CENTIVE
Conference on Electrical Engineering, Informatics
, Industrial Technology, and Creative Media
ISBN : 978-602-53004-0-0

Automated Detection of MRONJ Lesions in Panoramic Dental X-rays Using Candidate Region Identification and Semantic Segmentation

Manami Inoue[1*], Kento Morita[2], Yasuaki Sadakane[3],
Takumi Hasegawa[3], Masaya Akashi[3], Tetsushi Wakabayashi[2],
[1]Graduate School of Regional Innovation, Mie University
[2]Graduate School of Engineering, Mie University
[3]Department of Oral and Maxillofacial Surgery, Kobe University Graduate School of Medicine
[1,2]1577 Kurimamachiyacho, Tsu, Mie 514–8507, Japan
[3]7–5–2 Kusunokichou, Kobe Chuo-ku, Hyogo 650–0017, Japan
[*]624m001@m.mie-u.ac.jp

Abstract — Medication-related osteonecrosis of the jaw (MRONJ) is a severe adverse effect associated with the administration of bone-modifying agents, such as bisphosphonates (BP) and denosumab (Dmab), and angiogenesis inhibitors. Despite the advancements in therapeutic agents, the incidence of MRONJ has increased, as medication remains a primary risk factor. In most cases, MRONJ is diagnosed at an advanced stage, where portions of the jawbone become exposed in the oral cavity, interfering with both primary disease management and MRONJ treatment. Therefore, early detection and treatment prior to progression are critical for improving patient outcomes and reducing treatment complexity. In Japan, the low penetration of dental CT limits the feasibility of 3D diagnostic imaging in routine practice in dental clinics. Therefore, this study proposes a diagnostic method that relies solely on panoramic X-ray images to automatically predict MRONJ lesions. The proposed method first performs pre-processing to extract the mouth region, and then compares two approaches for MRONJ lesion segmentation. The first approach subdivides the mouth region into patches and utilizes patch-based classification to identify candidate regions before MRNOJ lesion segmentation. The second approach employs the masked vision transformer (Masked-ViT) to estimate the probability of MRONJ lesion presence across the image, and then segmentation is applied to high probability areas. On our panoramic X-ray image dataset consisting of 118 MRONJ patients, the patch-based method achieved a maximum Dice Similarity Coefficient (DSC) of 0.70, outperforming the method using Masked-ViT. Although promising, further enhancements are necessary to meet the requirements for clinical use.

Keywords – Deep learning, masked-ViT, panoramic dental X-ray, semantic segmentation

I. INTRODUCTION

Medication-related osteonecrosis of the jaw (MRONJ) is a severe adverse effect associated with the use of bone-modifying agents, such as bisphosphonates (BP) and denosumab (Dmab), and angiogenesis inhibitors. BP and Dmab are primarily administered to reduce the risk of skeletal complications in patients with bone loss due to long-term cancer treatment, osteoporosis, or malignant bone disease [1]. In 2003, osteonecrosis of the jaw (ONJ) was first reported in a patient treated with BP. Subsequently, similar cases were also observed in patients treated with Dmab. Initially, ONJ caused by BP was termed bisphosphonate-related osteonecrosis of the jaw (BRONJ), while that caused by Dmab was referred to as denosumab-related osteonecrosis of the jaw (DRONJ). Collectively, these were grouped under the term antiresorptive agent-related ONJ (ARONJ). However, following additional reports of ONJ associated with anti-angiogenic agents such as bevacizumab and sunitinib, the terminology was unified under MRONJ. More recently, ONJ has also been reported in patients treated with remosozumab, a drug that exhibits both bone formation-promoting and antiresorptive properties. As therapeutic agents continue to evolve, the types and administration routes of medications linked to ONJ are becoming increasingly diverse, and the incidence of MRONJ has shown a rising trend.

The number of patients diagnosed with MRONJ in

Japan increased significantly from fewer than 1,600 cases per year in 2013 to nearly 7,000 cases in 2019 [2]. Japan also has the highest average life expectancy in the world at 84.46 years, which contributes to the rising prevalence of age-related diseases such as cancer and osteoporosis. In clinical practice, treatment planning for patients with maxillofacial diseases often requires a trade-off between addressing the primary disease and managing MRONJ. This decision places a considerable burden on doctors. Furthermore, MRONJ is difficult to treat, as it typically manifests as exposed bone in the maxillofacial region. These challenges highlight the importance of early detection and intervention to reduce the clinical burden and improve patient outcomes.

Recent studies have reported that necrotic lesions observed intraoperatively are often more extensive than those identified in pre-operative radiographic examinations [3], [4], indicating a limitation in the accuracy of current image-based diagnosis. Although no definitive pathognomonic imaging features for MRONJ have been established [5], radiographic evaluation remains essential for pre-operative planning as well as for routine monitoring of disease progression and treatment response.

Three-dimensional modalities, such as cone-beam computed tomography (CBCT) and computed tomography (CT), have demonstrated superior capability in evaluating fine anatomical structures and assessing the extent of MRONJ lesions compared to two-dimensional panoramic X-ray radiographs [6]. However, in Japan, the adoption rate of dental CT systems remains low due to their high installation and operational costs. To address this limitation, this study proposes a diagnostic method that utilizes only panoramic X-ray radiographs to automatically predict MRONJ lesions without relying on CT-based imaging.

In this study, we employ YOLO [7] to automatically detect the mouth region as a bounding box (bbox) from panoramic X-ray radiographs of MRONJ patients. The detected region is then subdivided into multiple grids to localize potential MRONJ candidate areas. Subsequently, semantic segmentation is performed to delineate the MRONJ lesions at the pixel level.

II. PRELIMINARIES

A. Subjects

This study utilizes dental panoramic X-ray images from 118 patients diagnosed with MRONJ, collected at Kobe University Hospital. Under the supervision of medical experts, MRONJ lesion masks, including osteolysis, osteosclerosis, and sequestrum lesion, were manually annotated for all images. The use of these images was approved by the Ethics Committee of Kobe University Graduate School of Medicine. Figure 1 presents an example of a panoramic X-ray image along with its corresponding MRONJ lesion mask, highlighted by a yellow contour.

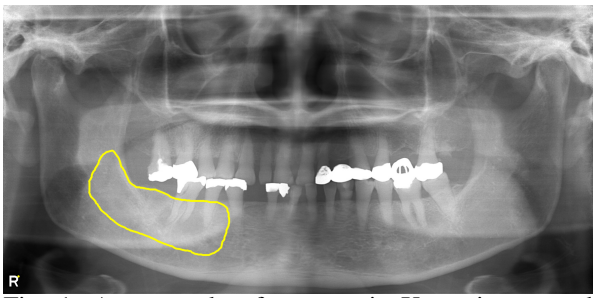


Fig. 1: An example of panoramic X-ray image and MRONJ lesion mask (highlighted by a yellow contour).

B. Masked Vision Transformer

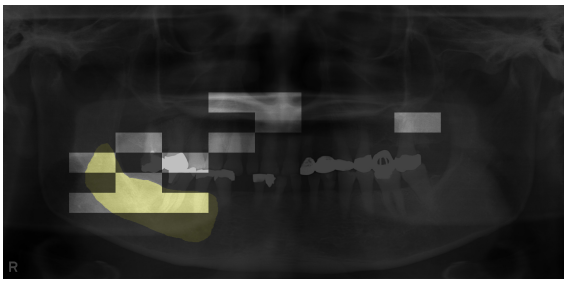
In the image recognition field, Masked Autoencoder (Masked AE) [13] has recently been proposed, drawing inspiration from the self-supervised learning approach in the natural language processing field, known as Masked Language Modeling (MLM). Masked AEs adopt the autoencoder architecture that learns to reconstruct randomly masked regions of input images, and have demonstrated improved performance compared to previous mask-based models such as BEiT [14]. Since random masking enables efficient learning from partially visible data, Masked AEs have gained attention as a promising approach for developing high-performance deep-learning models using limited training data.

In a related study, Lei *et al.* investigated a self pre-training paradigm with Masked AE for medical image analysis tasks, such as chest X-ray diagnosis, abdominal CT multi-organ segmentation, and MRI brain tumor segmentation [15], in which they demonstrated that Masked AE pre-training improved SOTA classification performance on a diverse set of medical image analysis tasks and the effectiveness of Masked AE on 3D medical images including both CTs and MRIs. However, Masked AE has not yet demonstrated its effectiveness on the X-ray images, which are 2D medical images used in our study.

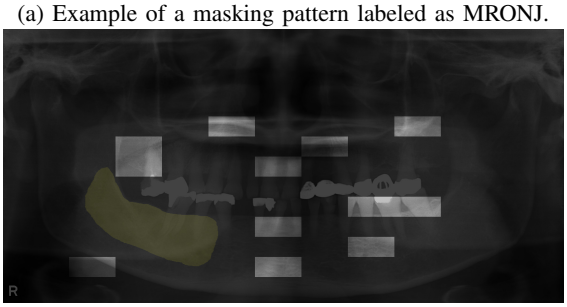
Therefore, this study utilizes the Masked Vision Transformer (Masked-ViT) inspired by the Masked AE, where image patches are randomly masked. The input image is first divided into a grid of patches, and a subset of these patches is randomly selected and masked. As shown in Fig. 2, applying the patch-wise random masking to an MRONJ image results in the generation of both masked MRONJ (Fig. 2a) images and masked non-MRONJ (Fig. 2b) images.

III. RESEARCH METHOD

Figure 3 shows a brief diagram of the proposed method, which automatically performs the preprocessing and MRONJ region prediction using dental X-ray images. The proposed method first determines



(a) Example of a masking pattern labeled as MRONJ.



(b) Example of a masking pattern labeled as normal.

Fig. 2: Examples of patch-wise random masked panoramic X-ray images. The dark patches indicate masked regions, while bright patches retain the original image content. The yellow-highlighted area denotes the MRONJ lesion.

the mouth region bounding box, which is the minimum required region for the MRONJ region prediction, using YOLO. After that, this paper proposes two methods of predicting the MRONJ region in the dental X-ray.

The first method predicts the MRONJ pixel existence for each subdivided 10 image patches using the generic binary classification Convolutional Neural Network (CNN). When a patch is classified as the MRONJ, it is fed into the semantic segmentation CNN, U-Net [16], to predict the MRONJ region pixel-by-pixel. In the second method, we divide the bounding box into 8×8, and $S\%$ of them are randomly blacked out to generate multiple mask images. Since the proposed ViT provides the degree of MRONJ or non-MRONJ class associations in the 0.0 to 1.0 range, the MRONJ class association is accumulated to enhance the MRONJ area. Finally, U-Net predicts the detailed MRONJ region from the obtained MRONJ high-risk area.

A. Mouth region detection using YOLO

Because the area around the eye socket and cervical spine are not necessary for predicting the MRONJ lesions, we need to extract the necessary region from the entire X-ray image, capturing a wide range from the eye socket to the mandibular base. As a preprocessing, this section applies the object detection CNN called YOLO to determine the bounding box (bbox in the following) of the mouth region, which is necessary for the MRONJ diagnosis. Since YOLO detects multiple bboxes in an image, the thresholding of the bbox confidence at a threshold of 0.5 suppresses unnecessary bboxes. In the experiment, manually determined bboxes are utilized for the model training, while YOLO-detected bboxes are used for the performance evaluation using the test dataset.

B. Method 1: Patch-based MRONJ detection and segmentation using CNN

This section describes the patch-based MRONJ detection using CNN and semantic segmentation of MRONJ pixels in patches classified as MRONJ patches.

The detected mouth region in the section III-A has approximately 1,500 pixels in width and 700 pixels in height. Because the detected bbox is relatively large compared to the input size of standard classification CNN architectures, and the MRONJ region occupies only a small portion (approximately 10-40%) of the bbox, the region is divided into 10 equal sections arranged in two rows and five columns.

We compare the performance of several fine-tuned CNNs, including VGG16 [8], Inception [9], EfficientNet [10], and MobileViT [11], for binary classification to determine whether each subdivided patch contains MRONJ lesions. For training, each patch is labeled as 1 when it contains any MRONJ lesion pixel, and 0 otherwise. These labeled patches are then used to train the classification models.

For each patch classified as containing MRONJ lesion pixels, pixel-level segmentation is performed using U-Net, a fully convolutional network (FCN) with an encoder-decoder architecture. The segmented patches are then stitched back into the input panoramic X-ray image to obtain the full-image MRONJ lesion.

C. Method 2: MRONJ lesion detection using Masked-ViT

Due to the limited number of MRONJ patients in Japan, our dataset comprised only 118 cases, which is insufficient for training conventional CNN models effectively. To address this limitation, we adopt a Masked-ViT trained on panoramic X-ray images with randomly masked regions.

The input to the proposed Masked-ViT consists of 64 tokens, requiring the YOLO-detected mouth region to be subdivided into an 8×8 grid. To generate training samples, we apply patch-wise random masking, where a fixed percentage of grid cells are randomly blacked out. During training, masked images are sampled to maintain a balanced number of MRONJ-positive and negative samples.

The ViT model is trained to predict whether an input randomly masked image contains any MRONJ lesion pixels. In this study, we use a pre-trained ViT and perform transfer learning to MLP (Multilayer Perceptron) Head consisting of layer normalization and fully-connected layers. Each blacked-out

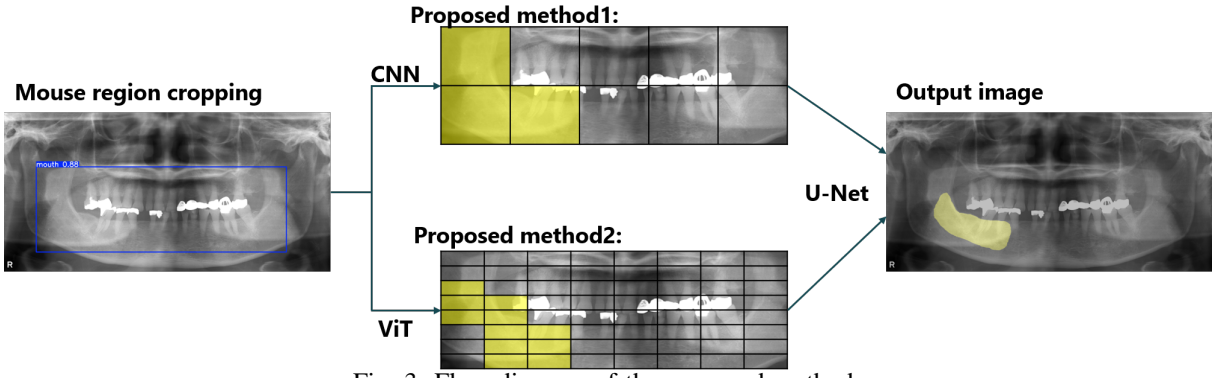


Fig. 3: Flow diagram of the proposed method.

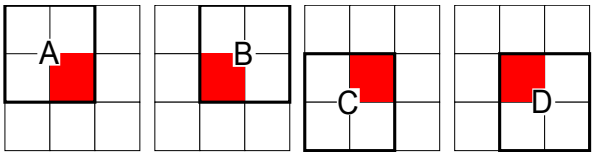


Fig. 4: Inference region of U-net based on Masked-ViT inference results. Region of inference using U-Net extract four types of 2×2 neighboring grid regions surrounding each MRONJ candidates (indicated in red) selected by Masked-ViT.

grid cell is represented by a `[MASK]` token, and the ViT outputs a probability score ranging from 0.0 to 1.0, indicating the likelihood of MRONJ presence. Although the randomly masked image is labeled as an MRONJ image, since it typically includes both MRONJ and non-MRONJ lesion pixels, aggregating predictions across multiple masked versions enables the construction of a pixel-wise MRONJ probability map across the image. To identify MRONJ candidate regions from the probability map generated by the Masked-ViT, min-max normalization is applied to scale the probability values to the range $[0, 1]$. Each grid cell's normalized values are then aggregated, and cells with average probability greater than the threshold (0.35 was obtained in the experiment) are selected as MRONJ candidates. For pixel-level MRONJ lesion segmentation, U-Net is applied to these candidate regions. However, each individual grid cell in the 8×8 division corresponds to a region of approximately 200×100 pixels, which is insufficient for stable training and inference. To mitigate this, we extract four types of 2×2 neighboring grid regions surrounding each high-probability MRONJ grid (indicated in red in Fig. 4). U-Net processes these expanded patches, and the resulting outputs are averaged to produce a pixel-level MRONJ segmentation map aligned with the original image resolution.

IV. EXPERIMENT

This section describes the details of the experimental settings and results of the two proposed MRONJ region detection methods and their comparison.

Table 1: Dataset summary

Dataset	Patients	Patch images	
		MRONJ	non-MRONJ
Training	58	208	372
Validation	30	108	192
Test	30	114	186

A. Dataset

The following experiments utilize 118 panoramic X-ray images collected from 118 patients diagnosed with MRONJ. These images are divided into training, validation, and test sets, as summarized in Table 1. The dataset was partitioned based on the size of MRONJ lesions to maintain a balanced number of MRONJ and non-MRONJ patches across all subsets.

B. Evaluation metrics

In the following experiments, binary classification results obtained using either CNNs or the proposed Masked ViT are evaluated using the precision, recall, and F1-score. The semantic segmentation performance is assessed using the Dice Similarity Coefficient (DSC), defined as follows:

$$DSC(P, Q) = \frac{2|P \cap Q|}{|P| + |Q|} \quad (1)$$

where, P represents the predicted MRONJ lesion, and Q denotes the manually annotated MRONJ lesion.

C. MRONJ lesion detection using CNN

1) MRONJ candidate patch identification results

To identify MRONJ lesion candidates, the first proposed method subdivides the YOLO-detected mouth region bbox into multiple image patches. Accordingly, this section evaluates the performance of various CNN-based binary classifiers in determining whether a given patch contains any MRONJ pixels.

Since the optimal parameters, such as the optimizer, its hyperparameters, and the loss function, vary across CNN architectures, these were experimentally selected for each model. The models were then compared using their best-performing parameters based on the F1-score obtained from the validation dataset. Table 2 summarizes the classification performance on the

Table 2: Patch-based MRONJ candidate identification results on the test set.

	Precision	Recall	F1-score
VGG16	0.55	0.85	0.66
Inception	0.62	0.60	0.61
EfficientNet	0.62	0.55	0.58
MobileViT	0.49	0.67	0.57

test dataset. Among the four evaluated CNN models, VGG16 achieved the highest F1-score of 0.66. Therefore, the VGG16 is utilized in the subsequent pixel-level segmentation experiments.

2) Pixel-level MRONJ lesions detection for MRONJ candidate patches.

This section describes pixel-level MRONJ lesion detection using the U-Net. The U-Net was trained on patch images containing MRONJ pixels, along with their corresponding ground truth masks. To construct the training dataset, the manually annotated mouth region bbox was divided into 10 patches, and patches containing 5% or more MRONJ pixels were selected. The trained U-Net achieved an average DSC of 0.55 on the test set. This model is subsequently employed for fully automated MRONJ lesion detection.

In the fully automated pixel-level MRONJ lesion detection, the trained U-Net is applied only to patch images predicted as MRONJ by the trained VGG16. Then, the predicted pixel-level MRONJ lesion masks are stitched back to the original X-ray image to measure the DSC for each original image.

The predicted class probabilities are typically binarized using a threshold of 0.5 to determine class labels. However, in the training dataset used for the MRONJ patch candidate classifier (VGG16), the MRONJ lesions occupy only a small portion of each subdivided image patch. This class imbalance may lead to a skewed probability distribution. To address this, we evaluated pixel-level MRONJ segmentation performance across various classification thresholds ranging from 0.2 to 0.8. Figure 5 presents the average, standard deviation, and maximum DSC values for each threshold. When applying the default threshold of 0.5, a high number of false-negative pixels were observed, resulting in an average DSC of 0.27 and a maximum DSC of 0.70. In contrast, lowering the threshold to 0.45 reduced the false-negative pixels and achieved the best performance in this experiment, with an average DSC of 0.30 and a maximum DSC of 0.70.

Figure 6 visualizes the MRONJ lesion segmentation results obtained using the optimal configuration, which achieved the highest DSC of 0.70 at a threshold of 0.45. The visualization reveals minor under-segmentation (shown in yellow) and over-segmentation (shown in red), primarily around the lesion boundaries. Despite these discrepancies, the proposed method accurately localized the overall extent and position of

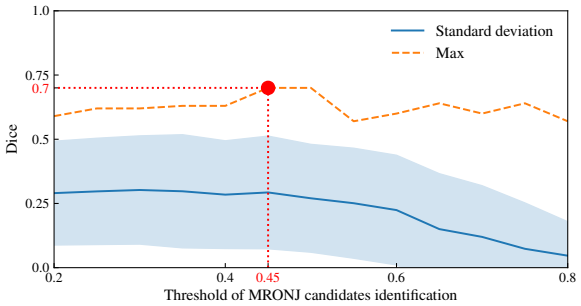


Fig. 5: DSC for MRONJ lesion segmentation across probability thresholds for patch candidate identification (CNN + U-Net).

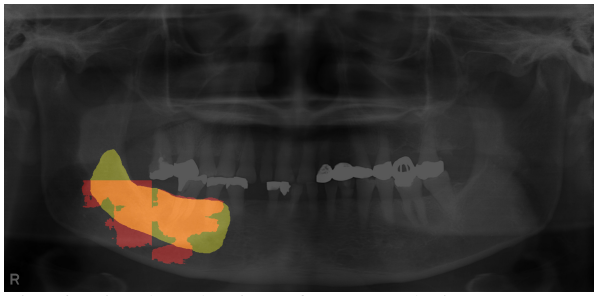


Fig. 6: Visual evaluation of MRONJ lesion segmentation result using CNN. Predicted MRONJ lesions are shown in red, ground truth MRONJ annotations are shown in yellow, and overlapping areas are displayed in orange.

the MRONJ lesions. Since the segmentation is only applied to patches classified as containing MRONJ pixels, misclassification at the patch level can propagate errors into the segmentation stage, leading to partial under- or over-segmentation. Enhancing patch-level classification performance is therefore essential for improving overall segmentation accuracy.

D. Results of MRONJ lesions prediction using masked ViT

1) MRONJ candidate region prediction using masked ViT

This section evaluates the performance of the Masked-ViT model in predicting the MRONJ candidate regions from randomly masked input X-ray images.

In this experiment, the Masked-ViT was trained using input images of 256×256 pixels with a batch size of 32. For each patient, 20 randomly masked images were generated to simulate variable masking patterns. In this study, an image is labeled as MRONJ-positive if the proportion of MRONJ lesion pixels within the visible (i.e., non-masked) area exceeds a predefined threshold. Since the masking rate directly influences the number of visible lesion pixels, it consequently affects the number of samples labeled as MRONJ-negative. To mitigate this imbalance, the threshold is adaptively set in the training set so that the numbers of MRONJ-positive and MRONJ-negative samples are equal. Table 3 summarizes the results of predicting whether the randomly masked input image contains

Table 3: Results of MRONJ candidate region prediction.

Masking rate (%)	Threshold (%)	Precision	Recall	F1-score
0.9	0.5	0.49	0.94	0.64
0.8	1.0	0.49	0.96	0.65
0.7	2.0	0.51	0.83	0.64
0.6	2.5	0.50	0.85	0.63

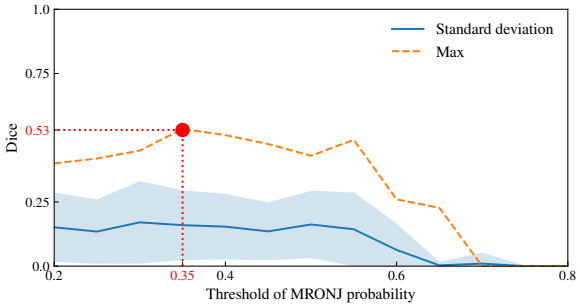


Fig. 7: DSC for MRONJ lesion segmentation across probability thresholds for MRONJ probability (Masked-ViT + U-Net)

MRONJ lesion pixels or not at various masking rates with the corresponding threshold. The highest F1-score of 0.65 was obtained at a masking rate of 0.8 and a threshold of 1.0%.

2) Automatic prediction results using masked-ViT

To perform pixel-level segmentation, the U-Net model is trained using 2×2 grid cells, that are occupied by 2% or more manually annotated MRONJ pixels. The trained U-Net model obtained an average DSC of 0.52 on the test set, which is used in the following fully automated segmentation experiments.

To determine the optimal inference-time threshold for identifying MRONJ candidate patches from the probability map, we evaluated the lesion segmentation performance using the best Masked-ViT model (as identified in the previous section) in conjunction with the trained U-Net. Figure 7 shows the DSC obtained at various probability thresholds. The highest DSC of 0.53 was obtained with the probability threshold of 0.35, indicating that this value yields the most effective trade-off between false positives and false negatives in candidate patch selection.

Figure 8 illustrates an example of the MRONJ segmentation result obtained using the proposed Masked-ViT and U-Net pipeline, with a corresponding DSC of 0.51. The relatively low DSC can be attributed to both under-segmentation (highlighted in yellow) and over-segmentation (highlighted in red) in the figure. One possible reason for this performance degradation is the insufficient number of masked samples generated during the inference stage, which led to incomplete suppression of false-positive grid cells in the MRONJ probability map. Consequently, patches without true MRONJ lesions were misclassified and further processed by U-Net, resulting in erroneous lesion segmentation in non-affected regions.

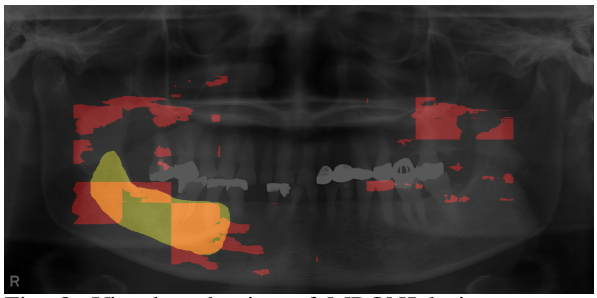


Fig. 8: Visual evaluation of MRONJ lesion segmentation result using Masked-ViT and U-Net. Predicted MRONJ lesions are shown in red, ground truth MRONJ annotations are shown in yellow, and overlapping areas are displayed in orange.

V. Conclusion

This study proposed a method for automatically predicting MRONJ lesions in panoramic X-ray images. Specifically, we introduced a Masked ViT to estimate the probability of MRONJ lesion presence across the image. This approach is compared with the baseline method, which identifies lesion-containing patches using a binary classification CNN. The detected candidate regions are subsequently refined using U-Net to perform pixel-level segmentation. Experimental results showed that the proposed CNN-based method achieved a maximum DSC of 0.7, outperforming the Masked-ViT-based method's DSC of 0.53. However, the average DSC of 0.3 indicated that the proposed method does not yet meet the performance required for clinical use, primarily due to limitations in lesion presence estimation accuracy during the binary classification step using Masked-ViT.

In the dataset, the number of non-MRONJ patches was approximately twice that of MRONJ patches, leading to a class imbalance that negatively impacted the training of the binary classification model and resulted in a biased recall. To mitigate this issue, the patch division process should be optimized to reduce the imbalance in class distribution. Furthermore, even in patches labeled as MRONJ, the actual lesion often occupies only a small portion of the patch, which introduces noise into the learning process. To address this, attention-based visualization techniques such as Score-CAM [17] can be employed to identify high-probability regions of MRONJ presence and refine candidate patch selection. Additionally, for improving the performance of the Masked-ViT, it is important to consider the number and consistency of generated masked variants during inference, as the current random masking scheme does not ensure uniform coverage across all patches.

References

[1] O. Nicolatou-Galitis, M. Schiødt, R. A. Mendes, C. Ripamonti, S. Hope, L. Drudge-Coates, D. Niepel and T. Van

den Wyngaert, "Medication-related osteonecrosis of the jaw: definition and best practice for prevention, diagnosis, and treatment," *Oral surgery, Oral medicine, Oral pathology and Oral radiology*, vol. 127, no. 2, pp. 117–135, 2019.

[2] M. Nashi, H. Kishimoto, M. Kobayashi, A. Tachibana, S. Hashitani, T. Shibatsuji, T. Nishida, K. Fujimura, S. Furudoi, Y. Ishida, S. Ishii, T. Fujita, S. Iwai, T. Shigeta, T. Harada, D. Miyai, D. Takeda, M. Akashi, K. Noguchi and T. Takenobu, "Incidence of antiresorptive agent-related osteonecrosis of the jaw: A multicenter retrospective epidemiological study in Hyogo Prefecture, Japan," *Journal of Dental Sciences*, vol. 18.3, pp. 1156–1163, 2023.

[3] A. Bedogni, S. Blandamura, Z. Lokmic, C. Palumbo, M. Ragazzo, F. Ferrari, A. Tregnaghi, F. Pietrogrande, O. Procopio, G. Saia, M. Ferretti, G. Bedogni, L. Chiarini, G. Ferronato, V. Ninfo, L. L. Russo, L. L. Muizo and P. F. Nocini, "Bisphosphonate-associated jawbone osteonecrosis: a correlation between imaging techniques and histopathology," *Oral surgery, Oral medicine, Oral pathology and Oral radiology, and Endodontology*, vol. 105, no. 3, pp. 358–364, 2008.

[4] N. Treister, N. Shehhy, E.H. Bae, B. Friedland, M. Lerman and S. Woo, "Dental panoramic radiographic evaluation in bisphosphonate-associated osteonecrosis of the jaws," *Oral Disease*, vol. 15, no. 1, pp. 88–92, 2009.

[5] P. Wongratwanich, K. Shimabukuro, M. Konishi, T. Nagasaki, M. Ohtsuka, Y. Suei, T. Nakamoto, R.G. Verdonschot, T. Kanesaki, P. Sutthiprapaporn and N. Kakimoto, "Do various imaging modalities provide potential early detection and diagnosis of medication-related osteonecrosis of the jaw? A review," *Dentomaxillofacial Radiology*, vol. 50, no. 6, 2021.

[6] Treister N. S., Friedland B. and Woo S. B., "Use of cone-beam computerized tomography for evaluation of bisphosphonate-associated osteonecrosis of the jaws," *Oral surgery, Oral medicine, Oral pathology and Oral radiology*, vol. 109, no. 5, pp. 753–764, 2010.

[7] J. Redmon, "You Only Look Once: Unified, Real-Time Object Detection," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[8] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *arXiv preprint arXiv*, 1409.1556, 2014.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[10] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *International conference on machine learning*, pp. 6105–6114, 2019.

[11] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv*, 2110.02178, 2021.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnilov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv*, 2010.11929, 2020.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollar and R. Girshick, "Masked autoencoders are scalable vision learners," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

[14] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv*, 2106.08254, 2021.

[15] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras and P. Prasanna, "Self pre-training with masked autoencoders for medical image classification and segmentation," *International Symposium on Biomedical Imaging (ISBI).IEEE*, pp. 1–6, 2023.

[16] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical image computing and computer-assisted intervention*, pp. 234–241, 2015.

[17] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel and X. Hu, "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020.