Oral Tissue Detection in Photographic Images using Deep Learning Technology

Ekawit Jaidee
Department of Computer Engineering
Faculty of Engineering
Chiang Mai University
Chiang Mai, Thailand, 50200
ekawit_jaidee@cmu.ac.th

Mansuang Wongsapai
Intercountry Centre for Oral Health
Department of Health
Ministry of Public Health
Chiang Mai, Thailand, 50000
mansuang.w@anamai.mail.go.th

Thawatchai Suthachai
Intercountry Centre for Oral Health
Department of Health
Ministry of Public Health
Chiang Mai, Thailand, 50000
thawatchai.s@anamai.mail.go.th

Sarit Theppitak
Department of Computer Engineering
Faculty of Engineering
Chiang Mai University
Chiang Mai, Thailand, 50200
sarit_theppitak@cmu.ac.th

Jitjiroj Ittichaicharoen
Department of Oral Biology and Diagnostic Sciences
Faculty of Dentistry
Chiang Mai University
Chiang Mai, Thailand, 50200
jitjiroj.itti@cmu.ac.th

Kritsasith Warin
Division of Oral and Maxillofacial Surgery
Faculty of Dentistry
Thammasat University
Pathum Thani, Thailand, 12121
warin@tu.ac.th

Siriwan Suebnukarn
Faculty of Dentistry
Thammasat University
Pathum Thani, Thailand, 12121
ssiriwan@tu.ac.th

Patiwet Wuttisarnwattana
Department of Computer Engineering
Faculty of Engineering
Chiang Mai University
Biomedical Engineering Institute
Chiang Mai University
Chiang Mai, Thailand, 50200
patiwet@eng.cmu.ac.th

***Abstract*—Oral tissue detection is an important task in many image-based computer aided analyses of oral healthcare. By detecting the shape of oral tissue in the photographic image, one can eliminate other non-essential parts, which can help optimize the further processes in the pipeline, for example, oral lesion detection, cancer classification, oral tissue segmentation, and dental health analyses. Manual labeling of data by humans can be inefficient, time-consuming, and error-prone, due to subjectivity, intra- and/or inter- observer variability between experts. Therefore, in this paper, we aimed to comprehensively evaluate deep learning-based models for detecting the oral tissue in various perspectives of photographic images. We studied four different state-of-the-art object detection models including Faster R-CNN, RetinaNet, SSD, and YOLOv5. Our models provided good results in the oral tissue detection application, with the Intersection over Union (IoU) over 90 % and F1-Score over 97%. We found that the Faster R-CNN with ResNet50 backbone and RetinaNet with ResNet50 backbone were the most accurate models in detecting objects. Alternatively, Faster R-CNN with MobileNetV3 backbone and SSD with VGG16 backbone were the best models in terms of processing speed, making them well-suited for handling large datasets. We hope that our findings here contribute to the improvement of the image-based oral healthcare analysis system in the future.**

Keywords— Compute Vision, Deep Learning, Dentistry, Image Processing, Oral Tissue Detection, Object Detection, Oral Cavity, Photographic image

I. INTRODUCTION

Oral disease is among the most prevalent diseases globall y. According to the World Health Organization (WHO), around 3.5 billion people, or 50% of the world's population, suffered from oral diseases in 2022 [1]. Common oral diseases include tooth decay, gingivitis, and oral cancer, which can be detected early through regular checkups. Failure to seek timely treatment for oral diseases can result in damage to the oral organs or even death. For instance, gingivitis can lead to tooth loss if left untreated in its advanced stages. Late treatment of oral diseases can also escalate treatment costs and the overall burden of disease management.

Deep learning technology has been widely used in clinics and research. Several studies have utilized the technology in the early detection of oral diseases, especially in photographic images of oral tissue. The images might include abnormalities such as oral potentially malignant disorders (OPMD) [2, 3] oral squamous cell carcinoma (OSCC) [4, 5], dental carries [6], and other lesions [7]. The technology can assist dental experts to analyze the images if they have any abnormality. This telemedical feature can help people who live in rural areas with limited access to oral healthcare.

However, in some studies, it was found that deep learning models had difficulties in searching for objects of interest such as oral lesions in the images. R. A. Welikala et al. developed an application called MeMoSa, which enabled dental professional s to analyze oral tissue images to see if there is any sign of the early stage of the oral cancers or not. The application incoporated artificial intelligence (AI) models for detecting and classifying the lesions. Although the model's performance was satisfactory for binary classification of oral lesions (any lesion or no?), but it fell short in detecting and multi-classifying different types of oral lesions [8]. G. Tanriver et al. [2] gave some opinions that the failure of the algorithm was largely due to the model training using the whole images instead of using smaller images or the regions of interest. To fix the problem, G. Tanriver cropped the areas of interest before training, which significantly improved the results.

979–8–3503–4210–9/23/$31.00 ©2023 IEEE

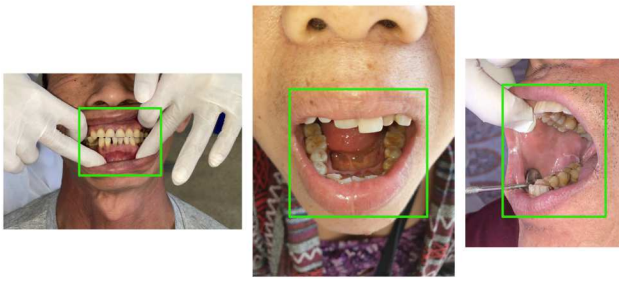

Fig. 1. Oral tissue images. The green box represents an area of interest for diagnosing oral disease. Any image contents outside the box is not an area of interest and should be disregarded for further processing.

Considering the issues observed in the previous studies, we anticipate similar challenges in our datasets. Dental experts in our research group routinely took photographic images of patients for oral health examination. They needed a system to assist their diagnosis in the oral health checkup or lesion analysis. However, the full images may not be suitable for further processing because they include irrelevant objects including patients' cloth, nose, doctor hands, dental equipment, etc. (Fig. 1). These objects are not of interest and irrelevant. If one leave them in the image for an AI system to consider during the oral analysis processing, they may cause errors in the results as reviewed previously.

The human effort to manually remove these objects in a large dataset of images is not efficient. Subjectivity, intraoperability, and interoperability significantly contribute to errors in the results created by human experts. Also, the manual labeling tasks are quite time-consuming, tedious, and expensive. Therefore, an automatic system for oral tissue region detection in photographic images is needed.

Mouth detection has been explored in past research, where most of these studies focus on detecting the mouth from a photograph that captures the entire face of the person, which is then analyzed further for various purposes such as classifying the person's emotions or detecting drowsiness while driving. An example of the studies is B. Reddy et al. [9], where a deep neural network-based model was proposed for detecting drowsiness in face images. The model consists of two steps: (1) face and landmarks detection, and (2) drowsiness classification. Most of the studies analyzed the entire person's face for determining emotions or behaviors. However, they did not focus on anatomical details of the oral cavity for the diagnostic purposes.

Therefore, in this paper, we propose a study to comprehensively evaluate the possibility of using deep learning models to detect an oral tissue region in photographic images for clinical purposes.

II. Oral Tissue Detection Models

Object detection is a computer vision task that uses deep learning models to identify and detect objects in photos or videos and is currently receiving a lot of attention. More than 3,000 papers were published on the topic in 2021. Object detection models can analyze an image to determine the location of a specific object and classify its type. These models are widely used in applications such as autonomous driving and traffic violation detection [10]. We will use object detection models to detect the shape of oral tissue in photographic images for clinical purposes.

The following sections describe in detail the models we used in the study. These models include Faster R-CNN, RetinaNet, SSD , and YOLO, which are commonly used in oral tissue analysis [2, 5, 6, 8] and other medical imaging applications [11, 12]. Again, these models have been published in more than 20,000 research papers.

A. Faster R-CNN

The Faster R-CNN algorithm is widely used in the field of computer vision for detecting objects. It was developed in 2015 by S. Ren et al. [13]. The architecture of the model has been developed based on previously developed R-CNN and Fast R-CNN models. There are three parts of the model that have important functions: a deep convolutional neural network (CNN) encoding part, Region Proposal Network (RPN) part, and Region-based Convolutional Neural Network (RCNN) part. The first part is the CNN model (sometimes called a backbone), which is responsible for generating feature maps encoding the spatial information at different scales of the input images. The second part is the RPN model, a crucial component used to propose regions likely to contain objects as bounding boxes. This allows the model to focus on a smaller set of candidate regions instead of exhaustively considering all possible image locations, significantly reducing computation time and making object detection more efficient. The last part is the RCNN model, designed to perform object detection in images. RCNN accurately localizes and classifies objects within an image by utilizing a combination of region proposals and deep convolutional neural networks. Due to its nature as a two-stage detection model with the additional RPN architecture, the model may exhibit superior performance as compared to other models, but that may cost with longer processing time [14]. However, the processing time usually depends on the size of the backbone.

B. RetinaNet

RetinaNet is a deep learning-based object detection model that was introduced by T.Y. Lin et al. in 2017 [15]. Detecting small objects in images can be a difficult task for traditional object detection models. In response, RetinaNet was designed to specifically address this problem. By leveraging Feature Pyramid Networks (FPN) and a new loss function known as Focal Loss, the RetinaNet model is able to significantly improve the accuracy of small object detection. This has made it a popular choice for various object detection applications in computer vision. The RetinaNet architecture consists of a backbone that extracts important image features and delivers semantic information at each layer to FPN. The FPN then extracts features from an input image at different scales. It applies a classification and regression subnetwork to each level of the feature pyramid to predict the presence of objects and their bounding boxes. Focal Loss tackles the issue of class imbalance in object detection, where the majority of image regions are non-object areas. It accomplishes this by assigning lower weights to correctly classified examples and prioritizing hard-to-classify examples during training. As a result, the model becomes more adept at handling challenging scenarios and exhibits improved detection performance, especially for small objects.

C. Single Shot MultiBox Detector (SSD)

SSD is a single-stage object detection model that was developed in 2016 by W. Liu et al. [16]. SSD is designed for real-time object detections such as self-driving vehicles, fire surveillance detection, oyster mushroom picking robot [17-19]. It is effective at detecting multiple objects in an image

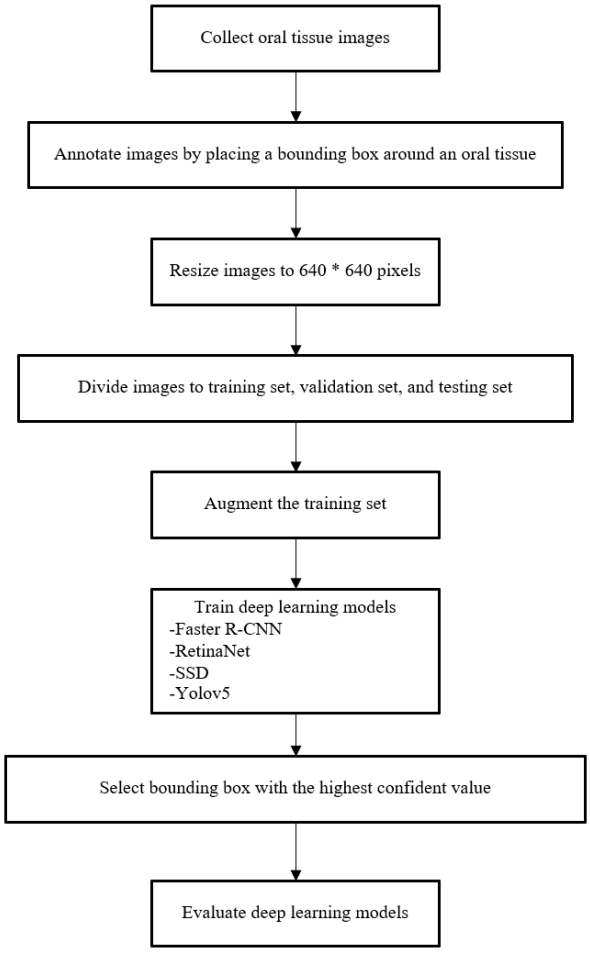

Fig. 2. Flowchart of the oral tissue detection experiment

in a single pass, making it faster than many other object detection algorithms [20]. Additionally, the implementation of multi-reference and multi-resolution detection techniques in SSD significantly improves the detection accuracy of a one-stage detector, especially for smaller objects [10]. The model has three main parts: Feature Extraction, Detection Head, and Non-Maximum Suppression. During the Feature Extraction stage, selective searches are conducted to identify object edges within the input image. Subsequently, the extracted features are mapped to the Detection Head for further processing. The Detection Head predicts bounding box coordinates, object classes, and confidence scores for each identified object. SSD uses a set of pre-defined default anchor boxes at different scales on multiple feature maps. These feature maps are derived from various layers in the network and allow the model to detect objects of different sizes effectively. The use of multi-scale feature maps enables SSD to handle objects of various scales in a single forward pass. The Non-Maximum Suppression algorithm then removes bounding boxes that are less likely to represent objects of interest based on the predicted scores.

D. You Only Look Once (YOLO)

YOLO is another well-known object detector model, especially for its processing speed. YOLO was originally developed by J. Redmon et al. [21]. but has been further developed into several versions by many groups [22-24]. It consists of three main components: the backbone network, the neck, and the head. The backbone network extracts important

Fig. 3. Selected bounding box with highest confident. blue = grouth truth

features from the input image to be used for detection. YOLOv5 introduces a new backbone network architecture called CSPDarknet53, which stands for Cross-Stage Partial Networks Darknet53. This novel backbone design improves the flow of information and enhances feature representation, contributing to better detection performance. The CSPDarknet53 backbone was inspired by previous works on Cross-Stage Aggregation in image classification tasks, making it more effective for object detection. The model neck extracts feature pyramids, helping the model in detecting objects of different sizes and scales effectively. The final operations of the model are carried out by the model head, which applies anchor boxes to feature maps and produces the ultimate output consisting of object classes, objectness scores, and bounding boxes. YOLO takes a distinct approach compared to two-stage detectors (such as Faster-RCNN) employing a single neural network across the full image. This network divides the image into regions and concurrently predicts bounding boxes and probabilities for each region. Despite its significant improvement in detection speed, YOLO experiences a decrease in accuracy when identifying locations, particularly when compared to two-stage detectors and especially for smaller objects [10].

III. EXPERIMENT

The aim of this work is to compare the performance of multiple detection models in locating a oral tissue in photographic images. The experimental steps are in the flowchart (Fig. 2).

A. Image Acquisition and Description

Dentists were the ones who collected the data. Patients were asked to sit on a chair. The dentists took images by capturing the patient oral tissue at 8 standard views. These include (1) Buccal Mucosa view, (2) Dorsal and Lateral Tongue view, (3) Floor of Mouth view, (4) Gingiva view, (5) Hard and Soft palate view, (6) Lips view, (7) Retromolar Pad view, and (8) Ventral Tongue view. There were 1,200 images in the dataset. The image data had a size ranging from 2,100 × 1,640 pixels to 4,032 × 3,024 pixels.

In this study, we used the secondary data that were collected in the previous works [4]. The data are kept at the Faculty of Dentistry, Thammasat University, and Intercountry Centre for Oral Health, Ministry of Public Health, Thailand. The study was endorsed by the ethics review committee of Thammasat University and was conducted in accordance with the principles of the Declaration of Helsinki. Since the images were fully anonymized and the study was retrospective, informed consent was not required.

B. Data Preparation

Ground truth was created by experts. All images were examined by the experts to identify the oral tissue location. Then, they placed a bounding box surrounding the organ, as

TABLE I. PERFORMANCE OF DEEP LEARNING MODELS

Model (Backbone)	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)	Parameter (M)	Average Processing Time (millisecond per image)
Faster R-CNN (MobileNetV3)	99.75 ± 0.32	99.75 ± 0.32	99.75 ± 0.32	91.19 ± 0.77	**19.40**	20.09
Faster R-CNN (ResNet50)	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	**93.53 ± 0.29**	41.80	132.33
RetinaNet (ResNet50)	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	93.40 ± 0.61	38.20	80.59
SSD (VGG16)	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	91.74 ± 0.82	35.60	**13.31**
YOLOv5 (CSPDarknet)	99.65 ± 0.18	95.67 ± 1.25	97.62 ± 0.72	90.73 ± 2.54	46.50	20.48
The numbers represent mean ± standard deviation across 4-fold cross-validation.						

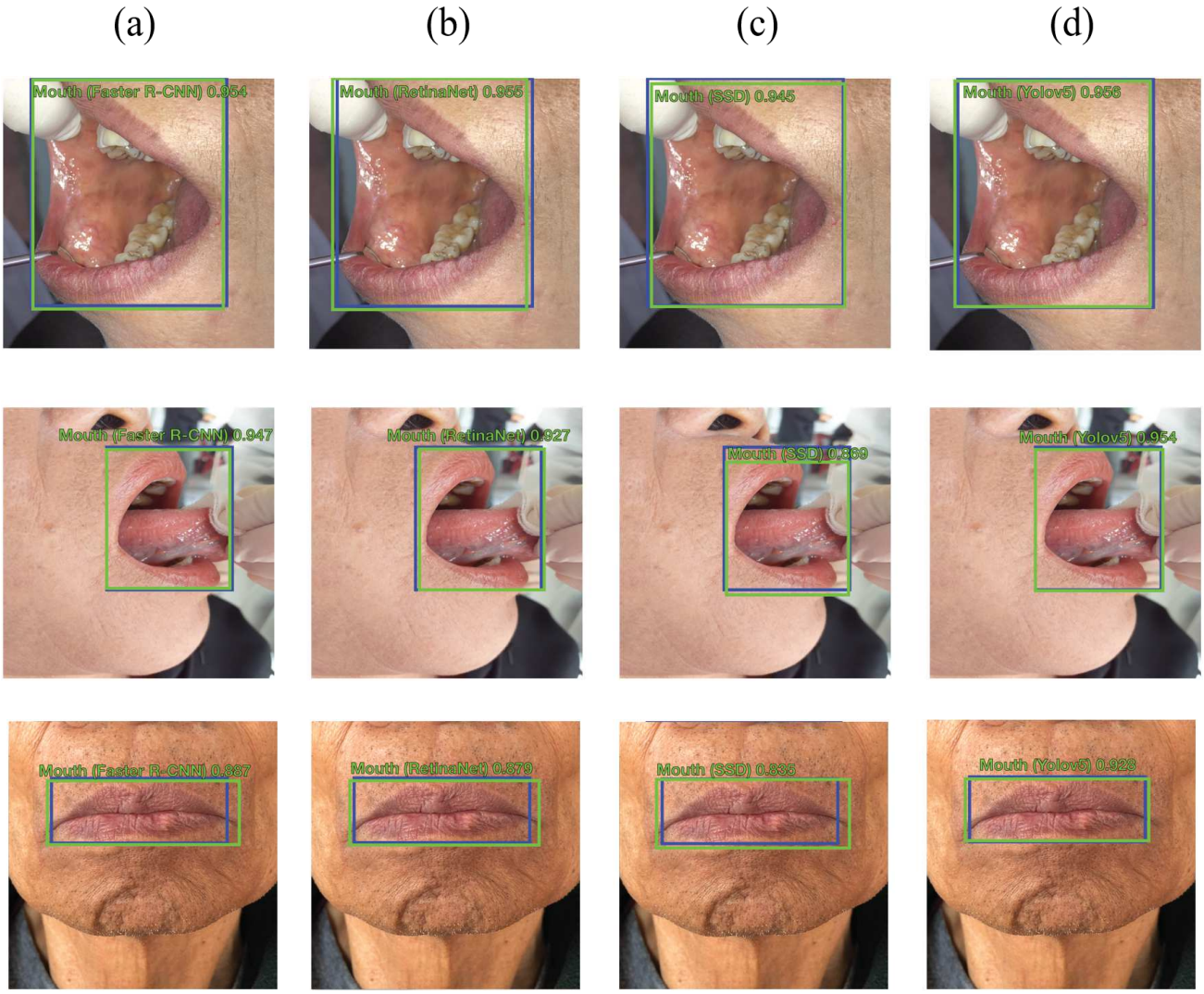

Fig. 4. Model detection results in three different anatomical views. In this example, buccal mucosa view (first row), dorsal and lateral tongue view (middle row), and lip view (last row) were tested. The images in columns (a-d) show the true positive results in different models: (a) Faster R-CNN, (b) RetinaNet, (c) SSD, and (d) Yolov5. The green box is prediction bounding box and the blue box is ground truth bounding box.

shown in Fig. 1. The tool used in this step was a web application called Labelbox.com [25]. All images including ground truth were resized to 640 × 640 pixels.

To determine the robustness of our training method, we performed 4-fold cross-validation and measured performance variability (standard deviation). Our dataset was divided into four subsets, with each subset being equally divided into training and test data. We performed cross-validation using the stratified approach to ensure that the ratio of anatomical views was maintained during the splitting process.

The dataset was divided into three sets: a training set, a validation set, and a testing set, consisting of 750 images (62.5%), 150 images (12.5%), and 300 images (25%) respectively. The training set was used to train the detection model. The validation set was used as a stopping criterion of the model to prevent the model overfit. The testing set was used to evaluate the model's performance.

C. Deep Learning Models

We used four famous object detection models in this study. These are state-of-the-art deep learning models that have been extensively used in many biomedical applications [13, 15, 16, 22]. The models were Faster R-CNN, RetinaNet, SSD, and YOLOv5.

We downloaded pretrained Faster R-CNN, RetinaNet, and SSD models from the PyTorch library for our experiments [26]. For the Faster R-CNN experiments, we used two types of pretrained backbones (encoding modules): ResNet50 and MobileNetV3. For RetinaNet, we used ResNet50 as the backbone. For SSD, we used VGG16 as the backbone. These backbones were pretrained using 330 thousands of images in the COCO val2017 public dataset [27]. Yolov5 model (Yolov5l) was downloaded from the Ultralytics official website (https://github.com/ultralytics/yolov5). The Yolov5 backbone architecture was CSPDarknet53, which again pretrained with the COCO dataset.

D. Training setup

To train deep learning models to detect a oral tissue in the images, we used the stochastic gradient descent optimizer with the following hyper-parameters: Number of maximum epoch was 1000, batch size of 16, learning rate of 0.01, momentum rate of 0.9, and weight decay of 0.0005. Except for the SSD model, the learning rate is 0.0001.

During training, the model performed inter-train augmentation by applying the following image adjustments: a horizontal flip with a 0.5 percent chance, a random vertical flip with a 0.5 percent chance, scaling with a size range of (-0.2, 0.2) gain, and shifting the Hue Saturation Value using the parameters (0.0015, 0.4, 0.2), respectively. In this section, we optimize the image data using the Python library Albumentations version 1.3.0 and OpenCV version 4.7.0.68.

We use early stopping to check if our model is learning optimally. During the training phase, we use a validation set to test the model's performance in each epoch. If the validation set loss does not decrease for 50 epochs, we stop training the model because it is assumed that our model has already learned the best it can.

E. Post Processing

A single image contains only one oral tissue. When utilizing the detection model, multiple bounding boxes may be identified. However, our images contained only one oral tissue per image. Therefore, we perform a post-processing step by selecting the most confident bounding box given by the model (Fig.3).

F. Performance Evaluation

To measure the performance of the models, we used the testing set that was prepared previously. Usually, an object detection model is evaluated using four standard key metrics: intersection over union (IoU), precision, recall, and F1-score.

the Intersection over Union is defined using the following equation:

$$IoU = \frac{A \cap B}{A \cup B} \quad (1)$$

where A is prediction bounding box and B is ground truth bounding box. The number ranges from 0.0 – 1.0 where 1.0 indicates the perfect match between the bounding boxes and 0.0 indicates absolutely no matching between the two. In this study, we define that if the bounding boxes have an IoU below 0.5, it also indicates no matching.

For other metrics, we need to measure the following quantities:

- True Positive (TP): the number of bounding boxes in the prediction that match the corresponding bounding boxes in the ground truth with IoU greater than or equal to 0.5.
- False Positive (FP): the number of bounding boxes in the prediction that match the corresponding bounding boxes in the ground truth but with IoU less than 0.5 (or do not match).
- False Negative (FN): the number of bounding boxes in the ground truth that match the corresponding bounding boxes in the prediction but with IoU less than 0.5 (or do not match).

Thus, Precision, Recall, and F1-Score equations are:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\text{-}Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

G. Workstation and Development Platform

The model was implemented/trained/run on a workstation named Deep Server, configured as follows: CPU Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz, NVIDIA Quadro RTX 8000 48Gb × 2 cards, CUDA version 12.0, with Ubuntu 20.04.5 LTS operating system. The development platform was Python version 3.8.11, and PyTorch ver. 1.12.1.

IV. Result And Discussion

The results show that oral tissue can effectively be recogni zed in the photographic images in all models. We evaluated four state-of-the-art object detection models: Faster R-CNN, RetinaNet, SSD, and YOLOv5 to perform the task. The qualitative results are shown in Fig. 4 and the quantitative results are shown Table 1.

We found that Faster R-CNN with the pretrained ResNet50 backbone yielded the highest IoU value of 93.53%, indicating that the model has predicted bounding boxes with the highest degree of overlap with the ground truth bounding boxes. It also achieved quite an ideal value F1-score of 100.00%, which suggests that the model is performing well in terms of detecting positive instances. Regardless of the good performance, it comes at the cost of prediction time, which seems to be slower than other models (10 times slower than the best one). We believe this was due to a high number of parameters (41.80 million). Interestingly, by changing the backbone to a smaller model, MobileNetV3, the prediction time was improved by 6 folds. However, the performance slightly decreased to 91.19% and 99.75% in terms of IoU and F1-score respectively. Notice

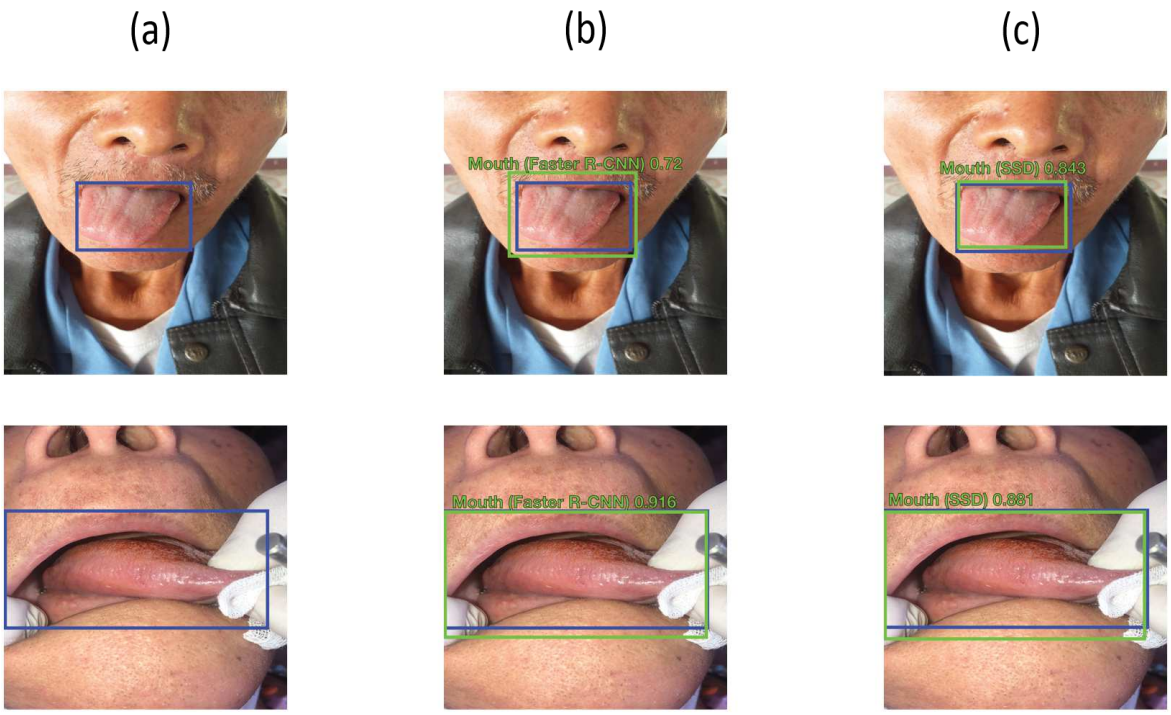

Fig. 5. YOLOv5 model tended to miss oral tissue in our dataset whiles other models did not. False negatives of the YOLOv5 (column (a)) are shown in two examples (top row and bottom row) where other model such as Faster R-CNN with MobileNetV3 (column (b)) and SSD with VGG16 (column (d)) provided a relatively good detections.

that the number of trainable parameters decreased from 41.80 million to 19.40 million for ResNet50 and MobileNetV3, respectively.

Both RetinaNet model (with ResNet50 backbone) and SSD model (with VGG16 backbone) performed well in terms of the detection performance (Table 1). The SSD model seems to perform best in terms of processing time, regardless of the relatively high number of parameters. The RetinaNet, SSD, and Faster R-CNN+ResNet50 models delivered a perfect score of F1-score. The reasons might be that all test images contained a well-defined oral tissue shape, reasonable zoom and orientation, minimal blur, and with sufficient light. Importantly, we considered that any overlap between prediction mask and the ground truth above 50% were treated as true positive.

Although the YOLOv5 model performed the oral tissue recognition task with an impressive processing time (20.48 ms per image), it did not perform relatively well under other metrics. The measured recall and IoU score were 95.67% and 90.73%, respectively. The failures of YOLOv5 are also illustrated in Fig. 5.

V. Conclusion

Recently, oral tissue image analysis using deep learning technology has been widely used in oral health care applications. However, most oral tissue images contain parts that are not relevant to the diagnosis, such patients' cloth, nose, doctor hands, dental equipment, which are outside the mouth. By leaving these components in the images, further processing may fail to recognize the potential lesions in the oral cavity. To address this issue, our model will help to automatically reduce the region of interest to focus only on the oral tissue area from the whole oral tissue image, as shown in Figure 1.

We evaluated four state-of-the-art deep learning models to find a cavity area in the image. We observe that in terms of detection performance, Faster R-CNN+ResNet50 and RetinaNet+ResNet50 models outperformed other models. In terms of the prediction speed, we discover that Faster R-CNN+MobileNetv3 and SSD+VGG16 model were the dominated ones. Regarding the YOLOv5, we think that the model may not be suitable for our application as they missed oral tissue in many instances whereas other models could do (Fig. 5). Also, we see that the YOLOv5 training produced an unstable model as observed in a relatively high standard deviation (Table 1).

In conclusion, this research presented a comprehensive study utilizing deep learning technology to recognize an oral tissue in a photographic image for the first time. With this knowledge, one can automate the oral tissue cropping process without relying on laborious human efforts. We believe that the subsequent oral healthcare analysis such as oral lesion detection, oral disease classification, tooth health evaluation, and others, will surely benefit from our work.

Acknowledgment

This research project was supported by the Graduate School (Chiang Mai University), the Department of Computer Engineering (Chiang Mai University), the Faculty of Engineering (Chiang Mai University), and Intercountry Centre for Oral Health (the Department of Health, the Thailand Ministry of Public Health). We also thank the following organizations for providing the oral tissue datasets: (1) Intercountry Center for Oral Health (Department of Health, Ministry of Public Health of Thailand), (2) the Faculty of Dentistry (Chiang Mai University), and (3) the Faculty of Dentistry (Thammasat University). Additionally, we would like to give thanks to Prof. Kasemsit Teeyapan, Ph.D for providing the high-performance workstation to be used in this study, and Mr. Kan Tippayamontri for his technical advices.

REFERENCES

[1] N. Jain, U. Dutt, I. Radenkov, and S. Jain, "WHO's global oral health status report 2022: Actions, discussion and implementation," (in eng), Oral Dis, Jan 20 2023, doi: 10.1111/odi.14516.

[2] G. Tanriver, M. Soluk Tekkesin, and O. Ergen, "Automated Detection and Classification of Oral Lesions Using Deep Learning to Detect Oral Potentially Malignant Disorders," (in eng), Cancers (Basel), vol. 13, no. 11, Jun 2 2021, doi: 10.3390/cancers13112766.

[3] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, and P. Jantana, "Performance of deep convolutional neural network for classification and detection of oral potentially malignant disorders in photographic images," International Journal of Oral and Maxillofacial Surgery, vol. 51, no. 5, pp. 699-704, 2022.

[4] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, and P. Jantana, "Automatic classification and detection of oral cancer in photographic images using deep learning algorithms," (in eng), J Oral Pathol Med, vol. 50, no. 9, pp. 911-918, Oct 2021, doi: 10.1111/jop.13227.

[5] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, P. Jantana, and S. Vicharueang, "AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer," PLOS ONE, vol. 17, no. 8, p. e0273508, 2022, doi: 10.1371/journal.pone.0273508.

[6] E. Y. Park, H. Cho, S. Kang, S. Jeong, and E.-K. Kim, "Caries detection with tooth surface segmentation on intraoral photographic images using deep learning," BMC Oral Health, vol. 22, no. 1, p. 573, 2022/12/07 2022, doi: 10.1186/s12903-022-02589-1.

[7] Y. Liang et al., "OralCam: Enabling Self-Examination and Awareness of Oral Health Using a Smartphone Camera," presented at the Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 2020. [Online]. Available: https://doi.org/10.1145/3313831.3376238.

[8] R. A. Welikala et al., "Automated Detection and Classification of Oral Lesions Using Deep Learning for Early Detection of Oral Cancer," IEEE Access, vol. 8, pp. 132677-132693, 2020, doi: 10.1109/ACCESS.2020.3010180.

[9] B. Reddy, Y.-H. Kim, S. Yun, C. Seo, and J. Jang, "Real-time driver drowsiness detection for embedded system using model compression of deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 121-128.

[10] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," Proceedings of the IEEE, 2023.

[11] R. Yang and Y. Yu, "Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis," Frontiers in oncology, vol. 11, p. 638182, 2021.

[12] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, and Z. Zeng, "CLU-CNNs: Object detection for medical images," Neurocomputing, vol. 350, pp. 53-59, 2019.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.

[14] A. K. Shetty, I. Saha, R. M. Sanghvi, S. A. Save, and Y. J. Patel, "A review: Object detection models," in 2021 6th International Conference for Convergence in Technology (I2CT), 2021: IEEE, pp. 1-8.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980-2988.

[16] W. Liu et al., "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 2016: Springer, pp. 21-37.

[17] N. N. A. Mangshor, N. S. Saharuddin, S. Ibrahim, A. F. A. Fadzil, and K. A. F. A. Samah, "A Real-Time Speed Limit Sign Recognition System for Autonomous Vehicle Using SSD Algorithm," in 2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), 2021: IEEE, pp. 126-130.

[18] A. Nguyen, H. Nguyen, V. Tran, H. X. Pham, and J. Pestana, "A visual real-time fire detection using single shot multibox detector for uav-based fire surveillance," in 2020 IEEE Eighth International Conference on Communications and Electronics (ICCE), 2021: IEEE, pp. 338-343.

[19] Y. Qian, R. Jiacheng, W. Pengbo, Y. Zhan, and G. Changxing, "Real-time detection and localization using SSD method for oyster mushroom picking robot," in 2020 IEEE International Conference on Real-time Computing and Robotics (RCAR), 2020: IEEE, pp. 158-163.

[20] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," Progress in Artificial Intelligence, vol. 9, no. 2, pp. 85-112, 2020.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[22] YOLOv5 by Ultralytics. (2020). [Online]. Available: https://github.com/ultralytics/yolov5

[23] C. Li et al., "YOLOv6: A single-stage object detection framework for industrial applications," arXiv preprint arXiv:2209.02976, 2022.

[24] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv preprint arXiv:2207.02696, 2022.

[25] Labelbox. (2023). Online, Online. [Online]. Available: https://labelbox.com

[26] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," Advances in neural information processing systems, vol. 32, 2019.

[27] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, 2014: Springer, pp. 740-755.