Clinically Guided Trainable Soft Attention for Early Detection of Oral Cancer

Roshan Alex Welikala[1], Paolo Remagnino[1], Jian Han Lim[2], Chee Seng Chan[2], Senthilmani Rajendran[3], 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], Sok Ching Cheong[3,4], and Sarah Ann Barman[1]

[1] Digital Information Research Centre, Kingston University, United Kingdom
r.a.welikala@kingston.ac.uk
[2] Centre of Image and Signal Processing, University of Malaya, Malaysia
[3] Head and Neck Cancer Research Team, Cancer Research Malaysia, Malaysia
[4] Department of Oral and Maxillofacial Clinical Sciences, University of Malaya, Malaysia
[5] Faculty of Dentistry, MAHSA University, Malaysia
[6] Centre for Research in Oral Cancer, University of Peradeniya, Sri Lanka
[7] Department of Oral Medicine and Radiology, BP Koirala Institute of Health Sciences, Nepal
[8] Oral and Maxillofacial Pathology, New York University, USA
[9] Faculty of Dentistry, Trisakti University, Indonesia
[10] Oral Medicine and Radiology, Jagadguru Sri Shivarathreeshwara University, India

Abstract. Oral cancer disproportionately affects low- and middle-income countries, where a lack of access to appropriate medical care contributes towards late disease presentation. Using artificial intelligence to facilitate the automated identification of high-risk oral lesions can improve patient survival rates. With image classification using oral cavity images and other forms of medical images, the information to be classified can often be extremely localized. To address this problem, we propose the use of convolutional neural networks with trainable soft attention. Further to this, we incorporate the use of localization loss to penalize the difference between attention maps and clinically annotated mask. This effectively allows clinicians to help guide soft attention. Improvements to the baseline were made, with an accuracy of 0.8333 and a ROC AUC of 0.8632, which equates to increases of 0.0245 and 0.0394, respectively. This accuracy corresponds to a sensitivity of 0.8469 and a specificity of 0.8208. Perhaps of more importance, is a model that demonstrates better capability at paying attention to the lesions in its decision making. Furthermore, visualizing resulting attention maps can help to strengthen clinical confidence in AI decision making.

Keywords: Deep Learning · Attention · Oral Cancer · Oral Potentially Malignant Disorders.

1 Introduction

Oral cancer has a major impact on global health, with an estimated 177,384 deaths in 2018 [1]. It is most prevalent in low- and middle-income countries (LMICs), where a lack of access to appropriate medical care contributes towards late disease presentation, and as a result survival rates are low. However, oral cancer presents unique opportunities, with oral lesions called oral potentially malignant disorders (OPMDs) preceding oral cancer for many patients. These lesions are visible for early detection and close monitoring without the need for invasive procedures.

Telemedicine can aid early diagnosis. The use of mobile phones has been field tested in a rural community [2], enabling two-way communication between primary healthcare practitioners and specialists located off-site. Integration of artificial intelligence (AI) into such approaches to facilitate the automated identification of high-risk oral lesions, will reduce the pressure on the limited number of specialists.

Currently, deep convolutional neural networks (CNN) provide the state-of-the-art results for many computer vision tasks. For medical image classification, the information to be classified can often be extremely localized. Therefore, the trainable attention mechanism can play a big role in medical image analysis, highlighting areas of interest whilst suppressing irrelevant parts of the image. This replicates the ability of clinicians to know where to look when making decisions.

Object detection equates to a type of hard attention and has been used to classify lesions in mammograms [3]. Hard attention can be effective, but in some cases may lead to loss of useful information when outside of the cropped region. Trainable soft attention offers the suppression of irrelevant background information without the need for cropped regions. It was introduced in machine translation [4] and later in image captioning. Jetley [5] demonstrated an increase in CNN image classification performance with the use of multi-scale soft attention. Pesce [6] interestingly improved the detection of chest radiographs containing pulmonary lesions, by penalizing differences in the soft attention maps and a subset of clinically annotated masks.

We first propose to adapt the multi-scale soft attention model derived by Jetley [5] for the novel application of attention for the early detection of oral cancer. Further to this, we incorporate multiple task learning, with the use of localization loss and classification loss. This adaptation is inspired by Pesce [6], with the localization loss quantifying the difference between soft attention maps and clinically annotated data; thus, effectively allowing clinicians to help guide soft attention. Instead of the attention maps indirectly driving the attention mechanism [6], our attention maps directly weight the feature vectors [5]. Whist an increase in image classification performance is the primary target, providing a model that reliably pays attention to the lesions in its decision making is of importance. Model interpretability to support clinical confidence in AI decision making can be aided by visualizing attention maps, with clinical confidence strengthened when attention demonstrates clinical guidance.

2 Related Work

The following reported studies, related to oral cancer, were CNN based. Aubreville [7] used InceptionV3 [8] to classify laser endomicroscopy images as clinically normal and carcinogenic. Halicek [9] used InceptionV4 for cancer detection from histology slides of excised tissue from the head and neck (included the oral cavity). Uthoff [10] used pairs of autofluorescence and white light mobile phone captured images as inputs to a VGG model [11] to perform suspicious vs. not suspicious classification. These reported studies utilized images obtained using specialized clinical procedures and advanced imaging systems.

Of more relevance to screening in LMICs are methods applied to standard oral cavity images. Welikala [12] explored/compared image classification and object detection to automate the early detection of oral cancer. Image classification was shown to be the more viable approach, whilst object detection struggled due to the indistinct nature of lesion boundaries. A study [13] demonstrated that simpler CNN architectures are more suitable when fine-tuning on an oral lesion dataset of limited size, VGG-19 [11] performed the best for the classification of referral vs. non-referral. Shamim [14] used VGG-19 for benign vs. pre-cancerous classification and Jubair [15] used a lightweight CNN for benign vs. suspicious classifications; both restricted to tongue lesions only.

3 Materials

A library of well-annotated images of oral lesions is currently being built. Accompanied by metadata of age, gender, and smoking, alcohol, and betel quid chewing status. Lesions in the images have been annotated with bounding boxes and each box has been assigned multiple labels that include lesion type, morphology, site, referral decision etc. At this initial phase of construction, the dataset included 2155 images. Each image has been separately annotated by 3-7 clinicians.

For this study we were only concerned with the bounding boxes and their referral decision labels. The annotations from multiple clinicians were combined with a novel strategy proposed by [12]. The annotated lesion's referral decision label was used as a single image label and if an image contained multiple annotated lesions then that with the highest referral decision severity was used. There were five classes in total, although the data was simplified to 'non-referral' vs. 'referral' as this would be the first step towards translation into clinical practice. The dataset was split into training, validation, and test sets (see Table 1).

Table 1. Image numbers according to class and set type.

Class	**Training**	**Validation**	**Testing**	**Total**
Non-referral	949	125	106	1180
Referral	795	82	98	975
Total	1744	207	204	2155

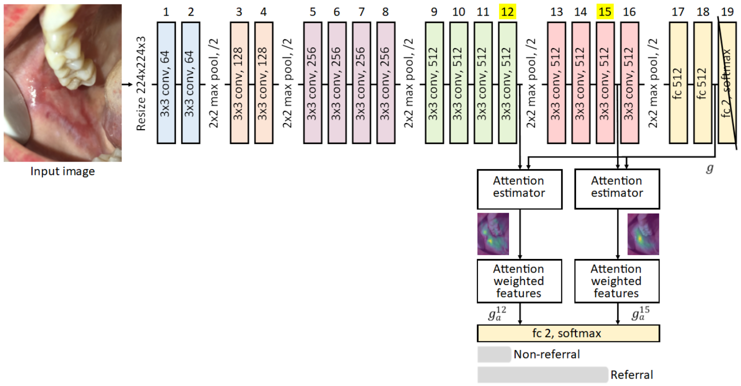

Fig. 1. Attention introduced at layer 12 and 15 of VGG-19.

4 Method

4.1 Attention Network

We adapted a soft trainable attention model proposed by Jetley [5], based on defining a compatibility score between local and global features. The intuition was that the compatibility score is intended to have a high value when the image patch described by the local features contained parts of the dominant image category. Therefore, a compatibility score assumed the role of attention values and was used to create a weighted combination of local features for performing image classification. This was a multi-scale approach achieved by leveraging local features from different intermediate stages of the CNN.

We implemented this approach using VGG-19 (reliable and outperforms newer architectures [13]), see Figure 1. The softmax classification layer was reduced to 2 neurons ('non-referral' vs. 'referral') and the prior fully connected layers were reduced to 512 neurons. Transfer learning was used to address the limited amount of data, with the VGG-19 model pre-trained on the ImageNet dataset [16]. We only fine-tuned from layer 12 and up to help avoid overfitting [13]. The attention mechanism was introduced at convolutional layers 12 and 15 (most effective), limited to fine-tuned layers.

Consider the local feature vector $l_{i,j}$ which was the output activations at the spatial location (i, j) of $n \times n$ spatial locations at a specific convolutional layer. Consider the global feature vector g which was normally fed into the final classification layer of the original VGG architecture.

The alignment model from [4] was re-purposed to calculate a compatibility score, defined as

$$c_{i,j} = f(l_{i,j} + g) \qquad (1)$$

where vectors $l_{i,j}$ and g were of equal dimensions (i.e. 1×512) and the function learnt using a single fully connected mapping to output a scalar compatibility score.

The compatibility scores were then normalized by a softmax operation to produce attention weights, defined as

$$a_{i,j} = \frac{exp(c_{i,j})}{\sum_{i=1}^{n}\sum_{j=1}^{n} exp(c_{i,j})} \tag{2}$$

Attention weights were then used to produce a single vector, defined as

$$g_a = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j} l_{i,j} \tag{3}$$

which was a weighted sum of the $l_{i,j}$ vectors. The g_a vector now replaced g as the global descriptor for the image.

This process was done separately for layers 12 and 15 (prior to max pooling), with 28×28 and 14×14 spatial locations, respectively. The resultant two vectors (g_a^{12} and g_a^{15}) were then concatenated into a single vector and passed through a softmax classification layer to produce class predictions.

4.2 Guided Attention

As standard, we minimized the classification loss (with cross entropy) of

$$L_{cls} = -\frac{1}{M}\sum_{m=1}^{M}\sum_{s=1}^{S} y_{s,m} log(\hat{y}_{s,m}) \tag{4}$$

where $y_{s,m}$ is the label from the one-hot class vector, $\hat{y}_{s,m}$ is the class probability from the prediction, S is the number of classes, and M is the number of images.

In addition to this, we also minimized attention based localization error, see Figure 2. If an image contained lesions, their bounding box annotations were converted to a binary mask, ones indicated pixels that belong to a lesion. The binary mask was then resized to the size of the attention map (either 28×28 or 14×14) for comparison. The attention map was rescaled to the range [0,1] using division by the maximum value of the map, to make it comparable to the binary mask. A pixel-wise mean square error was computed for attention based localization loss, which was then applied for all images that contained bounding boxes, defined as

$$L_{loc} = -\frac{1}{M_1}\sum_{m=1}^{M_1}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[b_{i,j,m} - a'_{i,j,m}]^2 \tag{5}$$

where b represents the binary masks, a' represents the rescaled attention maps, n represents both the height and width of the binary masks and attention maps, and M_1 is the number of images that contained bounding boxes. This quantified the difference between the binary masks and attention maps.

The network was then trained end-to-end to minimize a linear combination of classification loss and attention based localization loss, defined as

$$L = \lambda_1 L_{cls} + \lambda_2 L_{loc}^{12} + \lambda_3 L_{loc}^{15} \tag{6}$$

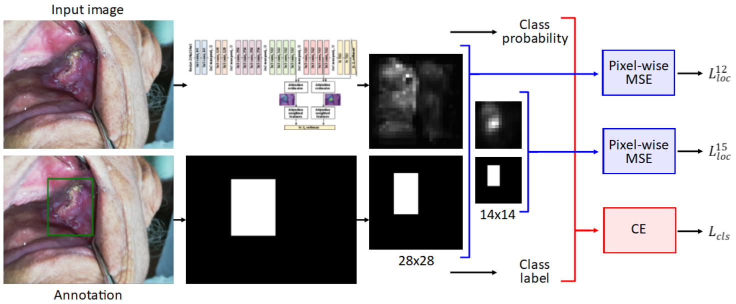

Fig. 2. Illustration of how the loss terms were calculated. MSE = mean square error, CE = cross entropy. Loss terms calculated across batch of images.

where attention based localization loss was calculated for the attention maps of convolutional layers 12 and 15. Empirically derived balancing parameters (λ) were used to weight the loss terms, with values of $\lambda_1 = 1.0$, $\lambda_2 = 2.5$, $\lambda_3 = 2.5$.

4.3 Technical Details

Backpropagation and stochastic gradient descent (SGD) with momentum was used for training. Images were rescaled to 224×224 pixels. Flipping, scaling, translation, and rotation were used to augment the training data (images/masks).

SGD mini-batch size was 128 images. Classification loss was class weighted to correct for the slight imbalance in the training data. We used a learning rate of 0.001, a momentum of 0.9, and a weight decay of 0.01. Batch normalization and a dropout of 0.5 were used on the first two fully connected layers. The network was trained for 100 epochs until convergence. The model was built on the training set and hyperparameters were derived from performance on the validation set.

A Nvidia GeForce RTX 2080 Ti graphics card with 11GB memory was used for training. This implementation used Keras and TensorFlow.

5 Results

Evaluation was performed on the test set, for the binary image classification task of 'referral' vs. 'non-referral'. We compared the performance of VGG with attention and VGG with clinically guided attention to the standard VGG (baseline); see Table 2. As the classes were approximately balanced in the test set, we used accuracy as a metric. For each approach, a confidence score threshold that produced the best operating point defined by the accuracy was selected. In addition, the ROC AUC is also provided to summarize the performance across different thresholds. The clinically guided attention model performed the best, with an accuracy of 0.8333 and a ROC AUC of 0.8632. This accuracy corresponds to a sensitivity of 0.8469, a specificity of 0.8208, a precision of 0.8137, and a recall of 0.8469. The precision-recall and ROC curves are provided in Figure 3.

Table 2. Image classification results.

Model	Accuracy	ROC AUC
Standard VGG	0.8088	0.8238
VGG with attention	0.8186	0.8498
VGG with clinically guided attention	0.8333	0.8632

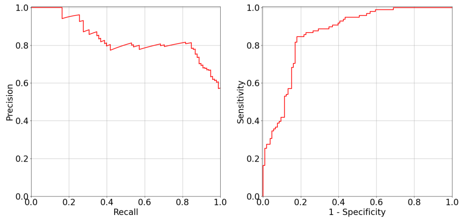

Fig. 3. Precision-Recall curve (AUC = 0.8367) and ROC curve (AUC = 0.8632) for the clinically guided attention model.

Outputs from the three approaches are provided in Figure 4 to enable a comparison of class predictions and attention maps. Figure 5 expands on further outputs of just the clinically guided attention model, providing a comparison of attention maps from layers 12 and 15.

6 Discussion and Conclusion

In this paper, we have demonstrated the performance of multi-scale trainable soft attention which has been clinically guided for the image classification of 'referral' vs. 'non-referral' with respect to oral cancer. The proposed model achieved an accuracy of 0.8333 and a ROC AUC of 0.8632, which is an improvement of 0.0245 and 0.0394 on the baseline model, and 0.0147 and 0.0134 on the model with attention (not clinically guided), respectively.

In image classification, the decision making process may not always use the most relevant parts of the images. This is evident from output examples of the baseline model shown in the second row of Figure 4, on occasions making decisions without even focussing on the lesions. The attention mechanism offers a much more targeted approach, highlighting areas of interest whilst suppressing irrelevant parts of the image. This is demonstrated in last two rows of Figure 4, showing the outputs for the two attention models, with lesions being more clearly highlighted. The bottom row does appear superior, with a greater coverage of the lesions, showing that attention benefits from clinical guidance, which also resulted in a higher classification performance.

Multi-scale attention allows complementary focus on different parts of the image at different scales, whilst still receiving clinician guidance on the general region of the lesion. Attention maps from layer 12 appear to attend to part details of lesions/surrounding area and those from layer 15 on whole lesions,

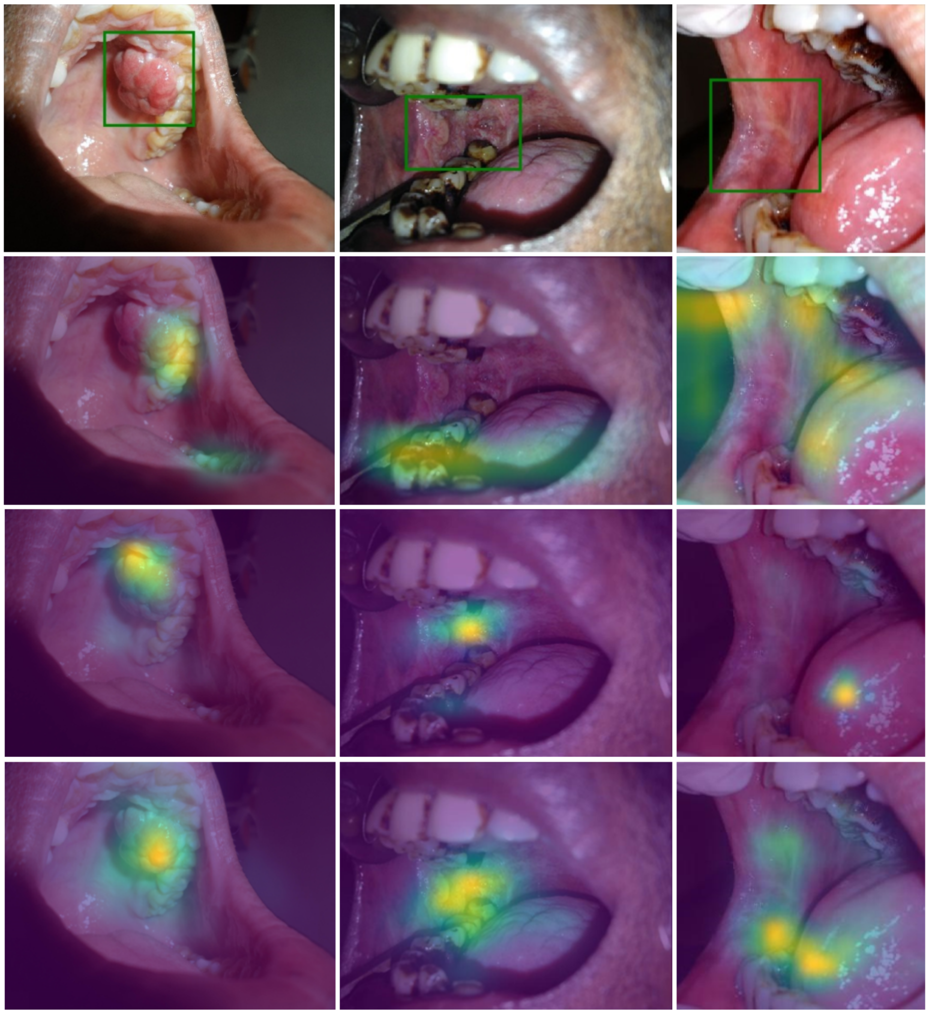

Fig. 4. Output comparison of the three models. Top row: input images, clinically annotated bounding boxes overlaid for visualization, class label (left to right) = [‘referral’, ‘referral’, ‘referral’]. Second row: Standard VGG, no trainable attention so Grad-CAM used for attention maps, class specific, predicted class and probability (left to right) = [‘referral’ 0.846, ‘referral’ 0.969, ‘non-referral’ 0.773]. Third row: VGG with attention, layer 15 attention maps, predicted class and probability (left to right) = [‘referral’ 0.901, ‘referral’ 0.991, ‘non-referral’ 0.536]. Bottom row: VGG with clinically guided attention, layer 15 attention maps, predicted class and probability (left to right) = [‘referral’ 0.739, ‘referral’ 0.990, ‘referral’ 0.662].

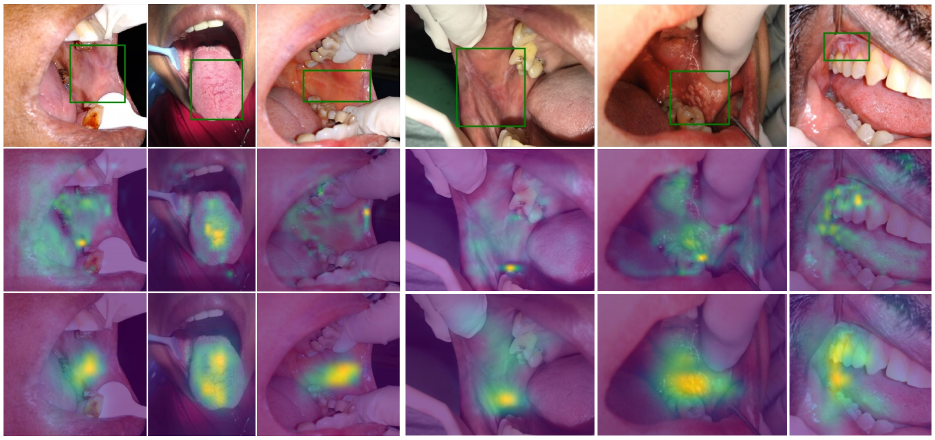

Fig. 5. Correct and incorrect outputs for the clinically guided attention model. Top row: input images, clinically annotated bounding boxes overlaid for visualization, class label (left to right) = [‘referral’, ‘non-referral’, ‘non-referral’, ‘referral’, ‘non-referral’, ‘referral’]. Middle row: layer 12 attention maps. Bottom row: layer 15 attention maps. Predicted class and probability (left to right) = [‘referral’ 0.738, ‘non-referral’ 0.836, ‘non-referral’ 0.929, ‘non-referral’ 0.687, ‘referral’ 0.820, ‘non-referral’ 0.881].

apparent from Figure 5. Incorrect classifications are shown in the last 3 columns of Figure 5, where localization still performs well (apart from the last column).

Related work reported values of 0.8500 and 0.8875 [10], 0.866 and 0.900 [7], 0.89 and 0.97 [14], and 0.867 and 0.845 [15], for sensitivity and specificity, respectively. Whilst we provide a comparative study of models in Table 2, currently, direct comparisons to related work can be difficult to make because their datasets and consequently their methodologies are designed to tackle different challenges. To drive competition, we aim to release a publicly available dataset.

Performances need to improve before translation into clinical practice. The future plan is to build a larger dataset (only a subset requires annotated bounding boxes), which is key to deep learning in order to improve results. A larger test set will be in place, whereby the generalizability of the model can be properly tested. An alternative is nested k-fold cross validation (keeping model selection in mind). We plan to make use of the metadata as input, and for models to output several of the other clinically assigned labels to gain further benefits of multi-task learning. High quality data will be promoted by putting constraints on what is acceptable and models will be built with a larger input image size.

In conclusion, this paper has demonstrated the use of trainable soft attention for the early detection of oral cancer, with improved performances when that attention was clinically guided. Importantly, the model demonstrates an improved ability to pay attention to the lesions in its decision making. In addition to post-hoc attention maps, trainable attention maps aid model interpretability, helping to strengthen clinical confidence in AI decision making, particularly when maps demonstrate clinical guidance on where to look.

Acknowledgments We would like to thank the Medical Research Council for providing funding (MR/S013865/1).

References

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **68**(6), 394–424 (2018).
2. Haron, N., Zain, R.B., Ramanathan, A., Abraham, M.T., Liew, C.S., Ng, K.G., et al.: m-Health for early detection of oral cancer in low-and middle-income countries. CA: Telemedicine and e-Health **26**(3), 278–285 (2020).
3. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. Scientific Reports **8**(1), 1–7 (2018).
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).
5. Jetley, S., Lord, N.A., Lee, N., Torr, P.H.S.: Learn to pay attention. arXiv preprint arXiv:1804.02391 (2018).
6. Pesce, E., Withey, S.J., Ypsilantis, P.P., Bakewell, R., Goh, V., Montana, G.: Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. Medical Image Analysis **53**, 26–38 (2019).
7. Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, C., Rodner, E., Denzler, J., et al.: Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. Scientific Reports **7**(1), 1–10 (2017).
8. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826 (2016).
9. Halicek, M., Shahedi, M., Little, J.V., Chen, A.Y., Myers, L.L., Sumer, B.D., et al.: Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. Scientific Reports **9**(1), 1–11 (2019).
10. Uthoff, R.D., Song, B., Sunny, S., Patrick, S., Suresh, A., Kolur, T., et al.: Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device with neural network classification for low-resource communities. PloS one **13**(12), e0207493 (2018).
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, (2014).
12. Welikala, R.A., Remagnino, P., Lim, J.H., Chan, C.S., Rajendran, S., Kallarakkal, T.G., et al.: Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. IEEE Access **8**, 132677–132693 (2020).
13. Welikala, R.A., Remagnino, P., Lim, J.H., Chan, C.S., Rajendran, S., et al.: Fine-Tuning Deep Learning Architectures for Early Detection of Oral Cancer. International Symposium on Mathematical and Computational Oncology, 25–31 (2020).
14. Shamim, M.Z.M., Syed, S., Shiblee, M., Usman, M., Ali, S.: Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer. arXiv preprint arXiv:1909.08987, (2019).
15. Jubair, F., Al-karadsheh, O., Malamos, D., Al Mahdi, S., Saad, Y., Hassona, Y.: A novel lightweight deep convolutional neural network for early detection of oral cancer. Oral Diseases,(2021).
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, 248–255 (2009).