From TMJ Localization to Detection of Osteoarthritic Features: An Anatomically Guided 3D Deep Learning Pipeline for Cone-Beam CT

Shradhdha Trivedi[0009-0002-8412-1258] [1], Vrundan Sojitra[2], and Mariela Padilla[0000-0003-3078-2544] [1]

[1] Herman Ostrow School of Dentistry, University of Southern California, Los Angeles, CA, USA
shradhdh@usc.edu

[2] Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA

Abstract. Automated assessment of temporomandibular joint osteoarthritis (TMJ OA) from cone-beam computed tomography (CBCT) remains limited by reliance on manually defined regions of interest and by uncertainty about how predicted anatomy should be encoded for 3D classification. This paper presents AutoTMJ-OA, a multi-stage CBCT framework for detecting radiographic osteoarthritic features as present or absent. Internal full-volume CBCT scans undergo sagittal TMJ localization, 3D ROI extraction, and SegResNet-based mandible segmentation; the public University of Michigan Deep Blue cohort enters directly at the ROI stage. The combined cohort included 216 patients and 432 joint-level samples, split at the patient level into 252 training, 87 validation, and 93 test joints. Four inputs generated from the same ROI and mandible mask—volume only, volume plus mask, masked volume, and masked volume plus mask—were evaluated at five cubic input sizes from 96^3 to 256^3 voxels using an identical classifier architecture and training protocol. Masked CBCT volume achieved the highest test-partition AUC at four of five input sizes, with the highest observed performance at 192^3 voxels (AUC 0.938; F1 0.926). Adding an explicit mask channel to an already masked volume did not improve performance. Because all configurations were evaluated on the same test partition, results are interpreted descriptively rather than as a pre-specified confirmatory comparison. These findings suggest that anatomical background suppression may contribute more to TMJ OA feature detection than an additional explicit shape channel.

Keywords: Temporomandibular joint · Osteoarthritis · Cone-beam CT · Anatomy-guided feature detection · Mandible segmentation · 3D deep learning

1 Introduction

Temporomandibular joint osteoarthritis (TMJ OA) is characterized by osseous remodeling of the mandibular condyle and adjacent articular structures. CBCT depicts erosion, osteophyte formation, sclerosis, subchondral cysts, and flattening [1, 19], but interpretation can be variable when changes are subtle or coexist with adaptive remodeling.

Deep learning has shown promise for TMJ OA detection and radiographic feature classification on sagittal CBCT images [12, 7, 20, 16]. Related work has also combined condyle localization with OA-feature detection on panoramic radiographs [11], quantitative 3D surface or imaging biomarkers have been investigated for earlier TMJ OA characterization [3], and patient-specific TMJ OA progression modeling has been reported using the Deep Blue cohort [2]. However, many approaches assume manual or pre-defined ROIs, and the optimal way to incorporate automatically predicted anatomy into a 3D CBCT prediction model remains unclear. An unmasked ROI preserves the fossa, eminence, and joint-space context but may contain irrelevant voxels and acquisition-specific shortcuts. Hard masking restricts the input to mandibular anatomy while preserving grayscale information describing cortical and trabecular structure. A second binary mask channel provides explicit geometry, although it may become redundant after hard masking.

Automated TMJ segmentation on CBCT has progressed from tracking-based slice methods to cascaded and multi-stage 3D networks for the condyle and glenoid fossa [14, 10, 21]. These studies support the feasibility of anatomy-aware TMJ analysis, but they primarily address segmentation rather than how predicted anatomy should be encoded for downstream detection of osteoarthritic features.

This paper presents AutoTMJ-OA, a framework for automated localization, segmentation, and detection of osteoarthritic features that tests how a predicted mandible mask should be presented to a 3D network and how volumetric input resolution affects binary feature-status prediction. For internal full-volume scans, it combines sagittal bounding-box detection, 3D mandible segmentation, and detection of osteoarthritic features. The Deep Blue cohort enters at the ROI stage because its released images are already TMJ-focused. Four anatomy-guided representations are evaluated at five cubic input sizes using identical data partitions, network architecture, and training protocol. This design isolates the effects of input representation and sampling resolution rather than network capacity.

2 Materials and Methods

2.1 Study Design, Ethics, and Cohorts

This retrospective study was approved by the Institutional Review Board at the University of Southern California (IRB #HS-25-00599). Informed consent was waived given the retrospective nature of the study and the use of de-identified imaging data. The cohort combined an internal institutional dataset with the

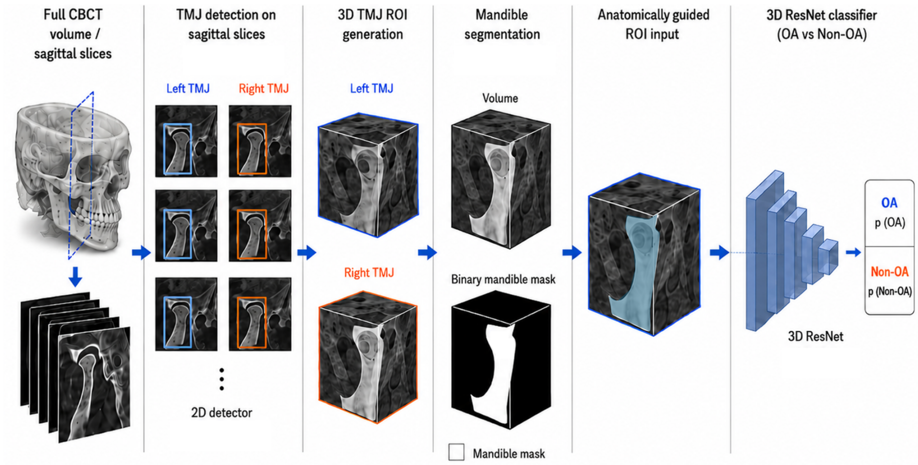

Fig. 1. Schematic overview of the proposed pipeline. Internal full-volume CBCT scans undergo sagittal TMJ localization and bilateral 3D ROI extraction. Deep Blue cases, which are already TMJ focused, enter at the ROI stage. Each ROI is processed by a 3D SegResNet to generate a binary mandible mask, from which four anatomically guided inputs are constructed for 3D ResNet prediction of OA features present versus absent.

Table 1. Patient-level partitions; counts are joint-level TMJ samples.

Site	Train	Validation	Test	Total
Internal institutional cohort	172	67	53	292
University of Michigan Deep Blue	80	20	40	140
Total	252	87	93	432

University of Michigan Deep Blue repository [4], totaling 216 patients and 432 evaluable TMJs. Left and right joints were separate joint-level samples, but all partitions were created at the patient level so that both sides remained together. The final split contained 252 training, 87 validation, and 93 held-out test joints (Table 1). Across the full set of joint-level volumes, approximately 55% were labeled OA-features-absent and 45% OA-features-present.

2.2 Joint-Level OA-Feature Labels

Labels were binary: radiographic OA features present (any grade) versus absent (Normal). The internal cohort used joint-level labels assigned under the institutional annotation protocol. Because Deep Blue provides patient-level OA status rather than separate labels for the left and right TMJs, a subject-matter expert reviewed each side independently on CBCT and assigned a joint-level OA-features-present or OA-features-absent label before model development. This side-specific relabeling aligned the reference standard with the joint-level prediction task and accounted for the fact that osteoarthritic findings may be asymmetric between the left and right TMJs.

2.3 Stage 1: TMJ Localization from Full-Volume CBCT

Automated anatomical localization was used only for the internal full-volume CBCT cohort. The Deep Blue release contains TMJ-focused ROI volumes and therefore did not require bounding-box detection. For the internal cohort, a slice-based two-dimensional detector analyzed sagittal CBCT slices, reducing computational burden relative to direct full-volume detection while matching the plane commonly used for TMJ interpretation.

Bounding-box localization was implemented in MMDetection [5] using Faster R-CNN [18] with a ResNet-50 backbone and feature pyramid network [13]. Manually annotated left- and right-TMJ boxes were stored in COCO format, with one TMJ box per positive sagittal slice. The detection dataset comprised 43,014 sagittal slices after three-fold augmentation, including 8,841 TMJ-positive slices (20.6%) and 34,173 negative slices (79.4%). The detection split was performed at the CBCT scan level before slice extraction and augmentation to avoid augmented or adjacent slices from the same scan appearing in both training and validation sets. After slice extraction and augmentation, this produced 34,410 training slices (80%) and 8,604 validation slices (20%), while preserving the positive-slice prevalence across partitions (20.7% and 19.9%, respectively). Detection-stage augmentation included random zoom-in and zoom-out transformations and shear transformations, applied jointly to each sagittal image and its bounding-box coordinates. These transformations were used to simulate variability in apparent TMJ scale, patient positioning, and slice geometry while preserving anatomically valid box supervision.

During inference, the detector was applied sequentially across sagittal slices. Predicted boxes from anatomically adjacent slices were aggregated, and their union defined the three-dimensional crop extent for each side. Each crop was expanded by a 5% margin and mapped back to the original volume coordinates to generate independent left and right TMJ ROIs; both sides were retained because OA-related findings can be unilateral or asymmetric. On the validation set, the detector achieved a mean IoU of 0.80 for left TMJs and 0.78 for right TMJs. Localization at IoU ≥ 0.5 demonstrated consistent detection across varying anatomical presentations and image-quality conditions.

2.4 Stage 2: Mandible Segmentation Training

Mandible annotations were initially generated from full CBCT volumes using the open-source DentalSegmentator model within 3D Slicer [6, 8]. DentalSegmentator was developed for multi-structure dento-maxillofacial CT/CBCT segmentation, including the mandible. All automated masks were reviewed and corrected by a subject-matter expert before model training. Corrected full-volume images and masks were cropped to TMJ ROIs using the corresponding ROI coordinates,

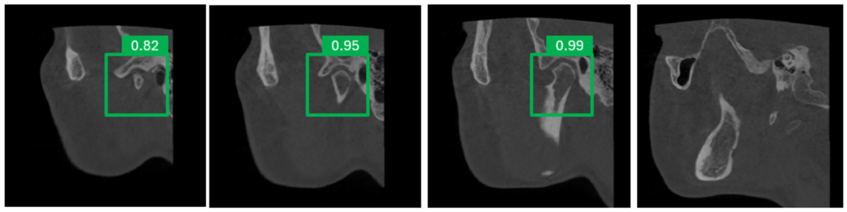

Fig. 2. Example outputs of the Stage-1 Faster R-CNN model on sagittal CBCT slices. Green boxes indicate predicted TMJ regions, and the displayed values are detector confidence scores. The rightmost panel illustrates a sagittal slice without a displayed TMJ detection.

and SegResNet was trained on ROI patches only. The segmentation cohort contained ROI patches from 93 internal CBCT volumes, divided into 68 training and 25 validation volumes.

A 3D SegResNet [17] was trained to produce a single binary mandible map using a combined Dice–binary-cross-entropy objective. The condylar head was not a separate class; within each TMJ ROI, the mask included the condylar process and visible superior ramus. Image and mask pairs underwent identical geometric preprocessing, with nearest-neighbor interpolation for masks. The largest connected component was retained and small holes were filled. On the 25-volume validation set, SegResNet achieved a Dice score of 0.876 and was then applied to ROIs from both cohorts.

2.5 Preprocessing and Input Representations

The internal CBCT cohort contained mixed native voxel spacings ranging from 0.15 to 0.30 mm. To standardize physical scale across acquisitions, internal volumes were resampled to 0.30-mm isotropic spacing before ROI generation and downstream analysis, using trilinear interpolation for image intensities and nearest-neighbor interpolation for masks. Deep Blue ROI volumes were reoriented and intensity-normalized using the same downstream preprocessing steps as internal ROIs. NRRD images and NIfTI masks were aligned in physical space. Intensities were clipped to $[-200, 1800]$ and normalized to $[0, 1]$.

Each extracted TMJ ROI retained the same anatomical field of view and was resized to one of five cubic network inputs: 96^3, 128^3, 160^3, 192^3, or 256^3 voxels. ROIs were cropped in physical space and resampled isotropically; cubic tensors were produced by resizing/padding the standardized field of view rather than intentionally deforming individual anatomical axes. Thus, the input-size experiment evaluated volumetric sampling resolution rather than different anatomical crops. Let I_s denote the normalized ROI at input size s and M_s the corresponding binary mandible mask. The masked image was $I_{M,s} = I_s \odot M_s + b(1 - M_s)$, where $b = -1$. Because image intensities were normalized to $[0, 1]$, this out-of-range constant provided an unambiguous suppressed-background value. Figure 3

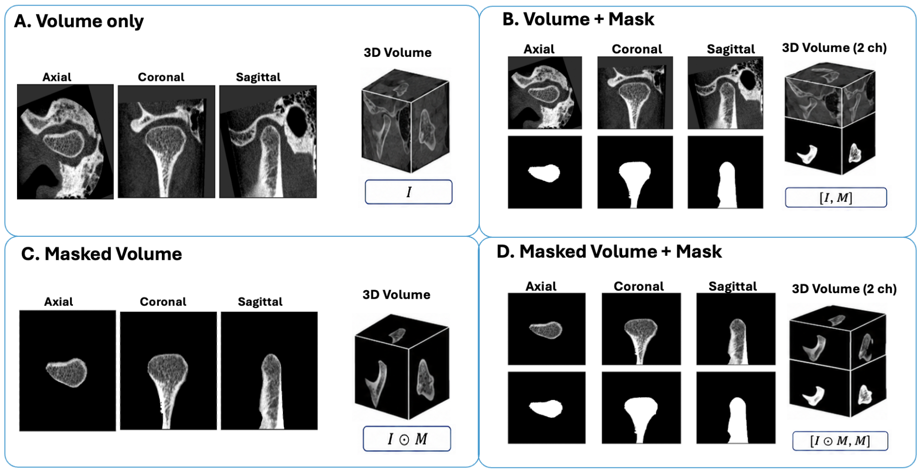

Fig. 3. Four inputs generated from the same TMJ ROI and mandible mask. At each cubic input size, the anatomical field of view is held fixed; only sampling resolution and representation change.

demonstrates a representative TMJ ROI CBCT volume, its corresponding binary mandible segmentation output, and the four derived input representations. Four inputs were generated at every size: (1) volume only, I_s; (2) volume plus mask, $[I_s, M_s]$; (3) masked volume, $I_{M,s}$; and (4) masked volume plus mask, $[I_{M,s}, M_s]$. The fourth representation was included as a redundancy ablation to test whether an explicit shape channel adds information after background voxels have already been suppressed.

2.6 3D Prediction Network and Controlled Experimental Protocol

All experiments used the same 3D ResNet architecture [9], comprising a $7 \times 7 \times 7$ stem, four residual stages with a base channel width of 32, global average pooling, and a single sigmoid output. The first convolution accepted either one or two channels according to the selected representation. Identical network architecture and training protocol were used across representations. A separate model was trained for every representation–input-size combination so that performance differences could be attributed to input encoding and volumetric sampling resolution rather than model capacity.

Training augmentation combined acquisition- and anatomy-aware perturbations: Gaussian blur followed by simulated down-sampling and up-sampling, random intensity scaling and offset perturbation, random mandible-mask erosion or dilation by 1–2 voxels, rotations up to $\pm 10°$, and translations up to ± 4 voxels. Models were trained on an NVIDIA A100 using AdamW [15] with learning rate 10^{-4}, weight decay 10^{-4}, bfloat16 mixed precision, linear warmup, and cosine decay. Class imbalance was handled with weighted binary cross-entropy,

$$\mathcal{L}_{\mathrm{WBCE}} = -w_1 y \log(\hat{p}) - w_0(1-y)\log(1-\hat{p}), \tag{1}$$

where $y \in \{0,1\}$ is the reference label, $\hat{p}$ is the predicted probability that OA features are present, and w_0, w_1 are class weights derived from the training partition.

The 252-joint training partition was used for model fitting and the 87-joint validation partition for checkpoint selection. For each of the 20 combinations of four representations and five input sizes, the retained checkpoint was evaluated on the 93-joint test partition. The primary endpoint was area under the receiver operating characteristic curve (AUC), with F1 score reported as a complementary prediction metric. Because the same test partition was used to compare all 20 configurations, these comparisons were treated as descriptive analyses; the configuration with the highest observed test AUC was not interpreted as a prespecified confirmatory winner, and small between-configuration differences were not considered evidence of statistical superiority.

3 Results

3.1 Cohort and Upstream Model Performance

The study included 216 patients and 432 joint-level samples, with 252 joints assigned to training, 87 to validation, and 93 to the held-out test set. Approximately 55% of joint-level volumes were OA-features-absent and 45% were OA-features-present. The internal test subset contained 53 joints and the Deep Blue test subset contained 40 joints. For internal full-volume CBCT scans, the Faster R-CNN detector achieved mean validation IoU values of 0.80 for left TMJs and 0.78 for right TMJs; localization at IoU ≥ 0.5 remained consistent across varying anatomical presentations and image-quality conditions. On the 25-volume segmentation validation set, the 3D SegResNet achieved a mandible Dice score of 0.876.

3.2 Test-Set Comparison Across Input Representations

Table 2 summarizes test-set AUC and F1 across four input representations and five cubic input sizes. Masked CBCT volume achieved the highest AUC at 96^3, 160^3, 192^3, and 256^3, whereas masked CBCT volume plus segmentation mask was marginally higher at 128^3 (0.911 versus 0.908). The highest observed performance was obtained with masked CBCT volume at 192^3, with an AUC of 0.938 and an F1 score of 0.926. At the same input size, CBCT volume plus segmentation mask achieved an AUC of 0.930 and F1 of 0.922; masked CBCT volume plus segmentation mask achieved an AUC of 0.934 and F1 of 0.920; and CBCT volume only achieved an AUC of 0.904 and F1 of 0.900.

Performance generally increased from 96^3 through 192^3 and remained similar at 256^3. The explicit segmentation-mask channel improved performance when

Table 2. Test-set AUC/F1 across input representations and cubic input sizes.

Input	$\mathbf{96^3}$	$\mathbf{128^3}$	$\mathbf{160^3}$	$\mathbf{192^3}$	$\mathbf{256^3}$
Masked vol.	0.884/0.868	0.908/0.892	0.931/0.914	**0.938/0.926**	0.935/0.924
Vol.+mask	0.848/0.859	0.871/0.865	0.902/0.910	0.930/0.922	0.928/0.925
Masked vol.+mask	0.873/0.864	0.911/0.890	0.919/0.902	0.934/0.920	0.932/0.923
Volume only	0.791/0.812	0.824/0.834	0.856/0.849	0.904/0.900	0.902/0.899

paired with the unmasked CBCT volume compared with CBCT volume only, and yielded the highest F1 at 256^3 (0.925 versus 0.924 for masked volume). In contrast, adding the mask to an already masked CBCT volume did not improve the highest observed overall performance. Because all configurations were assessed on the same test partition and interval estimates were not available, close numerical differences are interpreted descriptively.

4 Discussion

The descriptive test-set comparison indicates that input representation affects detection of OA-associated radiographic features. Masked volume produced the highest AUC at four of five sizes and the highest observed result at 192^3 (AUC 0.938; F1 0.926), while masked volume plus mask was marginally higher at 128^3. Background suppression may focus learning on cortical and trabecular information and reduce acquisition-related variation. Because all configurations were compared on the same test partition without interval estimates, small differences should not be interpreted as confirmatory superiority.

Adding the mask to an unmasked ROI improved performance over volume only, supporting explicit anatomical guidance when whole-joint context is retained. Adding the same mask to an already masked image did not improve AUC, suggesting that the masked volume already conveys the mandible's spatial support.

Performance increased from 96^3 to 192^3 and changed little at 256^3, suggesting a benefit from preserving fine osseous detail without establishing a universally optimal resolution. Internal scans were localized by Faster R-CNN and segmented by SegResNet (Dice 0.876); TMJ-focused Deep Blue images bypassed detection but underwent the same downstream stages. Prior CBCT studies have demonstrated automated condyle and glenoid-fossa segmentation [10, 21]; this study extends that work by using predicted mandibular anatomy to construct controlled inputs for OA-feature detection.

Limitations include multi-source rather than untouched external testing, a binary endpoint, and dependence on predicted masks. Deep Blue side-specific labels were assigned by one expert without second-reader reliability assessment. Table 2 reports point estimates without confidence intervals, and hard masking may remove temporal-bone or joint-space findings. Future work should include independent multi-vendor validation, multi-reader reference standards, feature-level prediction, calibration, and dual-stream models retaining whole-joint context.

5 Conclusion

AutoTMJ-OA combines automated TMJ localization, mandible segmentation, and binary OA-feature detection. In a descriptive comparison of 20 configurations, masked volume was generally strongest, with the highest observed result at 192^3 (AUC 0.938; F1 0.926). The findings support anatomical background suppression and adequate volumetric detail, but require independent confirmatory evaluation.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Ahmad, M., Hollender, L., Anderson, Q., et al.: Research diagnostic criteria for temporomandibular disorders (RDC/TMD): Development of image analysis criteria and examiner reliability for image analysis. Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology **107**(6), 844–860 (2009). https://doi.org/10.1016/j.tripleo.2009.02.023
2. Al Turkestani, N., Li, T., Bianchi, J., et al.: A comprehensive patient-specific prediction model for temporomandibular joint osteoarthritis progression. Proceedings of the National Academy of Sciences **121**(8), e2306132121 (2024). https://doi.org/10.1073/pnas.2306132121
3. Bianchi, J., de Oliveira Ruellas, A.C., Goncalves, J.R., et al.: Osteoarthritis of the temporomandibular joint can be diagnosed earlier using biomarkers and machine learning. Scientific Reports **10**(1), 8012 (2020). https://doi.org/10.1038/s41598-020-64942-0
4. Cevidanes, L.: A comprehensive patient-specific prediction model for temporomandibular joint osteoarthritis progression: Data set (2024). https://doi.org/10.7302/xc32-4d53, `https://deepblue.lib.umich.edu/data/concern/data_sets/2v23vv279`
5. Chen, K., Wang, J., Pang, J., et al.: MMDetection: Open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019), `https://arxiv.org/abs/1906.07155`
6. Dot, G., Chaurasia, A., Dubois, G., et al.: DentalSegmentator: Robust open source deep learning-based CT and CBCT image segmentation. Journal of Dentistry **147**, 105130 (2024). https://doi.org/10.1016/j.jdent.2024.105130
7. Eser, G., Duman, S.B., Bayrakdar, I.S., Celik, O.: Classification of temporomandibular joint osteoarthritis on cone beam computed tomography images using artificial intelligence system. Journal of Oral Rehabilitation **50**(9), 758–766 (2023). https://doi.org/10.1111/joor.13481
8. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., et al.: 3D Slicer as an image computing platform for the quantitative imaging network. Magnetic Resonance Imaging **30**(9), 1323–1341 (2012). https://doi.org/10.1016/j.mri.2012.05.001

9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
10. Jha, N., Kim, T., Ham, S., Baek, S.H., Sung, S.J., Kim, Y.J., Kim, N.: Fully automated condyle segmentation using 3D convolutional neural networks. Scientific Reports **12**(1), 20590 (2022). https://doi.org/10.1038/s41598-022-24164-y
11. Kim, D., Choi, E., Jeong, H.G., Chang, J., Youm, S.: Expert system for mandibular condyle detection and osteoarthritis classification in panoramic imaging using R-CNN and CNN. Applied Sciences **10**(21), 7464 (2020). https://doi.org/10.3390/app10217464
12. Lee, K.S., Kwak, H.J., Oh, J.M., Jha, N., Kim, Y.J., Kim, W., Baik, U.B., Ryu, J.J.: Automated detection of TMJ osteoarthritis based on artificial intelligence. Journal of Dental Research **99**(12), 1363–1367 (2020). https://doi.org/10.1177/0022034520936950
13. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017). https://doi.org/10.1109/CVPR.2017.106
14. Liu, Y., Lu, Y., Fan, Y., Mao, L.: Tracking-based deep learning method for temporomandibular joint segmentation. Annals of Translational Medicine **9**(6), 467 (2021). https://doi.org/10.21037/atm-21-319
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
16. Mourad, L., Aboelsaad, N., Talaat, W.M., Fahmy, N.M.H., Abdelrahman, H.H., El-Mahallawy, Y.: Automatic detection of temporomandibular joint osteoarthritis radiographic features using deep learning artificial intelligence: A diagnostic accuracy study. Journal of Stomatology, Oral and Maxillofacial Surgery **126**(4), 102124 (2025). https://doi.org/10.1016/j.jormas.2024.102124
17. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Lecture Notes in Computer Science, vol. 11384, pp. 311–320. Springer (2019). https://doi.org/10.1007/978-3-030-11726-9_28
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. vol. 28, pp. 91–99 (2015)
19. Schiffman, E., Ohrbach, R., Truelove, E., et al.: Diagnostic criteria for temporomandibular disorders (DC/TMD) for clinical and research applications. Journal of Oral & Facial Pain and Headache **28**(1), 6–27 (2014). https://doi.org/10.11607/jop.1151
20. Talaat, W.M., Shetty, S., Al Bayatti, S., Talaat, S., Mourad, L., Shetty, S., Kaboudan, A.: An artificial intelligence model for the radiographic diagnosis of osteoarthritis of the temporomandibular joint. Scientific Reports **13**(1), 15972 (2023). https://doi.org/10.1038/s41598-023-43277-6
21. Vinayahalingam, S., Berends, B., Baan, F., Anssari Moin, D., van Luijn, R., Bergé, S., Xi, T.: Deep learning for automated segmentation of the temporomandibular joint. Journal of Dentistry **132**, 104475 (2023). https://doi.org/10.1016/j.jdent.2023.104475