Journal of Dental Sciences 19 (2024) 254–260

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Deep learning algorithms for classification and detection of recurrent aphthous ulcerations using oral clinical photographic images

Mimi Zhou, Weiping Jie, Fan Tang, Shangjun Zhang, Qinghua Mao, Chuanxia Liu, Yilong Hao*

Stomatology Hospital, School of Stomatology, Zhejiang University School of Medicine, Zhejiang Provincial Clinical Research Center for Oral Diseases, Key Laboratory of Oral Biomedical Research of Zhejiang Province, Cancer Center of Zhejiang University, Hangzhou, China

Received 9 April 2023; Final revision received 19 April 2023
Available online 2 May 2023

KEYWORDS
Recurrent aphthous ulcerations;
Diagnosis;
Oral;
Deep learning;
Artificial intelligence

Abstract *Background/purpose:* The application of artificial intelligence diagnosis based on deep learning in the medical field has been widely accepted. We aimed to evaluate convolutional neural networks (CNNs) for automated classification and detection of recurrent aphthous ulcerations (RAU), normal oral mucosa, and other common oral mucosal diseases in clinical oral photographs.

Materials and methods: The study included 785 clinical oral photographs, which was divided into 251 images of RAU, 271 images of the normal oral mucosa, and 263 images of other common oral mucosal diseases. Four and three CNN models were used for the classification and detection tasks, respectively. 628 images were randomly selected as training data. In addition, 78 and 79 images were assigned as validating and testing data. Main outcome measures included precision, recall, F1, specificity, sensitivity and area under the receiver operating characteristics curve (AUC).

Results: In the classification task, the Pretrained ResNet50 model had the best performance with a precision of 92.86%, a recall of 91.84%, an F1 score of 92.24%, a specificity of 96.41%, a sensitivity of 91.84% and an AUC of 98.95%. In the detection task, the Pretrained YOLOV5 model had the best performance with a precision of 98.70%, a recall of 79.51%, an F1 score of 88.07% and an AUC of Precision-Recall curve 90.89%.

Conclusion: The Pretrained ResNet50 and the Pretrained YOLOV5 algorithms were shown to have superior performance and acceptable potential in the classification and detection of RAU lesions based on non-invasive oral images, which may prove useful in clinical practice.

* Corresponding author. Stomatology Hospital, Zhejiang University School of Medicine, 166 Qiutao North Road, Hangzhou, 310000, Zhejiang, China.
E-mail address: 7519040@zju.edu.cn (Y. Hao).

https://doi.org/10.1016/j.jds.2023.04.022

Introduction

Recurrent aphthous ulcerations (RAU) is the most common benign ulcerated lesion in the oral cavity with an uncertain etiology.[1] Surveys have suggested that at least 20% of the population suffers from this disease, and the prevalence of RAU can be as high as 50% in certain populations.[2,3] RAU is characterized by single or multiple painful and recurrent ulcerations in the oral cavity.[4] Patients experience discomfort when swallowing, eating, and communicating, thereby significantly affecting their quality of life.[5] It is worth noting that oral ulceration is quite complicated and diverse due to overlap of their clinical features, such as cancerous ulcers, tuberculosis ulcers, and eosinophilic ulcers.[6] An accurate diagnosis of RAU can significantly reduce unnecessary psychological and economic burdens on patients and greatly improve clinical diagnosis and treatment efficiency.

In recent years, artificial intelligence, particularly deep learning technology based on convolutional neural network (CNN), has enabled computer-aided diagnosis due to its excellent automatic feature extraction and generalization abilities, and its potential in the medical field has been increasingly recognized.[7,8] Studies have demonstrated that deep learning technologies, such as image recognition, have been extensively employed in intelligent diagnosis of skin diseases, caries detection, and early detection of lung and gastric cancers.[9–12] Oral mucosal diseases, such as RAU, which can be easily visualized without special instruments, possess significant potential for the application of artificial intelligence technology in the intelligent diagnosis of oral mucosal diseases.[13]

This study developed and trained four deep learning models to achieve a classification task based on clinical oral photographs, which can intelligently differentiate RAU, normal oral mucosa, and other common oral mucosal diseases. Furthermore, we constructed and trained three additional deep learning models to achieve intelligent recognition of lesion locations in RAU images and evaluated their performance. The implementation of this system is expected to eliminate variability in the diagnosis of RAU between clinical doctors and to benefit patients in disease follow-up and self-assessment. This study also establishes a foundation for the future development of more precise and convenient intelligent diagnosis systems for oral mucosal diseases.

Materials and methods

Data acquisition

This study was approved by the Institutional Review Board of the Stomatology Hospital of Zhejiang University School of Medicine.

The dataset of images utilized in this study was retrospectively collected from patients of the department of oral medicine between March 2022 and March 2023. A total of 785 images were included and divided into three distinct groups, comprising 251 images of RAU patients who met the diagnostic criteria for RAU, 271 normal oral mucosa images from healthy volunteers, and 263 images of patients with other common oral mucosal diseases.[3] The collection process was assisted by professional nurses using a professional camera to capture high-quality images of various areas of the oral cavity for each group. A detailed presentation of these areas is available in Table 1.

Two senior doctors manually annotated all images using the LabelImg software to label the image types and boundaries of the lesions in the RAU images.[14] To ensure the accuracy of the labeling results, the maximum intersection area of the labeling results of the two doctors was used to determine the boundaries of the final labeling results. The specific number of annotations is presented in Table 1, comprising 184 images with a single lesion and 67 images with multiple lesions. Examples of the oral photographic images from the collected dataset are presented in Fig. 1.

Experiments

The objective of this study was twofold: to perform an image classification task and an object detection task. The dataset of 785 photos was randomly divided into three distinct sets-training, validation, and testing-in an 8:1:1 ratio. The training set consisted of 628 photos to facilitate model learning, while the validation set comprising 78 photos was used to prevent overfitting. Finally, the testing set contained 79 photos and was utilized to evaluate the experimental results. To avoid anomalies caused by randomness, all experiments were conducted using 5-fold cross-validation. The machine model and program versions used for the classification and detection tasks were NVIDIA 3060(NVIDIA Corporation, Santa Clara, CA, USA), Pytorch 1.10 (Meta Platforms, Inc., Menlo Park, CA, USA), and CUDA 11.4 (NVIDIA Corporation) respectively. A detailed illustration of the entire research process is presented in Fig. 2.

Image classification

To train the classification model that automatically determines whether a picture belongs to RAU, normal oral mucosa, or other common oral mucosal diseases, three CNN-based models widely used in this field, namely DenseNet121, ResNet18, and ResNet50, were selected.[15,16] These models can introduce model parameters based on pretraining on big data through transfer learning. Additionally, to verify the effect of pretraining on the Imagenet dataset, a Not Pretrained ResNet50 was included as a comparison for the ResNet50 model.

Table 1 Baseline characteristics of our used photographic images of oral lesions.

Characteristics	RAU Data Set	Normal oral mucosa Data Set	Other common oral mucosal diseases
No. of images	251	271	263
No. of images with a single ulcer	184		
No. of images with multiple ulcers	67		
No. of oral ulcer lesions	577		
Distribution			
Lip mucosa	90	95	87
Buccal	44	48	55
Tongue	67	70	60
Palatal mucosa	11	14	12
Mouth floor	15	20	21
Retromolar trigone	10	10	10
Alveolar ridge	14	14	18
Types of other disease			
Oral mucosal patches striae diseases			148
Infectious diseases of oral mucosa			30
Bullous diseases of oral mucosa			15
Oral hypersensitivity disorders			20
Traumatic lesions of the oral mucosa			20
Carcinoma in situ or tumor-like lesions			30

Summary of image characteristics and available demographic information in the development and clinical validation data sets; RAU, Recurrent aphthous ulcerations; NO., number.

Figure 1 Examples of the oral photographic images from the collected dataset. (A) Normal oral mucosa images on the palatal mucosa. (B) A single ulcer lesion on the lip mucosa. (C) Multiple ulcer lesions on the tongue.

During the training process, the classification model automatically extracts the key features of each image and predicts the probability of each class using the Softmax function. The output results were compared with the doctor's annotated results, and the specific loss function (cross-entropy) was used to guide the model towards the next learning direction. To improve the training effect, the TorchVision (Meta Platforms, Inc.) program was used to enhance the annotated data, including rotating, cropping, scaling, and adjusting the saturation of images. This increases the generalization ability of the data used for training, which, in turn, helped to improve the model's performance. The training hyperparameters included a maximum of 25 epochs and a batch size of 16.

Objective detection

The experiment involved constructing an object detection model capable of accurately identifying the boundaries of ulcers in an oral image. Two widely used models in this field, namely You Only Look Once Version 5 (YOLOV5) and Faster Region-based Convolutional Neural Network (Faster R—CNN), were selected.[17,18] Both models used pretrained weights based on the COCO dataset and were further optimized through secondary training. To verify the effect of transfer learning, a Not Pretrained YOLOV5 was included as a control.

While the features extracted by CNN are used for classification, they also need to predict the possible target range and continuously compare the output results with the annotated results using regression for positioning. Therefore, the model's loss function includes both classification loss and localization loss. When the model predicts incorrectly, relevant feedback information is passed back through gradients to help correct the parameters in a more optimal direction. The key hyperparameters used for training were a maximum of 100 epochs and a batch size of 8. The Gradient-weighted Class Activation Mapping (Grad-CAM) method was also employed to draw Saliency Maps for YOLOV5's internal parameter results.[19]

Evaluation measures

For the classification task, we used several objective metrics including precision, recall (equal to sensitivity), F1

score, specificity, confusion matrix, and area under the receiver operating characteristics curve (AUC) to evaluate the performance of the classification model.[20] For the object detection task, we employed precision, recall, F1 score, and the area under the Precision-Recall (PR) curve AUC to evaluate the performance of the selected model.[21]

Results

Image classification

The performance of the image three-classification model on the test set is presented in Table 2. Among all the models, the Pretrained ResNet50 exhibited the best performance, achieving an precision of 92.86%, a recall (sensitivity) of 91.84%, an F1 score of 92.24%, a specificity score of 96.41%, and an AUC of 98.95%. In contrast, the Not Pretrained ResNet50 displayed the lowest performance in all indicators, achieving only an AUC of 86.69%. Fig. 3 illustrates the ROC curves and confusion matrices corresponding to four CNN models, as well as the changes in specificity scores for each model during the training process.

Objective detection

The performance of object detection on the test set is summarized in Table 3. Among the three CNN models, the Pretrained YOLOV5 delivered the best performance, achieving a precision of 98.70%, a recall of 79.51%, an F1 score of 88.07%, and an AUC of the PR curve of 90.89%. The Not Pretrained YOLOV5 model exhibited the least validation accuracy. Fig. 4 shows the AUC of PR Curve of each model. Two examples of the output of the object detection model are presented in Fig. 5.

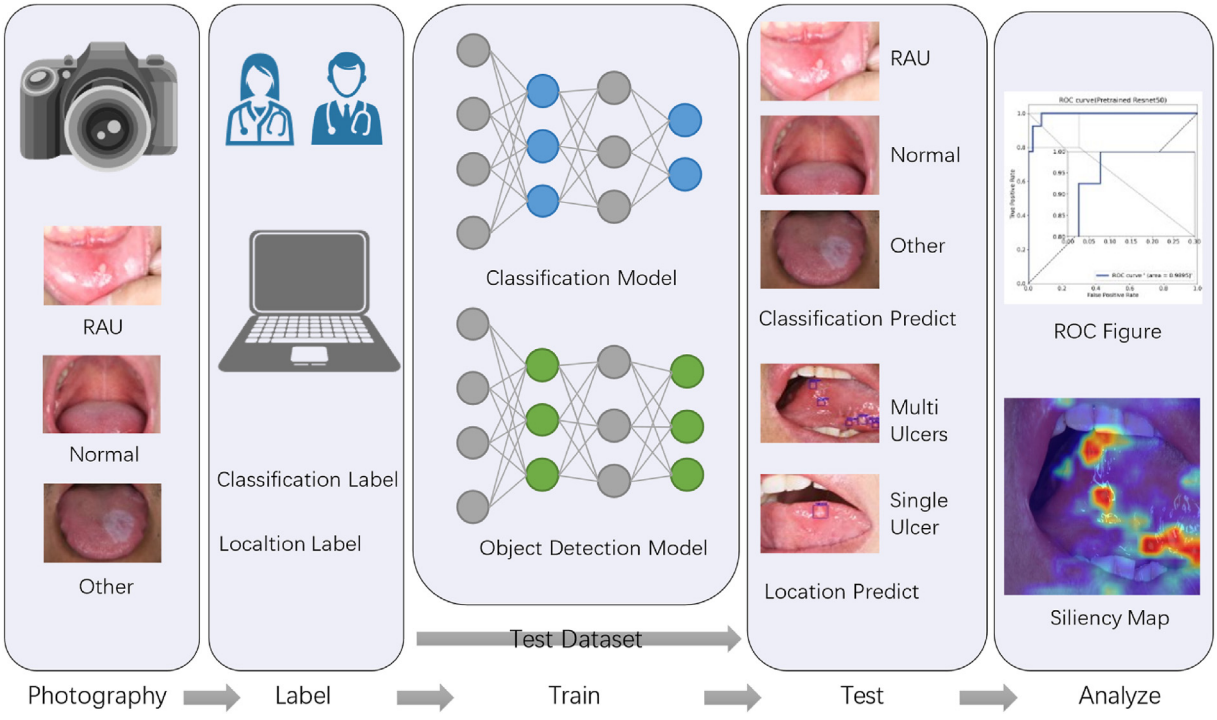

Figure 2 Workflow for computer-aided diagnosis of RAU based on deep learning: a certain amount of oral cavity images were collected through a camera, which were then annotated by senior doctors to classify samples and mark the boundaries of the lesions in the RAU images. The training and validation sets were used to train the deep learning model. Upon completion of the training, the testing set was used to predict and validate model performance. Finally, the results were analyzed quantitatively and visually to evaluate the accuracy of the model's predictions.

Table 2 Performance of the adopted models in the classification task.

Model	Precision (%)	Recall/Sensitivity (%)	F1 (%)	Specificity (%)	AUC
Pretrained Densent121	88.91	88.24	88.35	94.51	97.82
Pretrained Resnet18	88.68	88.19	88.30	94.73	98.33
Pretrained Resnet50	92.86	91.84	92.24	96.41	98.95
Not Pretrained Resnet50	69.07	68.10	67.85	85.44	86.69

AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic.

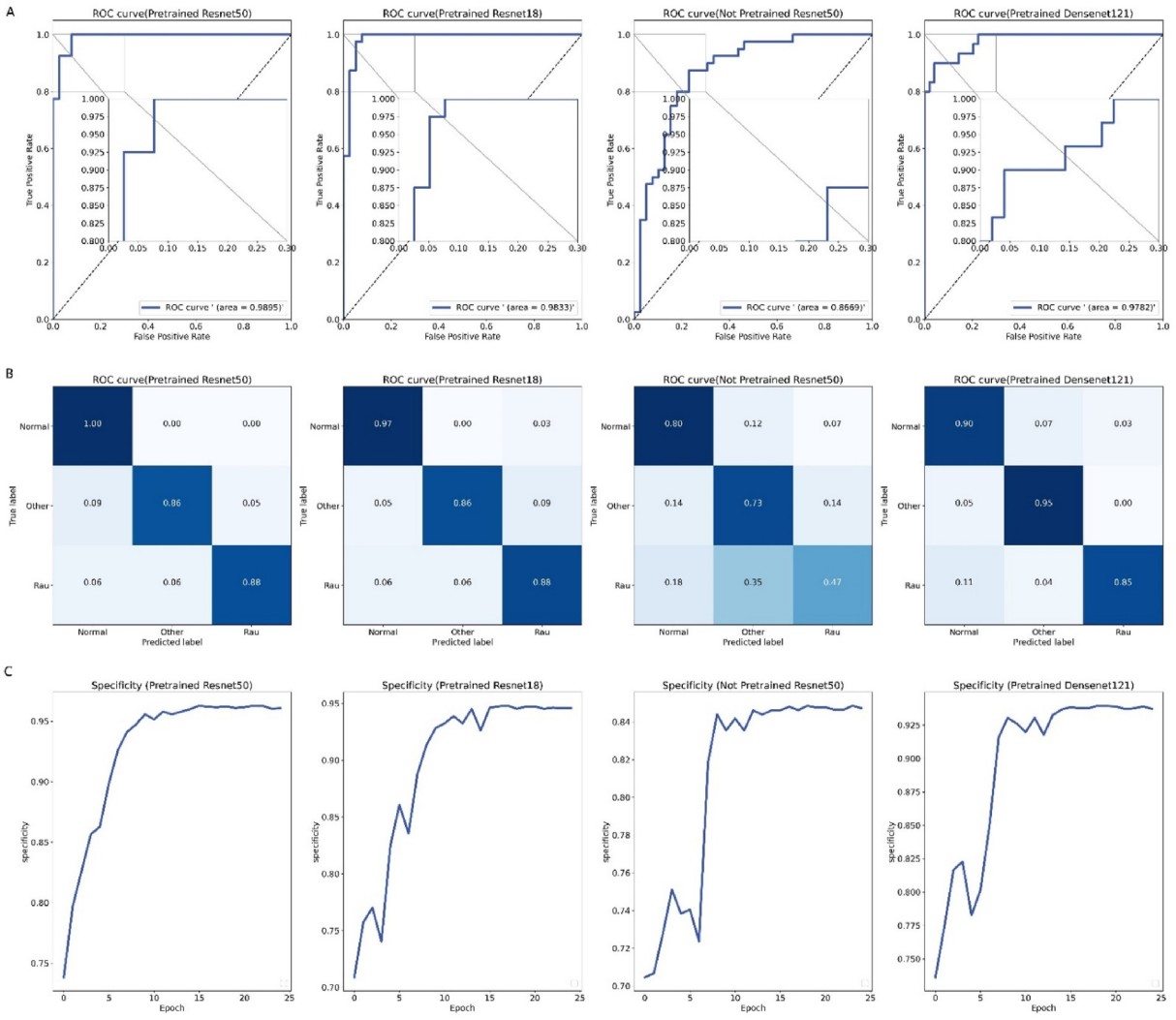

Figure 3 (A) ROC curves by each classification model, the upper-left corner of each image was magnified. (B) The confusion matrix obtained by each model on the test dataset. (C) The changes in specificity metric of the selected model on the validation set during the training process.

Table 3 Performance of the adopted models in the objective detection.

Model	Precision (%)	Recall (%)	F1 (%)	AUC of PR Curve
Pretrained Faster R—CNN	87.11	84.12	85.59	90.76
Pretrained YOLOV5	98.70	79.51	88.07	90.89
Not Pretrained YOLOV5	65.83	75.01	70.12	72.35

AUC: area under the receiver operating characteristic curve; PR: Precision-Recall; Faster R—CNN: Faster Region-based Convolutional Neural Network; YOLOV5: You Only Look Once Version 5.

Discussion

In clinical practice, the differential diagnosis of RAU from other ulcerative diseases, such as cancerous ulcers (carcinoma in situ or oral squamous cell carcinoma), traumatic ulcers, tuberculosis ulcers, etc., poses a considerable challenge due to their complexity and diversity.[6] The accurate identification and diagnosis of oral ulcerative disorders can alleviate the psychological and economic burden of patients, as well as improve their overall prognosis and survival rates. Deep learning algorithms based on CNN possess powerful data classification and prediction capabilities and have been widely applied in the field of medicine, such as intelligent-assisted diagnosis of liver and esophageal cancers and automatic detection of ophthalmic diseases.[22—25] These impressive findings inspire us to believe that deep learning may also have the potential to capture the characteristic features of oral mucosal

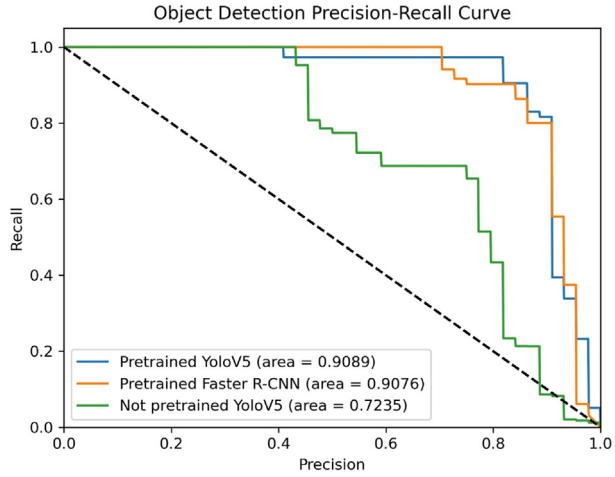

Figure 4 The Precision-Recall curves of the three selected models for the object detection task.

diseases, represented by RAU, providing a promising solution for designing intelligent diagnostic models of oral mucosal diseases.

In this study, we employed CNN-based classification and detection models to detect RAU in photographic images, resulting in a high degree of diagnostic accuracy. The Pretrained ResNet50 model exhibited the best performance in distinguishing RAU from normal oral mucosa and other common oral mucosal diseases, achieving an AUC of 98.95%. This performance was comparable to that reported in studies using CNN-based algorithms for skin lesion classification.[26] With regard to the object detection model, the Pretrained YOLOV5 model achieved an AUC of 90.89% for detecting ulcerative lesions in the image, which represents an acceptable level of accuracy.[15]

In the classification task, we utilized four CNN-based deep learning models that exhibited varying degrees of performance. Our study yielded three notable findings. Firstly, the use of Pretrained models significantly enhanced performance. Pretrained models are a typical application of transfer learning that reduces the amount of labeled data required for downstream tasks and improves generalization performance.[27] Secondly, in our experiments, both ResNet18 and ResNet50 outperformed the DenseNet model, which is consistent with the findings of previous studies.[15,25] This could be attributed to the introduction of residual learning and shortcut connections in the ResNet model, which effectively address the issue of gradient vanishing in deep networks, thereby enhancing the model's stability during training.[28] Thirdly, compared to ResNet18, ResNet50 exhibited significantly superior performance in multiple indicators, suggesting that larger models can learn more diverse features and achieve better performance.[29] Regarding the object detection task, YOLOV5 no longer uses the two-stage method of classification first and then detection, as in Faster R–CNN.[17,30] Instead, it employs a grid system to divide the image and performs end-to-end regression calculations directly, significantly simplifying the computation process.[31] Moreover, after several generations of optimization and transformation, the current YOLOV5 exhibits higher detection accuracy and faster training and prediction speeds.

Our study had several limitations. The algorithm only performed a three-class classification for oral mucosal diseases and cannot make clear predictions for a wider range of oral mucosal diseases. Future multi-center large-sample studies may help address this limitation. Furthermore, its potential clinical application needs to be carefully discussed. Risk and activity assessment of RAU, as well as the formulation of individualized medical strategies, still require evaluation by clinical physicians, which cannot be replaced by AI algorithms at present. Further research and development are needed to overcome these challenges.

In conclusion, our study demonstrated that the deep learning model based on the CNN algorithm, using oral images, can not only perform three-class classification for

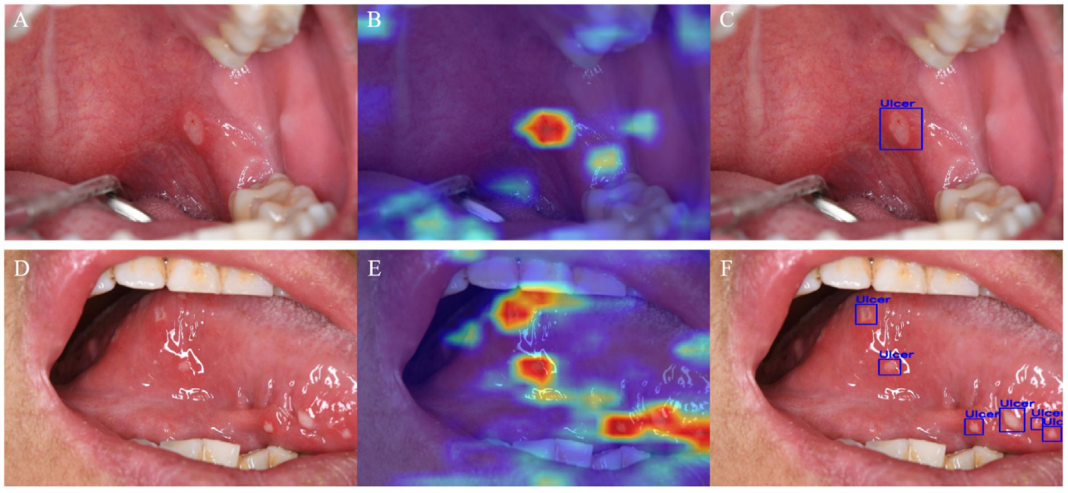

Figure 5 (A) The example image with a single ulcer lesion on the palatal mucosa. (B) Visualization results of deep learning features extracted by Grad-CAM method based on image A. (C) The true positive output of YOLOV5 detection based on image A. (D) The example image with multiple ulcer lesions on the tongue. (E) Visualization results of deep learning features extracted by Grad-CAM method based on image D. (F) The true positive output of YOLOV5 detection based on image D.

RAU, normal oral mucosa, and other oral mucosal diseases but also has the potential to accurately identify the location of lesions in RAU images. Based on our findings, we believe that deep learning technology has a promising future in the automatic diagnosis of oral mucosal diseases, represented by RAU. It may serve as a simple, non-invasive, and cost-effective screening tool to assist in clinical decision-making. However, further evaluation and validation are necessary before its widespread clinical application.

Declaration of competing interest

The authors have no conflicts of interest relevant to this article.

Acknowledgments

The National Natural Science Foundation of China (NSFC) (Grant Nos 82001047) and the Key Research and Development Program of Zhejiang Province (2020C03074).

References

1. Queiroz S, Silva M, Medeiros AMC, Oliveira PT, Gurgel BCV, Silveira É JDD. Recurrent aphthous ulceration: an epidemiological study of etiological factors, treatment and differential diagnosis. *An Bras Dermatol* 2018;93:341—6.
2. Hu Y, Guo H, He L, et al. The correlation between IFNG gene methylation and Th1|Th2 cell balance in ROU and the Interventional Study of Jiaweidaochi Powder. *Appl Biochem Biotechnol* 2023 (in press).
3. [Guidelines for the diagnosis and management of recurrent aphthous ulcers (draft)]. *Zhonghua Kou Qiang Yi Xue Za Zhi* 2012;47:402—4.
4. Scully C. Clinical practice. Aphthous ulceration. *N Engl J Med* 2006;355:165—72.
5. Hapa A, Aksoy B, Polat M, Aslan U, Atakan N. Does recurrent aphthous stomatitis affect quality of life? A prospective study with 128 patients evaluating different treatment modalities. *J Dermatol Treat* 2011;22:215—20.
6. Zeng X, Jin X, Zhong L, et al. Difficult and complicated oral ulceration: an expert consensus guideline for diagnosis. *Int J Oral Sci* 2022;14:28.
7. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med* 2020;26:900—8.
8. Min JK, Kwak MS, Cha JM. Overview of deep learning in gastrointestinal endoscopy. *Gut Liver* 2019;13:388—93.
9. Kühnisch J, Meyer O, Hesenius M, Hickel R, Gruhn V. Caries detection on intraoral images using artificial intelligence. *J Dent Res* 2022;101:158—65.
10. Dong D, Fang MJ, Tang L, et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. *Ann Oncol* 2020;31:912—20.
11. Venkadesh KV, Setio AAA, Schreuder A, et al. Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT. *Radiology* 2021;300:438—47.
12. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115—8.
13. Kolokythas A. Can Artificial Intelligence (AI) assist in the diagnosis of oral mucosal lesions and/or oral cancer? *Oral Surg Oral Med Oral Pathol Oral Radiol* 2022;134:413—4.
14. GitHub. LabelImg. Available from: https://github.com/heartexlabs/labelImg. [Accessed 1 April 2023].
15. Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P. Performance of deep convolutional neural network for classification and detection of oral potentially malignant disorders in photographic images. *Int J Oral Maxillofac Surg* 2022;51:699—704.
16. Su F, Sun Y, Hu Y, et al. Development and validation of a deep learning system for ascites cytopathology interpretation. *Gastric Cancer* 2020;23:1041—50.
17. Ren S, He K, Girshick R, Sun J, Faster R-CNN. Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell* 2017;39:1137—49.
18. Mushtaq M, Akram MU, Alghamdi NS, Fatima J, Masood RFJS. Localization and edge-based segmentation of lumbar spine vertebrae to identify the deformities using deep learning models. *Sensors* 2022;22:1547.
19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy: Institute of Electrical and Electronics Engineers, 2017:618—26.
20. Powers DMW. *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. arXiv preprint arXiv:2021.16010. 2020 (in press).
21. Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P. Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. *J Oral Pathol Med* 2021;50:911—8.
22. Huang L, Sun H, Sun L, et al. Rapid, label-free histopathological diagnosis of liver cancer based on Raman spectroscopy and deep learning. *Nat Commun* 2023;14:48.
23. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402—10.
24. Yang YJ, Bang CS. Application of artificial intelligence in gastroenterology. *World J Gastroenterol* 2019;25:1666—83.
25. Abdelmotaal H, Hazarbasanov R, Taneri S, et al. Detecting dry eye from ocular surface videos based on deep learning. *Ocul Surf* 2023;28:90—8.
26. Mahbod A, Schaefer G, Ellinger I, Ecker R, Pitiot A, Wang C. Fusing fine-tuned deep features for skin lesion classification. *Comput Med Imag Graph* 2019;71:19—29.
27. Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. *Proc IEEE* 2020;109:43—76.
28. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas*. NV: Institute of Electrical and Electronics Engineers, 2016:770—8.
29. Zhan H, Lin W-M, Cao Y. Deep model compression via two-stage deep reinforcement learning. In: *Machine Learning and Knowledge Discovery in Databases. Research Track: Proceedings of European Conference, Part I. Bilbao*. Spain: European Association for Artificial Intelligence, 2021:238—54.
30. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas*. NV: Institute of Electrical and Electronics Engineers, 2016:779—88.
31. Hafiz AM, Bhat GM. A survey on instance segmentation: state of the art. *Int J Multimed Inf R* 2020;9:171—89.