D'OraCa: Deep Learning-Based Classification of Oral Lesions with Mouth Landmark Guidance for Early Detection of Oral Cancer

Jian Han Lim[1(✉)], Chun Shui Tan[1], Chee Seng Chan[1], Roshan Alex Welikala[3], Paolo Remagnino[3], Senthilmani Rajendran[2], Thomas George Kallarakkal[4], Rosnah Binti Zain[4,5], Ruwan Duminda Jayasinghe[6], Jyotsna Rimal[7], Alexander Ross Kerr[8], Rahmi Amtha[9], Karthikeya Patil[10], Wanninayake Mudiyanselage Tilakaratne[4,6], John Gibson[11], Sok Ching Cheong[2], and Sarah Ann Barman[3]

[1] Centre of Image and Signal Processing, Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
[2] Head and Neck Cancer Research Team, Cancer Research Malaysia, 47500, Subang Jaya, Malaysia
[3] Digital Information Research Centre, Faculty of Science, Engineering and Computing, Kingston University, Surrey KT1 2EE, UK
[4] Department of Oral and Maxillofacial Clinical Sciences, Faculty of Dentistry, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
[5] Faculty of Dentistry, MAHSA University, 42610 Bandar Saujana Putra, Jenjarom, Malaysia
[6] Centre for Research in Oral Cancer, Department of Oral Medicine and Periodontology, Faculty of Dental Sciences, University of Peradeniya, Peradeniya 20400, Sri Lanka
[7] Department of Oral Medicine and Radiology, BP Koirala Institute of Health Sciences, Dharan 56700, Nepal
[8] Oral and Maxillofacial Pathology, Radiology and Medicine, New York University, New York, NY 10010, USA
[9] Faculty of Dentistry, Trisakti University, Kota Jakarta Barat, Jakarta 11440, Indonesia
[10] Oral Medicine and Radiology, Jagadguru Sri Shivarathreeshwara University, Mysuru 570015, Karnataka, India
[11] Institute of Dentistry, University of Aberdeen, Aberdeen AB25 2ZR, UK

Abstract. Oral cancer is a major health issue among low- and middle-income countries due to the late diagnosis. Automated algorithms and tools have the potential to identify oral lesions for early detection of oral cancer. In this paper, we aim to develop a novel deep learning framework named D'OraCa to classify oral lesions using photographic images. We are the first to develop a mouth landmark detection model for the oral images and incorporate it into the oral lesion classification model as a guidance to improve the classification accuracy. We evaluate the performance of five different deep convolutional neural networks and MobileNetV2 was chosen as the feature extractor for our proposed mouth

B. W. Papież et al. (Eds.): MIUA 2021, LNCS 12722, pp. 408–422, 2021.
https://doi.org/10.1007/978-3-030-80432-9_31

landmark detection model. Quantitative and qualitative results demonstrate the effectiveness of the mouth landmark detection model in guiding the classification model to classify the oral lesions into four different referral decision classes. We train our proposed mouth landmark model on a combination of five datasets, containing 221,565 images. Then, we train and evaluate our proposed classification model with mouth landmark guidance using 2,455 oral images. The results are consistent with clinicians and the F_1 score of the classification model is improved to 61.68%.

Keywords: Deep learning · Classification · Oral lesions · Mouth landmark

1 Introduction

Oral cancer is one of the most common malignant tumor with high risk in low- and middle-income countries (LMICs). There were an estimated 354,864 new cases of cancers of the oral cavity, and 177,384 deaths in 2018 [7]. Smoking, alcohol use and chewing of betel quid are the major risk factors for oral cancer [1,26,29,33]. Many people are unaware that cancer could arise in the oral cavity because of poor awareness of cancer-related symptoms. The early detection of oral cancer is essential for better survival. Oral cancer is often preceded by lesions termed as oral potentially malignant disorders (OPMDs), which are easily visible for early detection without the need of special instruments. Based on the appearances of oral lesions, specialists can make decisions on next course of action according to their clinical experience [40]. However, due to the limited effort towards screening and early detection, most patients affected by oral cancer are diagnosed at advanced-stages [29].

Artificial Intelligence (AI) has been adopted in various industries to improve the efficiency as well as to reduce the cost. Recent advances in deep learning techniques have improved the performance of AI models in various domains that can achieve or even outperform human level performance in cognition related tasks [28]. Deep learning has also gained popularity and made remarkable progress in the medical field to perform clinical diagnosis such as classifying skin lesions [12,34], detecting pneumonia from chest X-rays [4,31] and enhancing visualization of pathologies [15,23,25]. The development of deep learning techniques has yielded impressive results in the medical field, but it is not meant to replace humans, rather to assist humans and improve the efficiency.

For the past few years, early detection of oral cancer using deep learning techniques has been a significant research area. Specifically, deep learning algorithms are trained to capture fine-grained features of oral lesions and identify the specific visual patterns of oral cancer. The previous works are mainly based on different types of images such as multidimensional hyperspectral images [19], computed tomography (CT) images [44], microscopic images [3,13,22], autofluorescence images [37,39] and photographic images [14,41,42]. In this work, we propose

a novel deep learning framework to classify the oral lesions from photographic images into four different referral decision classes. Our proposed framework consists of a mouth landmark detection module and oral lesion classification module (Fig. 1). We design a new mouth landmark model to detect the location of the mouth and use it as an explicit feature to guide the classification model.

The contributions are twofold: i) To the best of our knowledge, we are the first to develop mouth landmark detection model that can detect the location of the mouth from the oral images. Existing facial landmark detection models do not work on oral images which do not consist of the entire human face. ii) We propose a novel oral lesion classification framework, namely D'OraCa with mouth landmark guidance for early detection of oral cancer. Experiments show that the performance of the classification model improves significantly with the mouth landmark guidance (Tables 4 and 5).

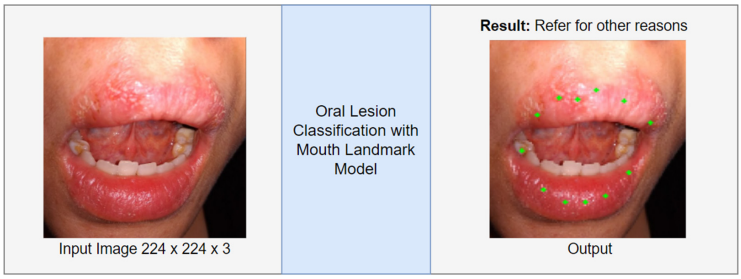

Fig. 1. Proposed oral lesion classification model with mouth landmark guidance that takes an oral images as input, detects the location of the mouth, and outputs the referral decision. On this example, proposed model correctly detects the mouth landmark and classifies the oral lesion as 'Refer for other reasons'.

2 Related Work

This section reviews the most relevant works related to the current research on oral lesion classification models and mouth landmark detection models.

2.1 Mouth Landmark Detection

There are no existing works on mouth landmark detection for oral images. However, there are a few studies on mouth features detection for front views of closed or slightly open mouth images. In [5], the authors focused on finding out the mouth candidates by segmenting the image based on skin-color. The input image must be a human face taken from the front view. It is not applicable for mouth images. Pantic et al. [30] proposed a mouth detection method with rule-based reasoning to extract the four mouth feature points based on template

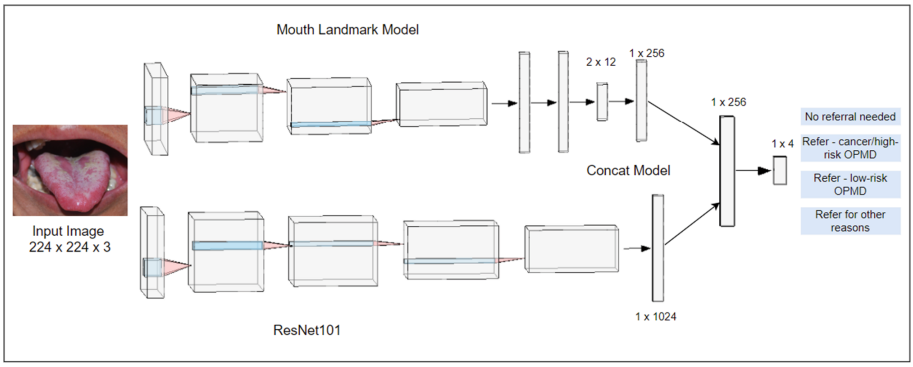

Fig. 2. Overview of the proposed architecture. Our network consists of two components. (Top) The first is the mouth landmark detection module, which detects the location of the mouth through a deep CNN and generates the mouth landmark feature. (Bottom) The second is the classification module. This module feeds oral images into ResNet-101 to obtain the fixed-size feature vector and fuses the two features to generate the final classification results.

matching. The proposed method depends on illumination conditions, assumes that the mouth-color pixel is red and segments based on it. The author mentioned that the proposed method can deal only with limited out-of-plane head rotations and expressionless mouth appearance. However, the oral images are usually taken at a different angle and the mouth is opened slightly larger to capture the oral cavity as shown in Fig. 4.

We also refer to facial landmark detection model as our related work due to the similar research area. The goal of facial landmark detection is to detect key points in human faces such as the eye corner, eyebrows, nose, chin and mouth. It is quite similar to our mouth landmark detection where the input image is only mouth area instead of the entire human face. Before the advent of deep learning, conventional facial landmark detection were mainly based on the template fitting method [2,45,48] and the cascaded regression-based method [8,20,38,43]. The template fitting method builds the face shape templates to fit the input images and estimates the landmark locations. While, the cascaded regression-based method estimates the landmark locations using image features with an initial guess and refines them using a cascade of machine learning models.

With the fast development of deep learning techniques in computer vision, deep learning based methods [9,11,21,27,46,47] have significantly boosted and outperformed both the template fitting method and cascaded regression-based method, creating a new state-of-the-art in facial landmark detection task. Most of them leverage deep convolutional neural networks (CNN) to learn facial features and predict the facial landmark in an end-to-end fashion. For instance, Yu et al. [46] proposed a deep deformation network and Lv et al. [27] presented a deep regression architecture with two-stage re-initialization for facial landmark detection. In [11], a style-aggregated network has been proposed to deal with the

large intrinsic variance of image styles for facial landmark detection. Chandran et al. [9] proposed an attention-driven architecture for facial landmark detection on very high resolution facial images without downsampling. Motivated by the development of facial landmark detection task, we develop a mouth landmark detection model based on deep CNN to detect the mouth key points in oral images.

2.2 Oral Lesion Classification

The previous works in oral lesion classification can be categorized based on the types of input images. We found that previous research was mainly limited to highly standardized images such as multidimensional hyperspectral images, CT images, microscopic images and autofluorescence images. Jeyaraj et al. [19] proposed a partitioned CNN algorithm to classify multidimensional hyperspectral images of the oral cavity into normal, benign or cancerous region. Xu et al. [44] developed a three-dimensional CNN algorithm for the early diagnosis of oral cancer. The proposed algorithm performed binary classification on CT images of oral cavity to profile oral tumors as benign or malignant. In [22], the authors showed that the fuzzy classifier were able to classify normal and oral cancer stages using the combination of texture based features from the histopathological images. Similar work has been done in [13] by using CNN to identify seven tissue classes from the histopathological images. Aubreville et al. [3] proposed a novel CNN-based approach for oral squamous cell carcinoma (OSCC) diagnosis on confocal laser endomicroscopy (CLE) images. Song et al. [37] and Uthoff et al. [39] presented a CNN binary classification method for oral cancer based on autofluorescence and white light images.

There are a few existing works involving the use of photographic images which is the most relevant to our work. The oral images can be captured directly using a smartphone and did not require specialized instruments. Fu et al. [14] developed a cascaded CNN model to perform binary classification on early detection of OSCC from photographic images. While, Welikala et al. [41] focused on detection and classification of oral lesions from photographic images using the Faster R-CNN [32] and ResNet-101 [16] network. Three separate models were built to explore different binary and multi-class image classification tasks. The authors further extended the work in [42] to compare the performance of five common CNN architectures on the binary classification of 'referral' vs. 'non-referral'. Transfer learning was applied on the CNN architectures pretrained on the ImageNet dataset [10] and fine-tuning to the smaller oral image dataset.

3 Methodology

In this section, we present our novel architecture for classification of oral lesions with mouth landmark guidance as shown in Fig. 2. In our proposed architecture, we integrate the mouth landmark detection model with the deep learning-based image classification model to classify oral lesions for the early detection of oral

cancer. Firstly, the mouth landmark detection model is employed to detect the location of mouth in an image. This explicit information is to tell where should the classification model focus on to look for the cancerous signs. The mouth landmark detection model will be further discussed in Sect. 3.1, then followed by the deep learning-based image classification model to predict the referral decision classes in Sect. 3.2. We compare the performance of the classification model with or without mouth landmark guidance. The objective is to prove and experiment on the hypothesis of the mouth landmark features might help the image classification model to focus on the mouth area in the image to increase the accuracy of the model.

3.1 Mouth Landmark Detection Module

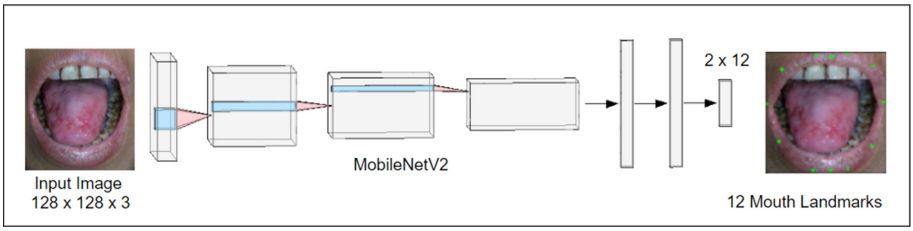

Fig. 3. Architecture of the proposed Mouth Landmark Detection Model: The oral image is fed into MobileNetV2, followed by two fully-connected layers to output 12 landmarks (green dots), indicating the location of mouth. Best viewed in color. (Color figure online)

The proposed mouth landmark detection module leverages the benefit of deep CNN to extract the image features and predict the mouth landmarks. This module is illustrated in Fig. 3. Technically, the input oral image I is fed into the deep CNN to extract the features. The features are then encoded by two fully-connected layers and a softmax layer to output the N number of mouth landmarks. The formula can be represented as:

$$p\left(M \mid I\right) = softmax(F_1(f_{CNN}(I))) \tag{1}$$

where f_{CNN} represents the deep CNN encoder, F_1 denotes the two fully-connected layers, I is the input oral image and $M = \{m_i\}_{i=1}^{Z}$ with $m_i \in \mathbb{R}$: $0 \leq m_i \leq 1$ is the output mouth landmark key points. The mouth landmark detection model is trained to minimize the mean squared error (MSE) loss L_{mse} as:

$$L_{mse}\left(Y_m, M\right) = \frac{1}{Z}\sum_{i}^{Z}(y_{m,i} - m_i)^2 \tag{2}$$

where Y_m represents the ground-truth landmark and $Z = 2 \times N$ is the number of mouth landmark key points, each landmark consists of two points to represent x-coordinate and y-coordinate.

3.2 Oral Lesion Classification Module

The classification module feeds the oral images into ResNet-101 to obtain the fixed-size feature vector. This feature vector is fused with the mouth landmark key points and classifies the image into the four different referral decision classes. The referral decision classes are "No referral needed", "Refer - cancer/high-risk OPMD", "Refer - low-risk OPMD" and "Refer for other reasons". The architecture of our network is summarized in Fig. 2. For the classification module, we have chosen ResNet-101 as the feature extraction network due to its superiority in the image classification tasks. ResNet-101 is a CNN with a much deeper layers, which consists of 101 layers with residual blocks that having shortcut connections to solve the vanishing gradient problem in training.

In order to guide the classification model with the proposed mouth landmark detection module, we encode the mouth landmark key points M using a fully-connected layer into a feature vector f_m with the size of 1×256. The ResNet-101 is then used to encode the oral image into a feature vector f_o with the size of 1×1024. Both feature vectors f_m and f_o are concatenated and processed through the last fully-connected layer followed by a softmax layer to output the final prediction. The formula can be represented as:

$$p\left(R \mid I\right) = softmax(F_2(f_m \oplus f_o)) \tag{3}$$

where $\oplus$ represents concatenation, F_2 denotes the last fully-connected layer and R is the predicted referral decision. The classification model is trained to minimize the cross-entropy loss L_{ce} as:

$$L_{ce}\left(Y_r, R\right) = -\sum_{i}^{C} y_{r,i} log(r_i) \tag{4}$$

where Y_r represents the ground-truth referral decision and C is the number of referral decision classes.

4 Experiments

4.1 Dataset and Metrics

There is no publicly available mouth landmark dataset for us to train our proposed mouth landmark detection model. Therefore, we make use of the existing facial landmark datasets, augmented the data to form our mouth landmark dataset for training and evaluation. We combine the face images from HELEN [24], 300W [35], AFW [48], IBUG [36], LFPW [6] and 300-VW [36] datasets to form a total of 221,565 face images. Each face images consists of 68 landmarks to indicate the location of eye corner, eyebrows, nose, chin and mouth. We preprocess the face images to extract only the mouth region with 20 mouth landmarks. We separate 180,000 images for training set, 20,000 images for testing set and 21,565 images for validation set.

To train and evaluate our proposed oral lesion classification model, we built a well-annotated oral image dataset which consists of 2,455 images collected from clinical experts from across the world. Each image was annotated by 1 to 7 expert clinicians to produce the referral decision, type of lesions, bounding box of the lesions, site, outline, etc. Each image was also linked to its metadata such as gender, age, smoking, alcohol use and chewing of betel quid. The annotations from multiple expert clinicians were processed with a novel strategy proposed by [41] to form a single set of annotations for the classification task. In this work, we only used the referral decision label as our classification objective. The dataset was split into 1,963 images for training set, 248 images for testing set and 244 images for validation set. The number of images for each referral decision class was shown in Table 1.

Table 1. Number of images according to the referral decision class

Referral decision	Training	Validation	Testing	Total
No referral needed	394	49	50	493
Refer - cancer/high-risk OPMD	509	63	64	636
Refer - low-risk OPMD	548	68	69	685
Refer for other reasons	512	64	65	641
Total	1963	244	248	2455

4.2 Mouth Landmark Detection Result

Table 2. Comparison between different deep CNNs on the mouth landmark testing set. The bold numbers represent the best result.

Methods	Mean square error (MSE)	
	(20 landmarks)	(12 landmarks)
Custom network	0.04567	0.04716
MobileNetV2	0.04454	**0.04239**
MobileNetV3	0.04658	0.05294
ResNet-50	0.04415	0.04948
ResNet-101	0.04543	0.04948

We evaluate the performance between different deep CNNs as the feature extractor for our proposed mouth landmark detection model. We compare the performance of MobileNetV2 [18], MobileNetV3 [17], ResNet-50, ResNet-101 [16] and

a custom network. The custom network is built using 5 convolutional layers, 2 fully-connected layers and we apply max pooling after convolutional layers. These models were evaluated using MSE in 12 and 20 landmarks to measure the average squared difference between the estimated landmark values and the actual landmark value.

As shown in Table 2, ResNet-50 and MobileNetV2 achieved the lowest MSE in 20 landmarks and 12 landmarks detection task respectively. MobileNetV2 (12 landmarks) was chosen as the deep CNN for our proposed mouth landmark model and integrated into the classification model. This was due to the lowest MSE achieved by MobileNetV2 and its lightweight model compared to the other methods. As the proposed mouth landmark detection model will be built into a mobile app in the future, a smaller size and faster inference time are required. As shown in Table 3, MobileNetV2 has the lowest number of parameters, smallest model size and fastest inference time.

Table 3. Comparison between different deep CNNs on the model size, number of parameters and inference time.

	Custom network	MobileNetV2	ResNet-50
No of parameters	7 million	2 million	23 million
Model size	30 MB	9 MB	90 MB
Inference time/Image	0.009 s	0.007 s	0.1 s

The qualitative results are shown in Fig. 4. Our proposed mouth landmark detection model can generate the correct mouth landmark in different angles of the mouth for oral images. For example, the top right image in Fig. 4 is showing the mouth captured from the right angle and the proposed model still can detect the correct mouth landmark. However, there are also some failure cases produced by our proposed model as shown in Fig. 5.

4.3 Oral Lesion Classification Result

Due to the lower number of oral images for the classification task, we implemented data augmentation on the dataset to generate more training samples through image pre-processing such as horizontal flip, horizontal shift and zoom. Note that data augmentation was not carried out on the validation and testing set. We used ResNet-101 pretrained on the ImageNet dataset as our deep learning model and performed transfer learning to our dataset, which as a result significantly reduced the training time and avoided overfitting the model.

To show the efficacy of our proposed mouth landmark guidance in the oral lesion classification model, we presented the quantitative result of the classification model with/without mouth landmark guidance on the test set in Table 4 and 5. Table 4 shows the result of the oral lesion classification model without mouth

Fig. 4. Qualitative results of the proposed mouth landmark detection model on a few images. It is noticed that the model is able to generate the correct mouth landmarks (green dots). Best view in color. (Color figure online)

Fig. 5. Incorrect mouth landmarks (green dots) generated by the proposed mouth landmark detection model. Best view in color (Color figure online)

Table 4. Oral lesion classification result without mouth landmark guidance, where TP, FP, TN and FN are true positive, false positive, true negative and false negative, respectively.

Class	TP	FP	TN	FN	Precision	Recall	F_1 score
No referral needed	20	13	185	30	60.61%	40.00%	48.19%
Refer - cancer/high-risk OPMD	49	37	147	15	56.98%	76.56%	65.33%
Refer - low-risk OPMD	34	27	152	35	55.74%	49.28%	52.31%
Refer for other reasons	40	28	155	25	58.82%	61.54%	60.15%
Macro-average					58.04%	56.84%	56.50%

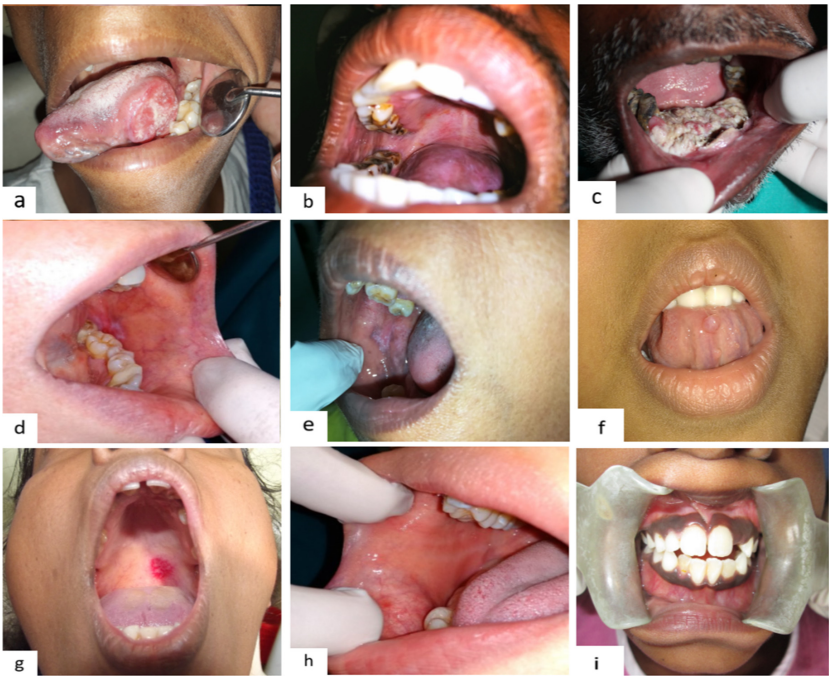

Fig. 6. Result of oral lesion classification model with mouth landmark guidance. (a), (b) and (c) are correctly classified as 'Refer - cancer/high-risk OPMD'. (d) and (e) are correctly classified as 'Refer - low-risk OPMD'. (f) and (g) are correctly classified as 'Refer for other reasons'. (h) and (i) are correctly classified as 'No referral needed'.

Table 5. Oral lesion classification result with mouth landmark guidance.

Class	TP	FP	TN	FN	Precision	Recall	F_1 score
No referral needed	24	18	180	26	57.14%	48.00%	52.17%
Refer - cancer/high-risk OPMD	46	25	159	18	64.78%	71.88%	68.14%
Refer - low-risk OPMD	43	25	154	26	63.24%	62.32%	62.77%
Refer for other reasons	42	25	157	23	62.69%	64.62%	63.64%
Macro-average					61.96%	61.70%	61.68%

landmark guidance. The model can achieve a precision of 58.04%, a recall of 56.84% and a F_1 score of 56.50%. Table 5 shows the result of the oral lesion classification model with mouth landmark guidance. The model can achieve a precision of 61.96%, a recall of 61.70% and a F_1 score of 61.68%. With mouth landmark guidance, the F_1 score of each referral decision classes were improved significantly, especially the F_1 score of "Refer - low-risk OPMD" class increased from 52.31 to 62.77 with a 20% improvement. The qualitative results from the

oral lesion classification model with mouth landmark guidance are provided in Fig. 6. The results are consistent with clinicians.

5 Conclusion

We presented a novel deep learning framework to classify the oral lesions from photographic images into four different referral decision classes. We also developed a mouth landmark detection model that can detect the location of the mouth from the oral images. We showed that the oral classification accuracy improved significantly with the guidance of the mouth landmark detection model. The model was trained and validated on a well-annotated oral image dataset containing 2,455 images. In conclusion, the initial results show the effectiveness of deep learning in early detection of oral cancer and we believe our proposed method can greatly contribute to the medical field. In future, we plan to improve the model by building a larger dataset with well-annotated labels and make use of the risk factors information to train the model.

Acknowledgements. This work was supported by the Medical Research Council under grant MR/S013865/1.

References

1. Amarasinghe, H., Johnson, N., Lalloo, R., Kumaraarachchi, M., Warnakulasuriya, S.: Derivation and validation of a risk-factor model for detection of oral potentially malignant disorders in populations with high prevalence. Br. J. Cancer **103**(3), 303–309 (2010)
2. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3444–3451 (2013)
3. Aubreville, M., et al.: Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. Sci. Rep. **7**(1), 1–10 (2017)
4. Ayan, E., Ünver, H.M.: Diagnosis of pneumonia from chest x-ray images using deep learning. In: 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), pp. 1–5. IEEE (2019)
5. Bao, P.T., Nguyen, H., Nhan, D.: A new approach to mouth detection using neural network. In: 2009 IITA International Conference on Control, Automation and Systems Engineering (case 2009), pp. 616–619. IEEE (2009)
6. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2930–2940 (2013)
7. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. **68**(6), 394–424 (2018)
8. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vis. **107**(2), 177–190 (2014)
9. Chandran, P., Bradley, D., Gross, M., Beeler, T.: Attention-driven cropping for very high resolution facial landmark detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5861–5870 (2020)