test set. Darker colors indicate higher performance. **B:** Permutation feature importance of seven clinical variables across internal (yellow) and external (blue) test sets.

Table 7
Performance comparison of M²4TMD, M³4TMD, and M⁴4TMD on the external test set.

ROC-AUC (95 % CI)	M²4TMD	M³4TMD	M⁴4TMD_Dec-Fus	M⁴4TMD_Fea-Fus
Task1	0.757 (0.663–0.849)	0.766 (0.676–0.858)	0.768 (0.670–0.866)	**0.782** (0.733–0.831)
Task2	0.908 (0.873–0.939)	0.915 (0.885–0.943)	0.914 (0.898–0.931)	**0.926** (0.892–0.958)
Task3	0.906 (0.860–0.953)	0.916 (0.864–0.972)	0.926 (0.895–0.960)	**0.930** (0.901–0.955)

Dec_Fus = decision fusion, Fea_Fus = feature fusion

could complicate care [3,4]. Furthermore, M⁴4TMD can facilitate TMD evaluations in regions with limited access to TMJ specialists. Beyond diagnostic accuracy, the clinical utility of M⁴4TMD is reinforced by its computational efficiency. The model achieves an inference speed of approximately 66 ms per case (minimum hardware requirements are listed in Supplementary Table 18), making it suitable for real-time deployment.

A notable aspect of this study is the specific focus on detecting DJD using MR images, a departure from the conventional gold standard of CT or CBCT. This approach is inherently challenging; MRI is known to have limited accuracy in detecting osseous changes of the TMJ, primarily due to its reliance on proton density imaging [16]. Hard tissues, with their low proton density, are less effectively depicted than soft tissues [36,37]. In this study, the difficulty of this task was underscored by the low accuracy of junior dentists when assessing DJD using MRI data. However, this challenge was pursued because MRI is the only technique capable of simultaneously evaluating both soft-tissue and bone abnormalities in the TMJ without ionizing radiation [15,38]. To overcome the limitations of MRI, a robust training strategy was employed: the model was trained on MRI data, but the ground truth labels for DJD were derived from corresponding CBCT scans. To our knowledge, this is the first study to develop and validate a method for MRI-based DJD detection on a large-scale dataset, offering clinical value by potentially reducing patient radiation exposure and examination costs.

While this training strategy established a strong baseline, the intrinsic physical limitations of MRI regarding hard tissue detection persisted in specific scenarios [39]. Our quantitative error analysis revealed that the model exhibited an elevated false-negative rate (FNR) for DJD, where “missed” diagnoses were predominantly characterized by isolated or minute pathological features. For instance, on the internal test set, 68.2 % of missed erosion cases presented as “subtle erosion”, and 57.1 % of missed osteophytes were graded as “slight (< 1 mm)”. To bridge the gap between MRI’s safety and the precision required for these subtle cases [40–41], we implemented a clinician-AI hybrid diagnostic simulation for Task 1. The results demonstrated that integrating AI predictions with the senior dentist yields performance superior to either alone (Supplementary Table 16). By dynamically adjusting the decision threshold (τ), the system could be tuned for different clinical priorities: setting $\tau = 0.65$ optimized the workflow for precision, achieving an overall accuracy of 80.6 %, which surpassed both the standalone model and the senior dentist. Furthermore, to translate this hybrid capability into a practical strategy for reducing radiation exposure, we evaluated the feasibility of this framework as a “Radiation Gatekeeper”. We proposed a two-step triage pathway (Strategy II) incorporating a dual-check safety net (requiring concordant negative assessments from both the AI

and the clinician to defer CBCT imaging). At a threshold of $\tau = 0.45$, this strategy achieved a sensitivity of 94.1 %, and a robust Negative Predictive Value (NPV) of 85.3 %, allowing 22.7 % of patients to safely defer CBCT imaging (Supplementary Table 17). Given that DJD is typically a chronic, non-life-threatening condition, this NPV (over 85 %) supports a safe protocol of conservative clinical monitoring (a "watch-and-wait" approach) for the "double-negative" group. In this pathway, CBCT imaging is deferred and performed only if clinical symptoms persist or worsen, effectively fulfilling the study's goal of minimizing unnecessary radiation and examination costs. The details of "Clinician-AI Hybrid Diagnostic Simulation for Task 1" and "Two-Step Pathway for CBCT Triage" are provided in the Supplementary Materials.

Our stepwise ablation studies demonstrated the value of integrating comprehensive imaging data, revealing distinct performance benefits from the addition of both multi-slice and multi-sequence MRI inputs. The transition from single-slice ($M^0$4TMD) to multi-slice inputs ($M^1$4TMD) yielded a significant performance enhancement, likely due to reduced image artifacts and the provision of richer anatomical context [42]. Building on this, the incorporation of multi-sequence MRI ($M^2$4TMD) provided a further boost in accuracy. This gain stems from the complementary nature of different MRI sequences; since each sequence offers unique tissue contrast to highlight different pathological features, their fusion compensates for the inherent limitations of any single sequence, enabling a more definitive identification of disease [43,44].

This study advances the field by moving beyond the conventional single-task paradigm of previous TMD research, which has largely focused on articular disc displacement, i.e., ADD, and rarely distinguishing between its subtypes. In contrast, this framework was designed to concurrently address three distinct yet often co-occurring abnormalities: DJD, ADD (with and without reduction), and joint effusion. This comprehensive scope was enabled by a multi-task learning strategy, which is particularly well-suited for TMD since it trains the model to leverage shared pathological features across related assessment tasks. As hypothesized, this approach improved overall performance and enhanced generalization. The benefits of this strategy were empirically evaluated on the external test set, where the multi-task model ($M^3$4TMD) demonstrated a clear performance advantage over its single-task counterpart ($M^2$4TMD).

The multimodal approach employed in this study aligns with the growing consensus that integrating diverse data sources is essential for complex diagnostic tasks [23]. This principle is particularly resonant in the assessment of TMD, which clinically requires a synthesis of imaging findings with a wide array of non-imaging data [1]. The architecture of $M^4$4TMD was expressly designed to mirror this clinical reality. To explicitly quantify the contribution of these non-imaging factors, we employed PI analysis, revealing task-specific dependencies that align with clinical pathology. For DJD (Task 1), "age" emerged as the dominant predictor, corroborating established literature regarding the progressive and irreversible nature of osteoarthritic changes over time [45]. Conversely, for ADD and effusion (Tasks 2 and 3), functional indicators such as "pain" and "opening limitation" demonstrated higher relative importance, reflecting their association with inflammatory states and disc displacement mechanics [46,47]. Crucially, our analysis uncovered a dynamic shift in feature reliance that explains the generalization capability of the framework. While the model prioritized high-frequency MRI details in the internal set, the importance of clinical variables (particularly "age" and "pain") markedly increased in the external evaluation. This suggests that clinical features function as domain-invariant anchors; when MRI feature reliability fluctuates due to domain shifts (e.g., scanner variations), the model adaptively leverages these robust clinical signals to stabilize diagnostic performance [23,48,49]. This compensatory mechanism validates that fusing MRI data with key clinical variables creates a more resilient representation than imaging alone.

Although this study presented promising results, several limitations require consideration and also outline avenues for future research. First, data were sourced from a single institution. Although we enhanced diversity by including two campuses with different MRI protocols, validation across multiple institutions is necessary to confirm the broader generalizability of the model. Second, the assessment scope was constrained in two ways regarding disease categorization. Consistent with previous DL studies based on TMJ CT / CBCT imaging [50,51], the DJD analysis was framed as a binary classification (DJD vs. non-DJD) by excluding cases classified as "indeterminate for DJD" per DC/TMD criteria [1]. This exclusion was a pragmatic decision driven by data characteristics and the need for robust model training. Given that indeterminate cases constituted a small minority in our initial cohort (103 of 1606 TMJs) and are inherently ambiguous even for experts, including them posed a risk of introducing label noise that could hinder model convergence. By focusing on a clear-cut binary ground truth, we ensured the model learned stable and distinct features of definitive degeneration. Additionally, the analysis of disc displacement was confined to its anterior form (ADD). While this simplified approach established a solid baseline, future studies should aim to leverage larger, multi-center datasets to develop a comprehensive 3-class DJD model (Non-DJD, Indeterminate, DJD) and integrate data from coronal MRI planes for a more comprehensive, three-dimensional assessment. Finally, the ground truth was established by human annotations. This introduced an element of inherent subjectivity, a fundamental characteristic of nearly all AI models trained on annotated data [52–54]. Despite efforts to enhance label accuracy and transparency through consistency training and expert adjudication, some subjectivity remained unavoidable. To further mitigate this limitation and create a high-fidelity reference standard for DJD, annotations were based on CBCT imaging, a modality widely recognized to have superior accuracy for detecting osseous changes over MRI [1,11]. Both the $M^4$4TMD framework and the human clinicians were then evaluated fairly on their ability to make assessment using only the MRI and clinical data. Given this methodological rigor, it is contended that the ground truth is more robust than that of many current AI medical imaging studies, and performance comparisons should be interpreted with this context in mind.

5. Conclusions

$M^4$4TMD framework was developed, establishing its efficacy as a multimodal, multi-task DL framework that integrates multi-sequence, multi-slice MRI findings with clinical data for the comprehensive assessment of TMD-related abnormalities. Results demonstrate that $M^4$4TMD can achieve assessment accuracy comparable to that of senior dentists, while also exhibiting excellent robustness and adaptability. These findings represent an important step toward automated, expert-level multimodal evaluation for TMD by enabling concurrent assessment of multiple abnormalities. Notably, this approach is the first to develop a method for detecting DJD from MRI, offering a clear clinical benefit by supporting an MRI-first triage workflow that potentially reduces patient radiation exposure and examination costs.

Data and code availability statement

The codes for $M^4$4TMD development and evaluation in this study are available at https://github.com/Dentist-Lang/M44TMD_Repository.

Our well-trained model can be downloaded from htt ps://drive.google.com/file/d/1zy20BmENEHNvOh2PbjVHIyPhwEe2_wtk/view?usp=drive_link.

Interested researchers may submit requests for de-identified data to the corresponding author, subject to approval and data sharing agreements.

Funding

This work was supported by the Pioneer and Leading Goose

Technology Project of Zhejiang Province (2024C03094).

CRediT authorship contribution statement

Xinrui Lang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Rundong Zhang:** Visualization, Validation, Resources, Investigation, Formal analysis, Data curation. **Zhouhang Yuan:** Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Zheqi Lyu:** Writing – review & editing, Visualization, Validation, Software, Project administration, Methodology, Conceptualization. **Jiawei Wang:** Visualization, Software, Methodology. **Bo Qiao:** Resources, Investigation, Data curation. **Yanzhen Zhang:** Writing – review & editing, Resources, Project administration, Data curation. **Zhengxing Huang:** Writing – review & editing, Conceptualization. **Fan Yang:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jdent.2025.106322.

References

[1] E. Schiffman, R. Ohrbach, E. Truelove, J. Look, G. Anderson, J.-P. Goulet, T. List, P. Svensson, Y. Gonzalez, F. Lobbezoo, A. Michelotti, S.L. Brooks, W. Ceusters, M. Drangsholt, D. Ettlin, C. Gaul, L.J. Goldberg, J.A. Haythornthwaite, L. Hollender, R. Jensen, M.T. John, A. De Laat, R. de Leeuw, W. Maixner, M. van der Meulen, G.M. Murray, D.R. Nixdorf, S. Palla, A. Petersson, P. Pionchon, B. Smith, C.M. Visscher, J. Zakrzewska, S.F. Dworkin, 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 (2014) 6–27.

[2] L.F. Valesan, C.D. Da-Cas, J.C. Réus, A.C.S. Denardin, R.R. Garanhani, D. Bonotto, E. Januzzi, B.D.M. de Souza, Prevalence of temporomandibular joint disorders: a systematic review and meta-analysis, Clin Oral Invest 25 (2021) 441–453, https://doi.org/10.1007/s00784-020-03710-w.

[3] S. Beaumont, K. Garg, A. Gokhale, N. Heaphy, Temporomandibular Disorder: a practical guide for dental practitioners in diagnosis and management, Aust. Dent. J. 65 (2020) 172–180, https://doi.org/10.1111/adj.12785.

[4] S. Kandasamy, D.J. Rinchuse, C.S. Greene, L.E. Johnston, Temporomandibular disorders and orthodontics: What have we learned from 1992-2022? Am J Orthod Dentofac. Orthop 161 (2022) 769–774, https://doi.org/10.1016/j.ajodo.2021.12.011.

[5] J.W. Busse, R. Casassus, A. Carrasco-Labra, J. Durham, D. Mock, J.M. Zakrzewska, C. Palmer, C.F. Samer, M. Coen, B. Guevremont, T. Hoppe, G.H. Guyatt, H. N. Crandon, L. Yao, B. Sadeghirad, P.O. Vandvik, R.A.C. Siemieniuk, L. Lytvyn, B. S. Hunskaar, T. Agoritsas, Management of chronic pain associated with temporomandibular disorders: a clinical practice guideline, BMJ 383 (2023), https://doi.org/10.1136/bmj-2023-076227.

[6] S.J. Scrivani, D.A. Keith, L.B. Kaban, Temporomandibular Disorders, N. Engl. J. Med. (2008), https://doi.org/10.1056/NEJMra0802472.

[7] J. Xu, D. Wang, C. Yang, F. Wang, M. Wang, Reconstructed magnetic resonance image-based effusion volume assessment for temporomandibular joint arthralgia, J. Oral Rehabil. 50 (2023) 1202–1210, https://doi.org/10.1111/joor.13551.

[8] J.M. Petscavage-Thomas, E.A. Walker, Unlocking the Jaw: advanced imaging of the temporomandibular joint, Am. J. Roentgenol. 203 (2014) 1047–1058, https://doi.org/10.2214/AJR.13.12177.

[9] R.L. Gauer, M.J. Semidey, Diagnosis and Treatment of Temporomandibular Disorders, Am. Fam. Physician 91 (2015) 378–386.

[10] J. Bianchi, J. Roberto Gonçalves, A. Carlos de Oliveira Ruellas, J. Vieira Pastana Bianchi, L.M. Ashman, M. Yatabe, E. Benavides, F.N. Soki, L.H.S. Cevidanes, Radiographic interpretation using high-resolution Cbct to diagnose degenerative temporomandibular joint disease, PLoS. One 16 (2021) e0255937, https://doi.org/10.1371/journal.pone.0255937.

[11] M. Ahmad, L. Hollender, Q.Anderson Odont, K. Kartha, R.K. Ohrbach, E. L. Truelove, M.T. John, E.L. Schiffman, Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD): Development of Image Analysis Criteria and Examiner Reliability for Image Analysis, Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod. 107 (2009) 844–860, https://doi.org/10.1016/j.tripleo.2009.02.023.

[12] E. Zain-Alabdeen, R. Alsadhan, A comparative study of accuracy of detection of surface osseous changes in the temporomandibular joint using multidetector CT and cone beam CT, Dentomaxillofac. Radiol. 41 (2012) 185–191, https://doi.org/10.1259/dmfr/24985971.

[13] S. Sang, N. Ameli, F.T. Almeida, R. Friesen, Association between clinical symptoms and MRI image findings in symptomatic temporomandibular joint (TMJ) disease: A systematic review, J. Craniomaxillofac. Surg. 52 (2024) 835–842, https://doi.org/10.1016/j.jcms.2024.04.006.

[14] Y.-H. Lee, Q.-S. Auh, S. Jeon, S.-W. Jang, T.-S. Kim, Distinguishing acute and chronic TMD in adolescent patients, Sci. Rep. 15 (2025) 37402, https://doi.org/10.1038/s41598-025-21302-0.

[15] S. Liu, L. Xu, S. Lu, M. Mao, L. Liu, B. Cai, Diagnostic performance of magnetic resonance imaging for degenerative temporomandibular joint disease, J. Oral Rehabil. 50 (2023) 24–30, https://doi.org/10.1111/joor.13386.

[16] S.C. Kiliç, N. Kiliç, F. Güven, M.A. Sümbüllü, Is magnetic resonance imaging or cone beam computed tomography alone adequate for the radiological diagnosis of symptomatic temporomandibular joint osteoarthritis? A retrospective study, Int. J. Oral Maxillofac. Surg. 0 (2023), https://doi.org/10.1016/j.ijom.2023.04.005.

[17] K. Yoon, J.-Y. Kim, S.-J. Kim, J.-K. Huh, J.-W. Kim, J. Choi, Explainable deep learning-based clinical decision support engine for MRI-based automated diagnosis of temporomandibular joint anterior disk displacement, Comput. Methods Programs Biomed. 233 (2023) 107465, https://doi.org/10.1016/j.cmpb.2023.107465.

[18] Y. Iwase, T. Sugiki, Y. Kise, M. Nishiyama, M. Nozawa, M. Fukuda, Y. Ariji, E. Ariji, Deep learning classification performance for diagnosing condylar osteoarthritis in patients with dentofacial deformities using panoramic temporomandibular joint projection images, Oral. Radiol. 40 (2024) 538–545, https://doi.org/10.1007/s11282-024-00768-0.

[19] L. Mourad, N. Aboelsaad, W.M. Talaat, N.M.H. Fahmy, H.H. Abdelrahman, Y. El-Mahallawy, Automatic detection of temporomandibular joint osteoarthritis radiographic features using deep learning artificial intelligence. A Diagnostic accuracy study, J. Stomatol. Oral. Maxillofac. Surg. 126 (2024) 102124, https://doi.org/10.1016/j.jormas.2024.102124.

[20] K.S. Lee, H.J. Kwak, J.M. Oh, N. Jha, Y.J. Kim, W. Kim, U.B. Baik, J.J. Ryu, Automated detection of TMJ osteoarthritis based on artificial intelligence, J. Dent. Res. 99 (2020) 1363–1367, https://doi.org/10.1177/0022034520936950.

[21] B. Lin, M. Cheng, S. Wang, F. Li, Q. Zhou, Automatic detection of anteriorly displaced temporomandibular joint discs on magnetic resonance images using a deep learning algorithm, Dentomaxillofacial Radiol 51 (2022) 20210341, https://doi.org/10.1259/dmfr.20210341.

[22] R. Rokhshad, H. Mohammad-Rahimi, F. Sohrabniya, B. Jafari, P. Shobeiri, I. A. Tsolakis, S.A. Ourang, A.S. Sultan, S.N. Khawaja, R. Bavarian, J.M. Palomo, Deep learning for temporomandibular joint arthropathies: A systematic review and meta-analysis, J. Oral Rehabil. 51 (2024) 1632–1644, https://doi.org/10.1111/joor.13701.

[23] J.N. Acosta, G.J. Falcone, P. Rajpurkar, E.J. Topol, Multimodal biomedical AI, Nat. Med. 28 (2022) 1773–1784, https://doi.org/10.1038/s41591-022-01981-2.

[24] Y.-H. Lee, S. Jeon, J.-H. Won, Q.-S. Auh, Y.-K. Noh, Automatic detection and visualization of temporomandibular joint effusion with deep neural network, Sci. Rep. 14 (2024) 18865, https://doi.org/10.1038/s41598-024-69848-9.

[25] M. Nozawa, M. Fukuda, S. Kotaki, M. Araragi, H. Akiyama, Y. Ariji, Can temporomandibular joint osteoarthritis be diagnosed on MRI proton density-weighted images with diagnostic support from the latest deep learning classification models? Dentomaxillofacial Radiol 54 (2025) 56–63, https://doi.org/10.1093/dmfr/twae040.

[26] F. Schwendicke, T. Singh, J.-H. Lee, R. Gaudin, A. Chaurasia, T. Wiegand, S. Uribe, J. Krois, Artificial intelligence in dental research: Checklist for authors, reviewers, readers, J. Dent. 107 (2021) 103610, https://doi.org/10.1016/j.jdent.2021.103610.

[27] P.M. Bossuyt, J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L. Irwig, J. G. Lijmer, D. Moher, D. Rennie, H.C.W. de Vet, H.Y. Kressel, N. Rifai, R.M. Golub, D.G. Altman, L. Hooft, D.A. Korevaar, J.F. Cohen, F. the S. Group, STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies, Radiology. (2015). https://pubs.rsna.org/doi/10.1148/radiol.2015151516.

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal Loss for Dense Object Detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988, https://doi.org/10.1109/ICCV.2017.324.

[30] T. Benil, R. Krishna, T.P. Sariki, P. Yashika, S. Saraogi, S. Saraogi, Detect pre-cancerous tongue lesions for early oral cancer diagnosis using deep learning algorithm, Sci. Rep. 15 (2025) 41828, https://doi.org/10.1038/s41598-025-25925-1.

[31] K.K. Bressem, L.C. Adams, C. Erxleben, B. Hamm, S.M. Niehues, J.L. Vahldiek, Comparing different deep learning architectures for classification of chest radiographs, Sci. Rep. 10 (2020) 13590, https://doi.org/10.1038/s41598-020-70479-z.

[32] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv preprint arXiv:1409.1556 (2014), https://doi.org/10.48550/arXiv.1409.1556.

[33] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision

and pattern recognition, 2017, pp. 4700–4708, https://doi.org/10.1109/CVPR.2017.243.
[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826, https://doi.org/10.1109/CVPR.2016.308.
[35] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626, https://doi.org/10.1109/ICCV.2017.74.
[36] A. Petersson, What you can and cannot see in TMJ imaging – an overview related to the RDC/TMD diagnostic system, J. Oral Rehabil. 37 (2010) 771–778, https://doi.org/10.1111/j.1365-2842.2010.02108.x.
[37] C. Lee, K.J. Jeon, S.-S. Han, Y.H. Kim, Y.J. Choi, A. Lee, J.H. Choi, CT-like MRI using the zero-TE technique for osseous changes of the TMJ, Dentomaxillofacial Radiol 49 (2020) 20190272, https://doi.org/10.1259/dmfr.20190272.
[38] M.A.Q. Al-Saleh, J.L. Jaremko, N. Alsufyani, Z. Jibri, H. Lai, P.W. Major, Assessing the reliability of MRI-CBCT image registration to visualize temporomandibular joints, Dentomaxillofac. Radiol. 44 (2015) 20140244, https://doi.org/10.1259/dmfr.20140244.
[39] A. Bria, C. Marrocco, F. Tortorella, Addressing class imbalance in deep learning for small lesion detection on medical images, Comput. Biol. Med 120 (2020) 103735, https://doi.org/10.1016/j.compbiomed.2020.103735.
[40] R. Emshoff, A. Bertram, L. Hupp, A. Rudisch, Condylar erosion is predictive of painful closed lock of the temporomandibular joint: a magnetic resonance imaging study, Head. Face Med. 17 (2021) 40, https://doi.org/10.1186/s13005-021-00291-1.
[41] T. Burt, K. Button, H. Thom, R. Noveck, M. Munafò, The Burden of the "False-Negatives" in Clinical Development: Analyses of Current and Alternative Scenarios and Corrective Measures, Clin. Transl. Sci. 10 (2017) 470–479, https://doi.org/10.1111/cts.12478.
[42] Y. Li, H. Zhang, C. Bermudez, Y. Chen, B.A. Landman, Y. Vorobeychik, Anatomical context protects deep learning from adversarial perturbations in medical imaging, Neurocomput 379 (2020) 370–378, https://doi.org/10.1016/j.neucom.2019.10.085.
[43] L. Han, T. Tan, T. Zhang, Y. Huang, X. Wang, Y. Gao, J. Teuwen, R. Mann, Synthesis-based imaging-differentiation representation learning for multi-sequence 3D/4D MRI, Med. Image Anal. 92 (2024) 103044, https://doi.org/10.1016/j.media.2023.103044.
[44] H. Wang, T. Zhu, S. Ding, P. Wang, B. Chen, Feature-enhanced multi-sequence MRI-based fusion mechanism for breast tumor segmentation, Biomed. Signal. Process. Control 90 (2023), https://doi.org/10.1016/j.bspc.2023.105886.
[45] Y.-H. Lee, S. Jeon, T.-S. Kim, H.-S. Kim, Q.-S. Auh, Y.-K. Noh, Clinical features and subgroup patterns in elderly and super-elderly TMD patients, Sci. Rep. (2025), https://doi.org/10.1038/s41598-025-29749-x.
[46] Y. Li, S.L.R. Han, Z. Xu, Q. Cheng, P. Fan, Y. Zheng, J. Wang, X. Xiong, Pain, Function and Quality of Life in Temporomandibular Disorder Patients With Different Disc Positions, J. Oral Rehabil. 51 (2024) 2622–2633, https://doi.org/10.1111/joor.13861.
[47] C. Li, B. Chen, R. Zhang, Q. Zhang, Comparative study of clinical and MRI features of TMD patients with or without joint effusion: a retrospective study, BMC. Oral Health 24 (2024) 314, https://doi.org/10.1186/s12903-024-04065-4.
[48] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, Npj Digit. Med 3 (2020) 136, https://doi.org/10.1038/s41746-020-00341-z.
[49] D.C. Castro, I. Walker, B. Glocker, Causality matters in medical imaging, Nat. Commun. 11 (2020) 3673, https://doi.org/10.1038/s41467-020-17478-w.
[50] W.-Y. Mao, Y.-Y. Fang, Z.-Z. Wang, M.-Q. Liu, Y. Sun, H.-X. Wu, J. Lei, K.-Y. Fu, Automated diagnosis and classification of temporomandibular joint degenerative disease via artificial intelligence using CBCT imaging, J. Dent. 154 (2025) 105592, https://doi.org/10.1016/j.jdent.2025.105592.
[51] W.M. Talaat, S. Shetty, S. Al Bayatti, S. Talaat, L. Mourad, S. Shetty, A. Kaboudan, An artificial intelligence model for the radiographic diagnosis of osteoarthritis of the temporomandibular joint, Sci. Rep. 13 (2023) 15972, https://doi.org/10.1038/s41598-023-43277-6.
[52] Z. Cui, Y. Fang, L. Mei, B. Zhang, B. Yu, J. Liu, C. Jiang, Y. Sun, L. Ma, J. Huang, Y. Liu, Y. Zhao, C. Lian, Z. Ding, M. Zhu, D. Shen, A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images, Nat. Commun. 13 (2022) 2096, https://doi.org/10.1038/s41467-022-29637-2.
[53] Y.-W. Chang, J.K. Ryu, J.K. An, N. Choi, Y.M. Park, K.H. Ko, K. Han, Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study, Nat. Commun. 16 (2025) 2248, https://doi.org/10.1038/s41467-025-57469-3.
[54] J. Wang, Y. Yu, Y. Tan, H. Wan, N. Zheng, Z. He, L. Mao, W. Ren, K. Chen, Z. Lin, G. He, Y. Chen, R. Chen, H. Xu, K. Liu, Q. Yao, S. Fu, Y. Song, Q. Chen, L. Zuo, L. Wei, J. Wang, N. Ouyang, H. Yao, Artificial intelligence enables precision diagnosis of cervical cytology grades and cervical cancer, Nat. Commun. 15 (2024) 4369, https://doi.org/10.1038/s41467-024-48705-3.

Glossary

Abbreviation: Full Term
ADD: Anterior Disc Displacement
ADDWR: Anterior Disc Displacement with Reduction
ADDWoR: Anterior Disc Displacement without Reduction
AUC: Area Under the Curve
CBCT: Cone-Beam Computed Tomography
CI: Confidence Interval
CNN: Convolutional Neural Network
CT: Computed Tomography
DC/TMD: Diagnostic Criteria for Temporomandibular Disorders
DJD: Degenerative Joint Disease
DL: Deep Learning
FNR: False-Negative Rate
Grad-CAM: Gradient-weighted Class Activation Mapping
IoU: Intersection over Union
MLP: Multilayer Perceptron
MRI: Magnetic Resonance Imaging
NPV: Negative Predictive Value
PI: Permutation Importance
PR-AUC: Precision-Recall Area Under the Curve
ROC: Receiver Operating Characteristic
ROI: Region of Interest
STARD: Standards for Reporting Diagnostic Accuracy guidelines
TMD: Temporomandibular Disorders
TMJ: Temporomandibular Joint
closed-PD: Closed-mouth Proton Density-weighted
closed-T2W: Closed-mouth T2-weighted