Oral Cavity Anatomical Site Image Classification and Analysis

Zhiyun Xue[*a], Paul C. Pearlman[b], Kelly Yu[c], Anabik Pal[a], Tseng-Cheng Chen[d], Chun-Hung Hua[e], Chung Jan Kang[f], Chih-Yen Chien[g], Ming-Hsui Tsai[e], Cheng-Ping Wang[d], Anil K. Chaturvedi[c], Sameer Antani[a]

[a]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

[b]Center for Global Health, National Cancer Institute, National Institutes of Health, Rockville, MD 20850

[c]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD 20850

[d]National Taiwan University Hospital, Taipei, Taiwan

[e]China Medical University Hospital, Taichung, Taiwan

[f]Chang Gung Memorial Hospital, Linkou, Taiwan

[g]Chang Gung Memorial Hospital, Kaohsiung, Taiwan

ABSTRACT

Oral cavity cancer is a common cancer that can result in breathing, swallowing, drinking, eating problems as well as speech impairment, and there is high mortality for the advanced stage. Its diagnosis is confirmed through histopathology. It is of critical importance to determine the need for biopsy and identify the correct location. Deep learning has demonstrated great promise/success in several image-based medical screening/diagnostic applications. However, automated visual evaluation of oral cavity lesions has received limited attention in the literature. Since the disease can occur in different parts of the oral cavity, a first step is to identify the images of different anatomical sites. We automatically generate labels for six sites which will help in lesion detection in a subsequent analytical module. We apply a recently proposed network called ResNeSt that incorporates channel-wise attention with multi-path representation and demonstrate high performance on the test set. The average F1-score for all classes and accuracy are both 0.96. Moreover, we provide a detailed discussion on class activation maps obtained from both correct and incorrect predictions to analyze algorithm behavior. The highlighted regions in the class activation maps generally correlate considerably well with the region of interest perceived and expected by expert human observers. The insights and knowledge gained from the analysis are helpful in not only algorithm improvement, but also aiding the development of the other key components in the process of computer assisted oral cancer screening.

Keywords: Oral Cancer, Deep Learning, Location Classification, Image Classification

1. INTRODUCTION

Oral cavity cancers are one of the most common cancers in the world, especially in Asia. According to WHO [1], the number of annual incidences and mortalities are around 380,000 and 180,000. One main contributing risk factor to oral cancers is the use and consumption of tobacco and alcohol, particularly smokeless tobacco and betel quid, which are major carcinogenic exposures in many low- and middle-income countries. The majority of oral cancers are squamous cell carcinomas (SCC), and the development of oral SCC are associated with a number of precursor lesions, such as leukoplakia and erythroplakia [2]. Like other cancers, early detection and prevention are important for reducing morbidity and death rate of oral cancer. In fact, oral cancer is often detected at late stage and has one of the lowest five-year survival rates (50% or less) among the major cancer types. Successful therapeutic management depends on a definitive, accurate, and timely diagnosis, so it is therefore crucial to identify and characterize precursor lesions. However, visual inspection of oral lesions is challenging since benign lesions can be easily confused with those that might be potentially malignant, particularly in early stage of the disease. Moreover, dysplasia/micro-invasive carcinomas can be present in clinically normal-appearing mucosa. Care-provider performance on predicting dysplastic oral lesions and oral SCC through visual inspection has been observed to be low and demands significant training and expertise [3].

One way to improve visual inspection quality is to use automated computer-assisted techniques that leverage machine learning and image processing techniques. While these techniques have been proposed for photographic images of the oral cavity for the automatic detection of oral cancers and precursor lesions in recent years [4-6], the number of such studies in the literature is limited, especially for those that use a large dataset with biopsy confirmed cases. Following promising results in proof-of-concept studies reported in the literature for cancer screening applications (such as the automated visual evaluation (AVE) work for cervical cancer screening and triage [7]), we aim to investigate AVE for triage of screen detected oral lesions. Led by the US National Cancer Institute with performance sites in Rockville, MD and the four cities in Taiwan, a study aiming to thoroughly investigate the natural history of oral SCC was initiated, through which clients with and without precursor lesions are longitudinally followed and biopsies and photographic images of the oral cavity are serially obtained. In this study, for each patient, pictures of different anatomical sites, viz. right buccal mucosa, left buccal mucosa, top of the mouth, dorsal tongue, ventral tongue, floor of the mouth, as well as pictures of several most severe lesions were taken. Care-providers who take the pictures may label them with the name of the anatomical sites and information for lesions. However, labeling is tedious and may be inconsistent. To reduce the labeling labor cost, and improve data quality, as a first step, we have developed a classifier to automatically separate the images into different categories based on the main captured anatomical location sites using the pilot batch of 250 patients from the dataset. This anatomical site classifier is also one preprocessing step in our pipeline of the AVE. In addition, since it is relatively easy for human to tell/explain the differences of images of different sites which are related with spatial information that can be specified in the images, we consider it as a good application for analyzing and demonstrating whether the classifier makes correct prediction based on the information that aligns well with human understanding/interpretation and why the classifier might fail for certain images. In this paper, we report and present our effort and results on the site image classification since, to the best of our knowledge, there is no such work reported in the literature for oral cavity photographic images. In the following sections, we will provide detailed descriptions on the dataset and its ground truth labeling, the classification network and the network visualization method, the results of experiments conducted, the discussion and analysis of the results, and the future work inspired by the analysis.

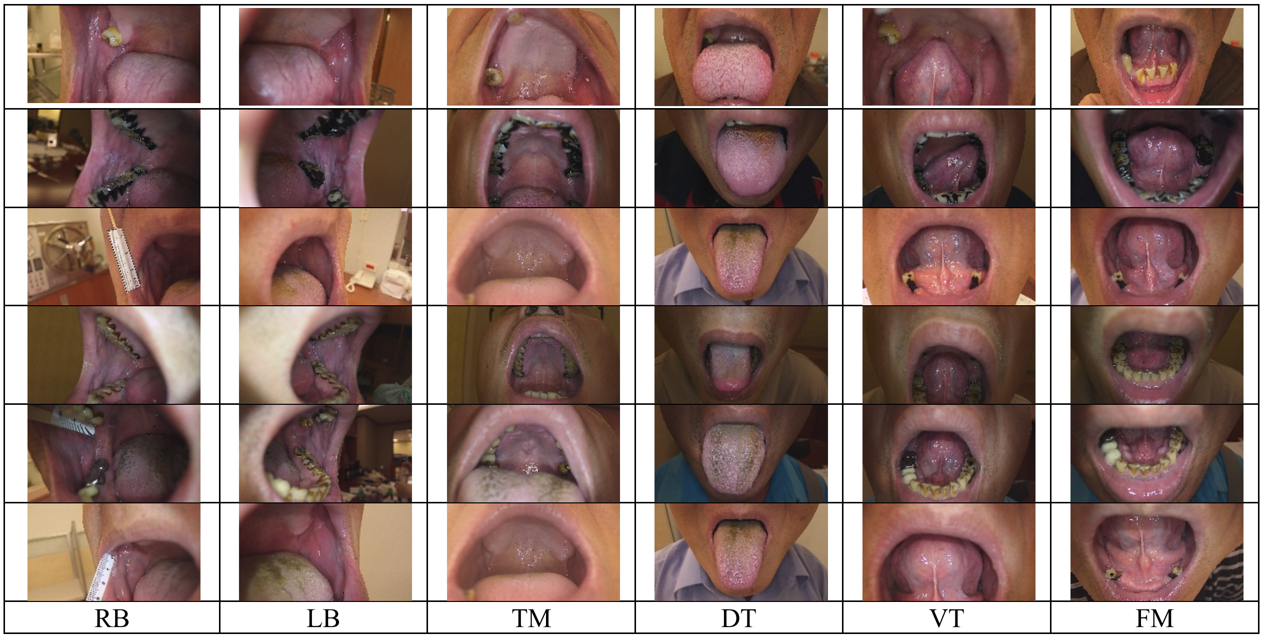


Figure 1. Example images from several patients

2. DATASET

The dataset we used for this task is the first batch consisting of 250 patients. According to the image collection protocol, for each precursor or cancer subject, the care-provider takes at least six pictures of the patient's mouth during each clinic visit. There is one picture for each of the six sites: 1) right buccal mucosa (labeled as RB); 2) left buccal mucosa (labeled as LB); 3) top of the mouth (labeled as TM); 4) dorsal tongue (labeled as DT); 5) ventral tongue (labeled as VT); and, 6) floor of the mouth (labeled as FM). If there are lesions, a picture of each of the top three most severe lesions are captured with a ruler to aid in lesion size assessment. In follow-up visits, pictures of all old and new lesions are taken and recorded on the oral examination form. The care-provider also names the files of the pictures with appropriate initials to indicate the (deidentified) patient ID, visit number, location site, and lesion number if applicable. For example, an image with name "HNC1001_B_RB_L1" is for subject ID HNC1001 at baseline from the right buccal mucosa and contains lesion number one. Based on this rule, we extracted the ground truth site label for each image. Not all images have location site specified (or specified with the defined site abbreviations) in the file name, and a patient may have several images of the same site from one visit or acquired during multiple follow-up visits. Figure 1 shows examples of images of several patients from different oral cavity sites. The images have variations in non-clinical factors such as illumination, view area and angle, ruler inclusion, in addition to factors related with subject or disease (such as the location, color and shape of the lip, teeth, tongue, and lesions).

3. METHODOLOGY

Deep learning has been actively studied to a great extent in the general image domain, especially for image classification. The deep networks usually have a huge number of parameters (in the order of millions) that are manipulated to achieve good performance. Therefore, large datasets are needed to train the network from the scratch to avoid overfitting in classification predictions. Like other medical domain applications which are limited by relatively small datasets, we started work with the networks that have ImageNet pretrained model available in order to take advantage of transfer learning. We selected the deep classification network called ResNeSt [8], a recent ResNet [9] variant network, to fine tune with our dataset in order to classify these images into six categories: RB, LB, TM, DT, VT, and FM, respectively. ResNeSt was proposed as an improvement over ResNet to aid in classification and downstream tasks with comparable computation cost. It incorporates channel-wise attention with multi-path representation and introduces a new block module called the "split attention" block. Specifically, in each "split attention" block, feature maps are first divided into several "cardinal groups" (as was done in one previous ResNet variant – ResNeXt [10]). Next, the feature maps in each cardinal group are separated channel-wise into subgroups ("splits"). The features across subgroup splits are combined ("attention") before being concatenated for all the groups. ResNeSt also applies network tweaks and several training strategies (such as augmentation, label smoothing, drop out regularization, and batch normalization) to improve its generalizability. Detailed information on network hyperparameters and specific augmentation methods that we used is provided in the next section.

Aside from high performance, we are also interested in examining whether the classification network makes the prediction based on the appropriate visual content in the image and understanding why the network does not classify certain images correctly. Deep network interpretation is a research area of growing interest. Surveys of various techniques are given in [11, 12]. Representative methods based on Convolutional Neural Networks (CNNs) include: saliency visualization map in [13], Local Interpretable Model-agnostic Explanations (LIMEs) [14], Class Activation Mapping (CAM) [15], and SHapley Additive exPlanations (SHAP) [16]. We applied GradCAM (Gradient-weighted Class Activation Mapping) [17] for the explanation of the classification results obtained by our ResNeSt model, given its ability to create high-resolution class-discriminative visualizations and provide insights into misclassified cases of the models. GradCAM is an improvement and generalization of CAM by using the gradient information flowing into the last convolutional layer of the CNN to indicate the neuron importance to the model's decision. It can be applied to a large range of CNN architectures without the need to modify the architectures and retrain the models. We extracted the GradCAM heatmaps of all the images in each set. Specifically, we analyzed the success and failure cases in the test set and discussed them with epidemiologists. In addition to GradCAM heatmaps, we also extracted the features in the ResNeSt model and examined the feature separability among different classes using t-SNE plots.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Table 1 lists the total number of images for each site in the dataset selection that contains 250 patients. In the experiments, we divided the dataset into training, validation, and test set at patient level (with 180, 20, and 50 patients, respectively). Table 2 lists the number of images of each category in each set.

Table 1. The number of images of each location site in the dataset

	RB	LB	TM	DT	VT	FM
Num. of images	609	623	396	404	396	393

Table 2. The number of images in each category in the training/validation/test set

	RB	LB	TM	DT	VT	FM
Train Set	433	448	284	289	284	283
Validation Set	52	55	35	37	34	34
Test Set	124	120	77	78	78	76

We used the ResNeSt50 network in our experiment. Since the dataset size is relatively small, we fine-tuned the ImageNet pretrained model with our training set. We also applied augmentation methods to increase the dataset size and improve the robustness of the model. We included random resized crop, color jittering, PCA normalization, random small rotation, and center crop. Note that we did not use the random horizontal/vertical flip, an augmentation method that is frequently used in other applications, as it would introduce confusion in differentiating class RB from LB. The loss function was categorical cross-entropy with label smoothing. The optimizer was Adam ($\beta 1 = 0.9$, $\beta 2 = 0.999$) with a learning rate of $5x10^{-4}$. The image size for model training was 224 x 224 pixels, and the batch size was 64. The model was trained for 100 epochs and the model at the epoch with the highest performance on validation set was selected.

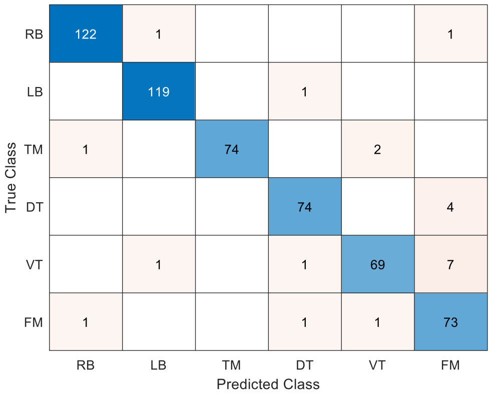
Table 3 lists the precision, recall, and F1 score of each class in the test set, as well as the macro average and weighted average of all classes for each of those three metrics. For precision, all classes have value higher than 0.95 except FM class whose value is 0.86. For recall, all classes have value no lower than 0.95 except VT class whose value is 0.88. The F1-score of each class is higher than 0.91 and the average F1-score on all classes is 0.96. The value of accuracy is 0.96. Figure 2 shows the confusion matrix chart obtained on the test set. From the confusion matrix chart, we can see, among these classes, the biggest confusion occurs between FM and VT. As some example images shown in Figure 1, the difference between images of these two classes might be very subtle and very challenging to differentiate.

Table 3. Precision, recall, and F1 score on the test set

	Precision	Recall	F1-score	Num. of Images
RB	0.98	0.98	0.98	124
LB	0.98	0.99	0.99	120
TM	1.00	0.96	0.98	77
DT	0.96	0.95	0.95	78
VT	0.96	0.88	0.92	78
FM	0.86	0.96	0.91	76
macro avg	0.96	0.96	0.96	553
weighted avg	0.96	0.96	0.96	553

To examine the representation and differentiation capability of features learned by the network, we extracted the features from the global average pooling layer in the model for the test images and visualized them with t-SNE plot. As shown in Figure 3, the features in different classes are very well separated.

We created GradCAM heatmaps to examine which areas in the image contribute the most to the classification decision the network makes for an input image. Figure 4 shows several examples of images that were classified correctly by the model and their corresponding heatmaps. For images correctly classified as TM, DT, VT or FM, as shown in Figure 4 (c) – (f), the highlighted regions in heatmaps generally concentrate on the right region of interest in each class. For images correctly classified as either RB or LB, the highlighted regions in heatmaps are relatively more diverse than those of the other four classes. As shown in Figure 4 (a) and (b), it appears that information both inside the mouth and outside the mouth were considered by the network. In addition, when there is a ruler which was put to indicate the approximate region of suspected lesion, the ruler area is usually highlighted, suggesting the characteristics of the ruler (may be its location or perspective) may also be used as a clue by the network to make decisions. We plan to develop algorithms to detect or segment the ruler for the subsequent task of lesion detection as there are no lesion manual markings other than the information of the existence of a lesion in the dataset at present and it is time-consuming and tedious to mark the lesion region.



True Class	RB	LB	TM	DT	VT	FM
RB	122	1				1
LB		119		1		
TM	1		74		2	
DT				74		4
VT		1		1	69	7
FM	1			1	1	73

Predicted Class

Figure 2. The confusion matrix on the test set

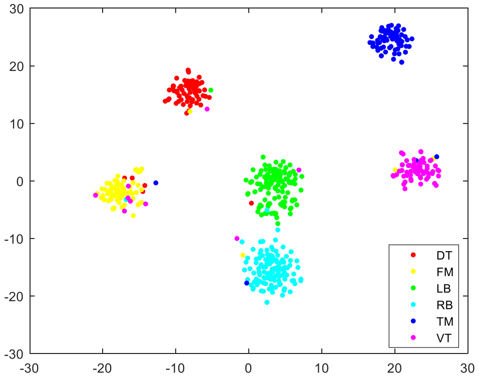


Figure 3. The t-SNE plot of the features

We also examined the heatmaps of the 22 images that were misclassified. These heatmaps may shed a light on why the network does not classify the image correctly. Several representative examples are given in Figure 5. For example, in Figure 5(a) and (c), the highlighted area is not at the right location, indicating the network does not use the right information to make decision. Comparing Figure 5(c) and the last image in Figure 4(e) (which is from the same patient but classified correctly by the model), the highlighted areas are in different sites. In Figure 5(b), the heatmap shows it concentrates mainly on the tongue which takes a significant large area in the image. This may be the reason why the network considered it as DT instead of LB. Figure 5 (d) image contains a large portion of both TM and DT area and the network misclassified it as VT. The image in Figure 5(e) is actually incorrectly annotated. It should be VT and the network classification result is correct. We also noticed there were a few images in validation set and training set that were mis-annotated while examining the results and heatmaps. It is common to have labeling errors in data collection due to human fatigue. A hybrid annotation approach in which users browse the automatic generated labels that are correct for the majority of images and only correct those misclassified ones would not only save time and effort but also reduce labelling errors. Some highlighted areas in the heatmap in Figure 5(f) are not within face/mouth but the background. We may consider removing background border regions in images for performance improvement in future.

5. CONCLUSIONS

In this paper, we report an important pre-step in developing an automatic approach toward improving the accuracy of visual evaluation for triage of screen detected oral lesions. Specifically, we applied deep learning technique to classify images captured from different major anatomical locations of the mouths to be inspected for oral cavity cancer assessment. The network achieves high classification accuracy on a randomly selected test set that is a subset from the pilot batch of data. In addition, we considered the network explanation and discussed the visualization results. Analysis and examination of both the correctly classified images and misclassified images lead to insights on where the network may concentrate and provide reasonable explanation on why the model derives certain predictions. They also demonstrate the effectiveness of the applied classification method. Based on the analysis, future work will explore topics such as background removal, ruler detection/segmentation, in addition to other main modules in the pipeline including lesion detection/segmentation, image quality control, and lesion classification.

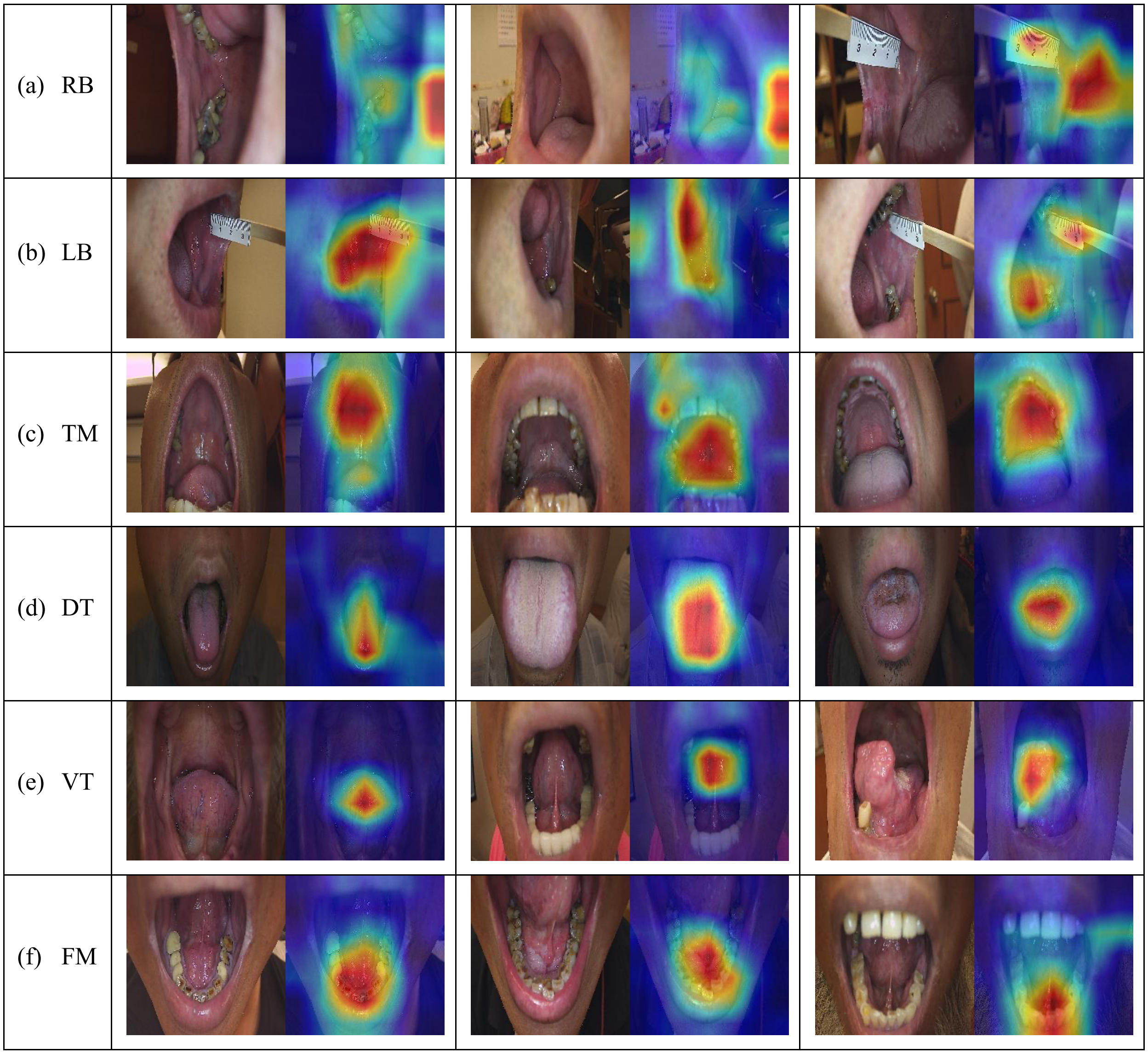


Figure 4. GradCAM heatmaps of correctly classified images

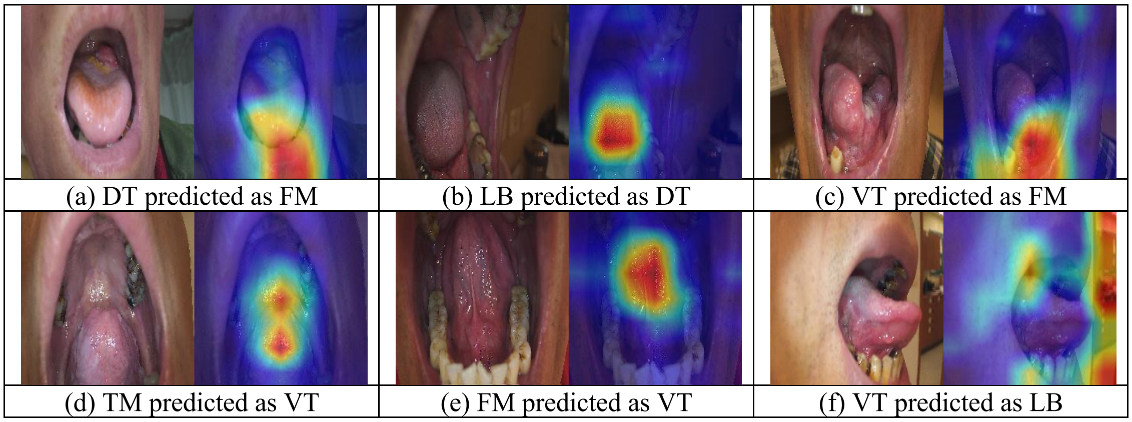

(a) DT predicted as FM
(b) LB predicted as DT
(c) VT predicted as FM
(d) TM predicted as VT
(e) FM predicted as VT
(f) VT predicted as LB

Figure 5. GradCAM heatmaps of in-correctly classified images

ACKNOWLEDGEMENT

This research was supported by the Intramural Research Program of the National Library of Medicine and the National Cancer Institute, both part of the National Institutes of Health (NIH). The authors want to thank Sandeep Angara at the National Library of Medicine for his valuable help and discussion on software programming.

REFERENCES

1. World Health Organization International Agency for Research on Cancer (IARC). Cancer Fact Sheets: Lip, oral cavity. Global Cancer Observatory 2020; Available from: https://gco.iarc.fr/today/data/factsheets/cancers/1-Lip-oral-cavity-fact-sheet.pdf.
2. P. H. Montero, and S. G. Patel, “Cancer of the oral cavity”, Surg Oncol Clin N Am. 2015 July; 24(3): 491–508. doi:10.1016/j.soc.2015.03.006.
3. J.B. Epstein, et al., “The limitations of the clinical oral examination in detecting dysplastic oral lesions and oral squamous cell carcinoma”, J Am Dent Assoc, 2012. 143(12): p. 1332-42.
4. K. Horio, et al., “Discrimination of oral mucosal disease inspired by diagnostic process of specialist”, Journal of Medical and Bioengineering, 2013. 2: p. 57-61.
5. M.Z.M. Shamim, et al., “Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer”, The Computer Journal, 2020.
6. Q. Fu, et al., “A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study”, EClinicalMedicine, 2020. 27: p. 100558.
7. A. Pal, Z. Xue, B. Befano, A. C. Rodriguez, L.R. Long, M. Schiffman, S.K. Antani, “Deep metric learning for cervical image classification”, IEEE Access, vol. 9, pp. 53266-53275, 2021, doi: 10.1109/ACCESS.2021.3069346.
8. H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Muller, and R. Manmatha, M. Li, A. Smola, “ResNeSt: split-attention networks”. https://arxiv.org/abs/2004.08955.
9. K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition”, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
10. S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, “Aggregated residual transformations for deep neural networks”, https://arxiv.org/abs/1611.05431.
11. Q. Zhan, S.-C. Zhu, “Visual interpretability for deep learning: A survey”, Front. Inf. Technol. Electron. Eng. 2018, 19, 27–39.
12. Y. Zhang, P. Tino, A. Leonardis, and K. Tang, “A survey on neural network interpretability”, arXiv preprint arXiv: 2012.14261 (2021)
13. K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps”, arXiv preprint arXiv:1312.6034 (2013).
14. M.T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier”, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM (2016).
15. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning deep features for discriminative localization”, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921-2929, doi: 10.1109/CVPR.2016.319.
16. S.M. Lundberg, and S. Lee, “A unified approach to interpreting model predictions”, Advances in Neural Information Processing Systems, 2017
17. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.