IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 1, 2020, accepted July 6, 2020, date of publication July 20, 2020, date of current version July 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010180

Automated Detection and Classification of Oral Lesions Using Deep Learning for Early Detection of Oral Cancer

ROSHAN ALEX WELIKALA[1], PAOLO REMAGNINO[1], (Senior Member, IEEE), JIAN HAN LIM[2], (Graduate Student Member, IEEE), CHEE SENG CHAN[2], (Senior Member, IEEE), 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], JOHN GIBSON[11], SOK CHING CHEONG[3,4], AND SARAH ANN BARMAN[1]

[1]Digital Information Research Centre, Faculty of Science, Engineering and Computing, Kingston University, Kingston upon Thames KT1 2EE, U.K.
[2]Centre of Image and Signal Processing, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[3]Head and Neck Cancer Research Team, Cancer Research Malaysia, Subang Jaya 47500, Malaysia
[4]Department of Oral and Maxillofacial Clinical Sciences, Faculty of Dentistry, University of Malaya, Kuala Lumpur 50603, Malaysia
[5]Faculty of Dentistry, MAHSA University, Jenjarom 42610, Malaysia
[6]Centre for Research in Oral Cancer, Faculty of Dental Sciences, University of Peradeniya, Kandy 20400, Sri Lanka
[7]Department of Oral Medicine and Radiology, B. P. 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, Jakarta 11440, Indonesia
[10]Department of Oral Medicine and Radiology, Jagadguru Sri Shivarathreeshwara University, Mysuru 570 015, India
[11]School of Medicine, Medical Sciences and Nutrition, Institute of Dentistry, University of Aberdeen, Aberdeen AB25 2ZD, U.K.

Corresponding author: Roshan Alex Welikala (r.welikala@kingston.ac.uk)

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

ABSTRACT Oral cancer is a major global health issue accounting for 177,384 deaths in 2018 and it is most prevalent in low- and middle-income countries. Enabling automation in the identification of potentially malignant and malignant lesions in the oral cavity would potentially lead to low-cost and early diagnosis of the disease. Building a large library of well-annotated oral lesions is key. As part of the MeMoSA® (Mobile Mouth Screening Anywhere) project, images are currently in the process of being gathered from clinical experts from across the world, who have been provided with an annotation tool to produce rich labels. A novel strategy to combine bounding box annotations from multiple clinicians is provided in this paper. Further to this, deep neural networks were used to build automated systems, in which complex patterns were derived for tackling this difficult task. Using the initial data gathered in this study, two deep learning based computer vision approaches were assessed for the automated detection and classification of oral lesions for the early detection of oral cancer, these were image classification with ResNet-101 and object detection with the Faster R-CNN. Image classification achieved an F_1 score of 87.07% for identification of images that contained lesions and 78.30% for the identification of images that required referral. Object detection achieved an F_1 score of 41.18% for the detection of lesions that required referral. Further performances are reported with respect to classifying according to the type of referral decision. Our initial results demonstrate deep learning has the potential to tackle this challenging task.

INDEX TERMS Composite annotation, deep learning, image classification, object detection, oral cancer, oral potentially malignant disorders.

I. INTRODUCTION

Oral cancer is one of the most common cancers worldwide and is characterized by late diagnosis, high mortality rates and morbidity. GLOBOCAN estimated 354,864 new cases and 177,384 deaths in 2018 [1]. Two-thirds of the global incidence of oral cancer occurs in low- and middle-income countries (LMICs), half of those cases are in South Asia [2]. Tobacco use, in any form, and excessive alcohol use are the major risk factors for oral cancer. A factor most prominent in

The associate editor coordinating the review of this manuscript and approving it for publication was Essam A. Rashed.

South and Southeast Asia is the chewing of betel quid which generally is comprised of areca nut, slaked lime, betel leaf and may contain tobacco [3]. Nowadays, these quids are available commercially in sachets and are popular in public due to vigorous marketing strategies. Oral cancer is typically associated with late presentation, particularly in LMICs, where more than two-thirds present at late stages and as a result survival rates are poor [4]. Management of cancers, especially at the late stages, is very costly [5]. The lack of public awareness and the lack of knowledge of health professionals concerning oral cancer is an important reason for late detection [6].

Late diagnosis does not need to be a defining attribute as oral cancer is often preceded by visible oral lesions termed as oral potentially malignant disorders (OPMDs) which can be detected during routine screening by a clinical oral examination (COE) performed by a general dentist. If a suspicious lesion is identified the patient is referred to a specialist for confirmation of diagnosis and further management. Previous studies in India reveal screening has resulted in early diagnosis, down-staging of the disease and reduction in mortality amongst individuals who use tobacco and alcohol [7]. With most of the burden of oral cancer falling on LMICs due to the limited number of specialists and health resources, it is vital that screening programs must offer a low-cost and efficient approach to diagnosis. Such a viable approach would be the use of telemedicine. Haron *et al.* [8] showed a moderate to high concordance between the clinical diagnoses made by specialists performing a COE compared to when they review images captured from mobile phones. This remote consultation by specialists may improve the referral accuracy of screening programs. Taking this concept one step further by incorporating an automated detection system linked to artificial intelligence to analyze mobile phone images would be greatly beneficial.

Methods related to the automated diagnosis of oral cancer, OPMDs and benign lesions are largely based on microscopic images [9]–[12]. Other literature covers the use of multi-dimensional hyperspectral images of the mouth [13], the use of CT (computed tomography) images [14], the use of autofluorescence [15], [16] and fluorescence imaging [17] which focused on relative close-ups of the oral lesions and, finally, standard white light images which captured oral cavity structures [18]–[20].

Early publications in the field focused on texture based features, Thomas *et al.* [18] used the grey level co-occurrence matrix and grey level run-length, whilst Krishnan *et al.* [9] made use of higher order spectra, local binary pattern and laws texture energy. The more recent papers [10]–[17], [19], [20] have made the shift towards employing deep learning, which are artificial neural networks that consist of many layers of neurons and rely on large datasets and fast computing power to enable them to learn complex patterns. More specifically these publications made use of the deep convolutional neural network (CNN) whose architectures made the explicit assumption that the inputs were in the form of images. Since winning the ImageNet [21] image classification competition in 2012 with AlexNet [22], CNNs have gained wide popularity in the field of computer vision. A summary of related work is provided in Table 1.

Whilst CNNs are primarily used for image classification (an image classified into a certain class), building frameworks based around CNNs has shown considerable progress in the field of object detection (predicting bounding boxes and each box was classified into a certain class) for natural image datasets such as Pascal VOC (Visual Object Classes) [23] and COCO (Common Objects in Context) [24] which contained object classes such as cats, dogs, cars, bicycles etc. The highest accuracy object detectors to date were based on a two-stage approach popularized by the R-CNN family, which were region-based CNN approaches and included R-CNN [25], Fast R-CNN [26], Faster R-CNN [27] and most recently the Mask R-CNN [28] which could also output object instance segmentation. One-stage detectors such as YOLO (You Only Look Once) [29] and SSD (Single Shot Detector) [30] had the potential to be faster, at the cost of accuracy. Object detection frameworks have been explored in the medical imaging domain, with the Faster R-CNN being applied to colon polyp detection [31] and detection and classification of lesions on mammograms [32]. Anantharaman *et al.* [20] applied the Mask R-CNN to oral images using a dataset of 40 images, to detect the benign oral lesions of cold sores (herpes labialis) and canker sores (aphthous ulcers). Their evaluation was based on instance segmentation as opposed to bounding box detection.

To find a solution to the early detection of oral cancer, gathering reliable clinically labelled data is key to enable automated systems to be built. This has to be done at large scale to take advantage of deep learning. This paper presents a multidisciplinary collaboration that intends to build this dataset, providing clinical experts with the tools required to produce rich annotations. We also introduce a novel strategy to combine bounding box annotations from multiple clinicians. Using this data, we decided to assess two different approaches for the automated detection and classification of oral lesions, a deep learning based image classification framework and a deep learning based object detection framework. The former made use of ResNet-101 [33] which was a powerful CNN. The latter built on other work [20], but used the Faster R-CNN rather than the Mask R-CNN in order to focus on bounding box detection performance and, more importantly, this framework was now applied to oral cancer and OPMDs. Fig. 1 demonstrates a simplified version of the outcome expected from each approach. The initial dataset had 2155 images, despite being larger than the majority of cases presented in Table 1, this was small particularly when one considers the variation of the oral disease presentations. Despite this, the intention was to demonstrate proof of concept before proceeding with further investigation. This study was approved by the respective institutional review boards.

TABLE 1. Summary of related work.

Author	Title	Year	Description
Krishnan [9]	Automated oral cancer identification using histopathological images: a hybrid feature extraction paradigm	2012	Histopathological images. Texture based features with a fuzzy classifier used for 3-class image classification. 158 images from 42 individuals.
Thomas [18]	Texture analysis based segmentation and classification of oral cancer lesions in color images using ANN	2013	Standard white light images of oral cavity structures. Semi-automated active contour used for lesion segmentation. Followed by texture based features with a neural network used for 6-class image classification. 16 images.
Aubreville [10]	Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning	2017	Confocal laser endomicroscopy providing in vivo cell structure images. CNN used for binary image classification. 7894 images from 12 individuals.
Rana [17]	Automated segmentation of gingival diseases from oral images	2017	Fluorescence outputs used to colour augment white light images of close-ups of oral lesions captured using an oral imaging camera. Fully convolutional network used for lesion segmentation. 405 images from 150 individuals.
Anantharaman [19]	Oro vision: Deep learning for classifying orofacial diseases	2017	Standard white light images of oral cavity structures. CNN used for binary image classification. 75 images.
Anantharaman [20]	Utilizing Mask R-CNN for Detection and Segmentation of Oral Diseases	2018	Standard white light images of oral cavity structures. 2-class CNN based object detection and instance segmentation of lesions. 40 images.
Folmsbee [11]	Active deep learning: Improved training efficiency of convolutional neural networks for tissue classification in oral cavity cancer	2018	Histopathological images. CNN and fully convolutional network used for 7-class segmentation. 143 images.
Song [15]	Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning	2018	Autofluorescence and white light images of close-ups of oral lesions captured with a mobile phone attachment. CNN used for binary image classification. Image pairs from 170 individuals.
Uthoff [16]	Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device with neural network classification for low-resource communities	2018	In-depth details of the mobile phone attachment device used to capture images in [15].
Gupta [12]	Tissue Level Based Deep Learning Framework for Early Detection of Dysplasia in Oral Squamous Epithelium	2019	Histopathological images. CNN used for 4-class image classification. 2688 images from 52 individuals.
Jeyaraj [13]	Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm	2019	Hyperspectral images of the oral cavity. CNN used for 3-class image classification. 500 images.
Xu [14]	An Early Diagnosis of Oral Cancer based on Three-Dimensional Convolutional Neural Networks	2019	CT images of the oral cavity. Three-dimensional CNN used for binary image classification. 7000 images, not specified how many CT image sequences.

II. MATERIALS

With the long term goal of using telemedicine to facilitate the management of patients, Cancer Research Malaysia has developed a mobile phone App called MeMoSA® (**Mo**bil**e** **Mo**uth **S**creening **A**nywhere) [34]. MeMoSA® allows for the easy documentation of oral lesions through a mobile phone camera and enables seamless two-way communication between primary healthcare practitioners and specialists located off-site. The future scope would be the integration of automated detection systems to further assist with patient triaging at the primary care level.

In addition to this, MeMoSA® Annotate is a separate browser-based annotation tool, created to build a library of well-annotated images of oral lesions which can be used both for a better understanding of disease appearance and the development of artificial intelligence algorithms specifically geared to the early detection of oral cancer. Images are currently in the process of being gathered from clinical experts with image capture protocols in place to help with standardization. Metadata will also accompany the images and includes age, gender, and their status with respect to risk factors (i.e. smoking, alcohol and betel quid chewing). Each clinician can annotate multiple lesions per image using multiple rectangular bounding boxes and labels for each bounding box that include lesion type, lesion description, disease type, referral decision and numerous other labels.

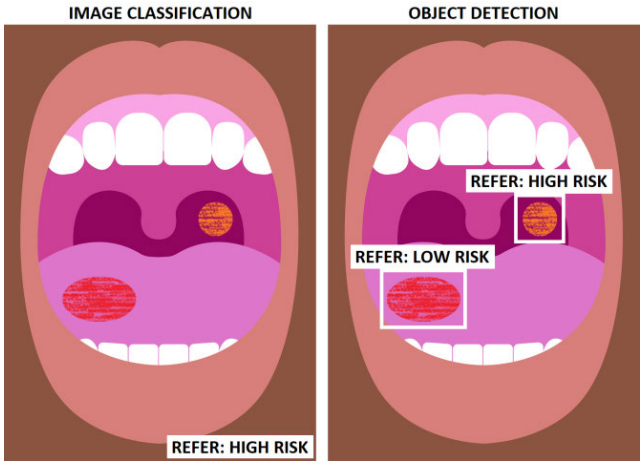


FIGURE 1. Image classification outcome versus object detection outcome.

The intention is for each image to be separately annotated by multiple clinicians to provide a richer data source, and to use agreement between clinicians which provides high concordance to a COE (to be presented separately).

For this initial study we had access to a set of 2155 oral cavity images from 1085 individuals, which were a mixture of images with and without lesions. This included images captured during the testing of the MeMoSA® App, images received from clinical experts and images downloaded from web search engines. The images were comprised of different oral cavity structures including buccal mucosa, tongue, palate, floor of the mouth etc. The images were of varying size, the largest was 5472 x 3648 pixels and the smallest was 119 x 142 pixels. Using MeMoSA® Annotate, 800 images were separately annotated by between 3-7 expert clinicians and to boost the size of the dataset, a further 1355 images were annotated by a single expert clinician. The aforementioned patient related metadata was not available for the entire dataset and therefore, was not included at this stage of the study. Also, we focused only on the referral decision label of the lesion. Table 2 describes what each referral decision represents in terms of disease type of the lesion.

TABLE 2. Referral decisions and corresponding disease types. NOS = not otherwise specified.

Referral Decision	Disease
Refer - cancer/high risk OPMD	Cancer
	Non-homogeneous leukoplakia
	Erythroplakia
	Tumor (NOS)
	Ulcer (NOS)
	Palatal lesions associated with reverse smoking
	Discoid lupus erythematosus (lip only)
	Oral submucous fibrosis
Refer - low risk OPMD	Homogeneous leukoplakia
	Lichenoid lesion/Lichen planus
	Tobacco Keratosis
	Actinic keratosis (lip only)
	Discoid lupus erythematosus (others)
Refer for other reasons	Benign
	Developmental abnormalities
No referral needed	Benign
	Developmental abnormalities
	Normal anatomical variant

III. METHOD

A novel strategy to combine bounding box annotations from multiple clinicians is introduced in Section III.A. The resultant annotated data was used to build and assess two computer vision approaches to tackle the automated detection and classification of oral lesions for the early detection of oral cancer. Firstly, deep learning based image classification which is covered in Section III.B, followed by deep learning based object detection which is covered in Section III.C.

A. COMPOSITE ANNOTATION

Given that images in our dataset have been annotated by several clinicians, the focus was to combine these multiple annotations into a single annotation which we referred to as a composite annotation. A composite annotation could still list separate bounding boxes (e.g. for separate lesions in an image), but no longer contained individual inputs for each clinician. Composite annotations were used for both training and evaluation purposes.

The task of annotating lesions has a degree of subjectivity leading to disagreement amongst the clinicians. As such, using a combination of their annotations would be more reliable and stable. The combination of conventional annotations could be achieved with a variety of schemes [35], among which the ''voting policy'' was the most common and simplest, and had been demonstrated to be as effective as more complicated strategies [36]. The majority vote would need to be adapted to this task (detailed below). As all our clinicians were of senior level, we opted not to take a weighted voting approach (e.g. based on the number of years of experience) and instead we simply used majority voting which assumed all clinicians were equally valued.

Although the majority vote was a simple approach, the nature of this task was not trivial, as the data took on the structure of that used for object detection frameworks (bounding boxes and class labels). Hence, the majority agreement had to be performed with respect to not only the bounding box class labels, but also the bounding box location. Using non-maximum suppression (NMS) was not viable, as we had no scores involved and it defeated the purpose of finding an agreement among the clinicians. Finding all bounding boxes that had an intersection over union (IoU) greater than 0.5 and then combining them by averaging their bounding box coordinates would be one approach to take to find agreement on location and would work for well-defined objects. For our task this would not be sufficient, given that some of the

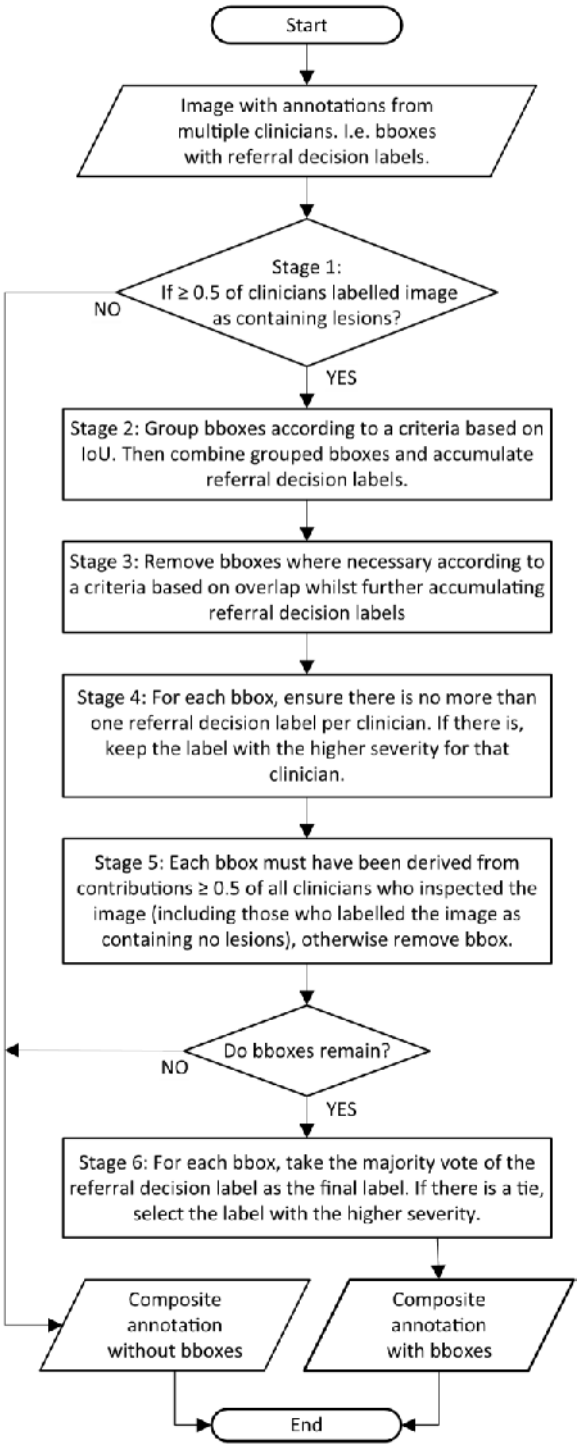


FIGURE 2. Strategy to combine annotations from multiple clinicians into a composite annotation. Bbox = bounding box.

lesions were not distinct in terms of what constitutes their boundary (image quality and composition were also issues). This led to the following inconsistencies among clinicians for some lesions: (I) considerable disagreement on the bounding box size, (II) clinicians annotating a lesion with a single bounding box versus clinicians that used multiple smaller bounding boxes.

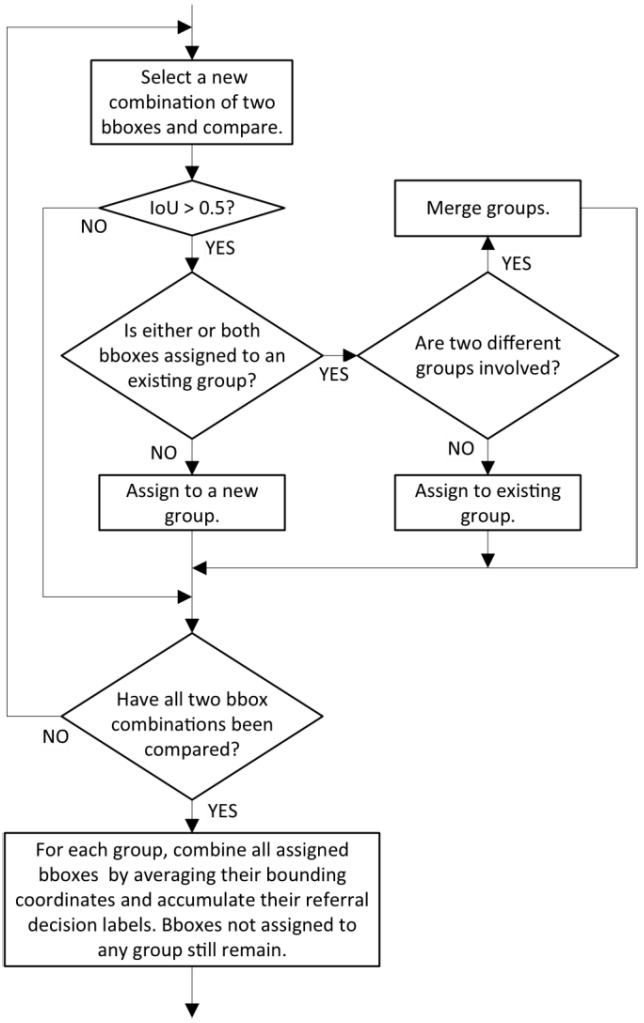


FIGURE 3. Group bounding boxes according to a criteria based on IoU. Then combine grouped bounding boxes and accumulate referral decision labels. This is an expansion of stage 2 of the strategy detailed in Fig. 2. Bbox = bounding box.

We propose a strategy to combine annotations from multiple clinicians into a composite annotation, focusing on location and the referral decision label of the bounding boxes. This strategy was applied to each image and handled the inconsistencies discussed so far for our dataset. Initially focusing on grouping and then combining bounding boxes that were deemed similar according to a criteria based on IoU, followed by a second opportunity that brought bounding boxes together according to a criteria based on simple overlap. A detailed breakdown is provided in Fig. 2, 3 and 4, with the latter two figures expanding on specifics of Fig. 2. It must be understood before proceeding that this task was unlike conventional object detection tasks [24] where objects regularly lie within another object (e.g. a person and their handbag) or where objects significantly overlap, these scenarios are less likely for oral lesions and there were no such cases in our current dataset. Several examples of the application of this strategy are provided in Fig. 5.

Following completion of the strategy above, the dataset of 2155 images was split into 1744 training images,

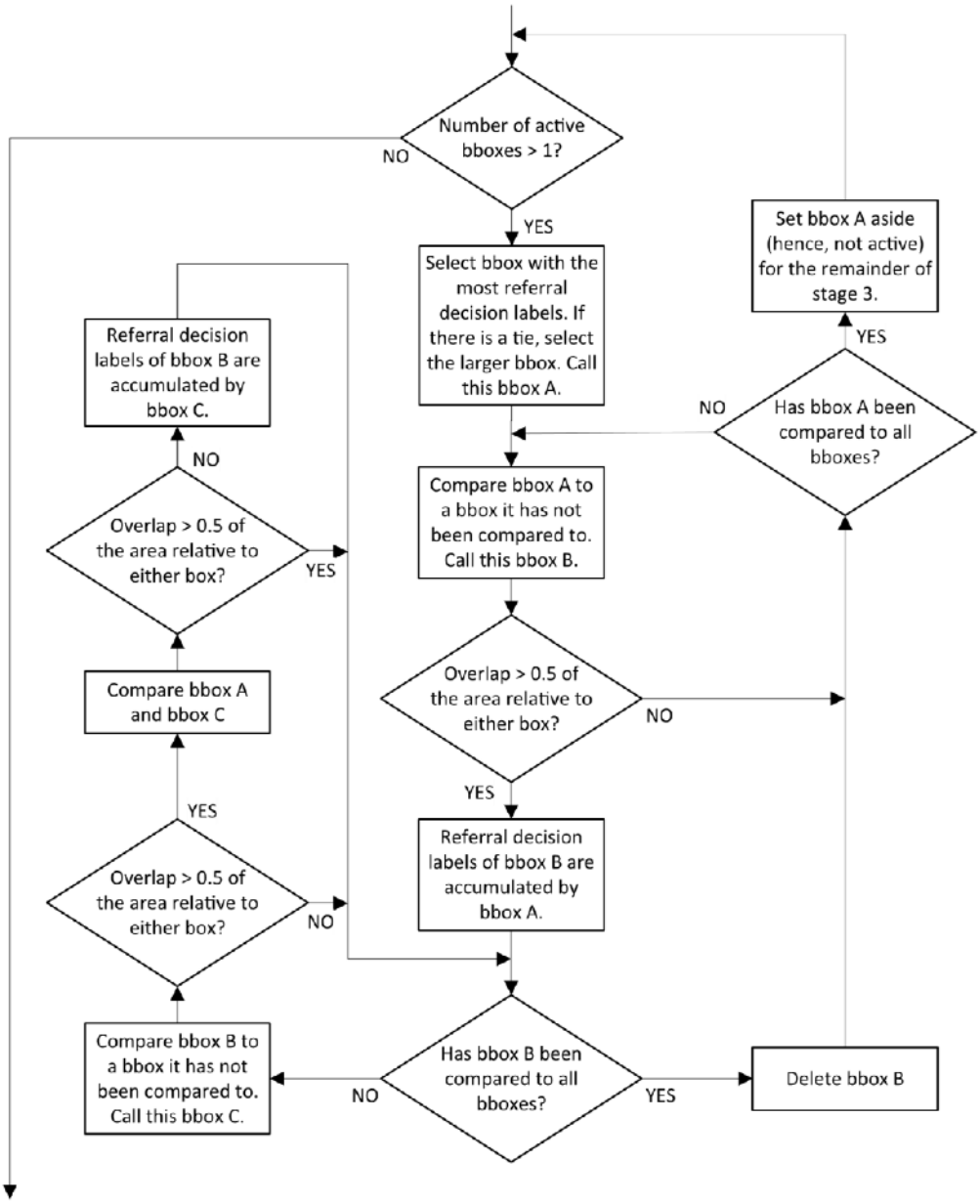


FIGURE 4. Remove bounding boxes where necessary according to a criteria based on overlap whilst further accumulating referral decision labels. This is an expansion of stage 3 of the strategy detailed in Fig. 2. Bbox = bounding box.

207 validation images and 204 testing images, images with no annotated lesions were not removed from the dataset. The split was random aside from the constraint that images from the same individual were confined to the same set. The total dataset amounted to 1433 annotated lesions whose breakdown is shown in Table 3. This equated to 1341 images which contained annotated lesions and 814 images without (808 images did not progress beyond step 1 and 6 images did not progress beyond step 5). The annotated data was geared towards object detection (Section III.C), the annotated data was simplified to make it applicable to image classification (Section III.B). In which 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. If no annotated lesions exist, then the image was labelled as 'no lesion'. A breakdown of the data on this image basis is shown in Table 4.

B. IMAGE CLASSIFICATION

Image classification refers to a process in computer vision that can classify an image into a certain class according to its visual content. We made use of deep learning based image classification, which used deep neural networks (deep

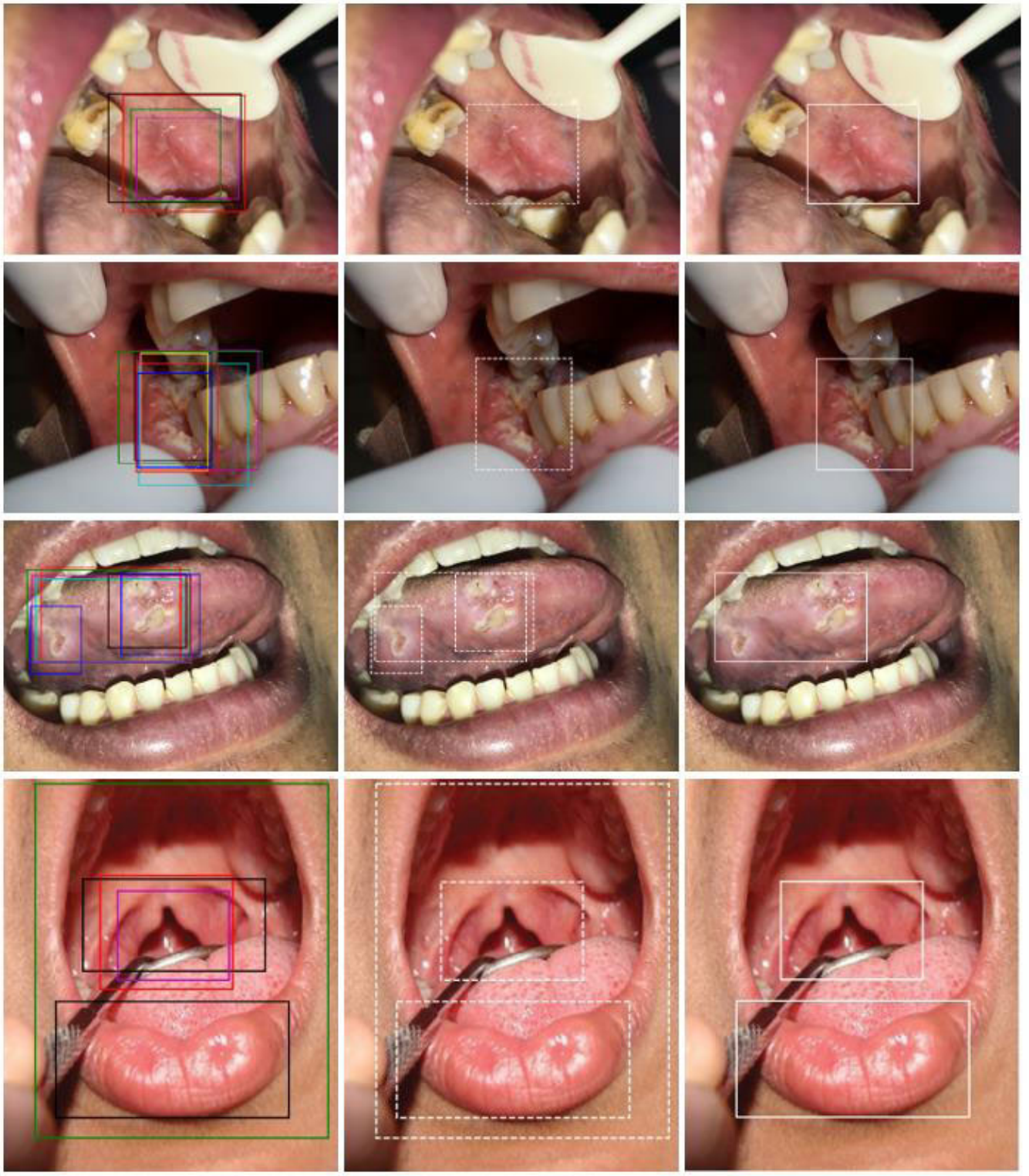


FIGURE 5. Combination of annotations from multiple clinicians to produce composite annotations. Left column: original bounding boxes, each clinician is represented by a colour. Middle column: bounding boxes after step 2. Right column: final bounding boxes with the following derived referral decisions from the top to the bottom row, 'refer - low risk OPMD', 'refer - cancer/high risk OPMD', 'refer for other reasons', 'refer for other reasons' (for both bounding boxes).

referring to the number of layers in the network). Specifically, CNNs whose architecture made the explicit assumption that the inputs were images, to exploit the strong spatially local correlation present in natural images. Automatically learning features at multiple levels of abstraction allowed a deep network to learn complex functions mapping the

TABLE 3. Annotated lesions numbers according to referral decisions and dataset type.

Dataset	No referral needed	Refer for other reasons	Refer - low risk OPMD	Refer - cancer/high risk OPMD	Total
Training	294	301	352	213	1160
Validation	39	24	33	29	125
Testing	46	36	41	25	148
Total	379	361	426	267	1433

TABLE 4. Image numbers according to referral decisions and dataset type.

Dataset	No lesion	No referral needed	Refer for other reasons	Refer - low risk OPMD	Refer - cancer/ high risk OPMD	Total
Training	665	284	290	299	206	1744
Validation	88	37	23	31	28	207
Testing	61	45	35	39	24	204
Total	814	366	348	369	258	2155

input to the output directly from data, without depending completely on hand-crafted features. The levels of features could be enriched by the number of stacked layers (depth). ResNet-101 [33] was a CNN with a depth of 101 layers (used residual blocks to combat training issues associated with very deep networks). We made use of this architecture due to its widespread use and high reported performances.

Transfer learning is a machine learning technique where a model trained on one task is re-purposed on a second related task. Transfer learning is popular in deep learning given the enormous resources required to train deep learning models, and we used transfer learning to address the fact that we currently have a limited amount of data. ResNet-101 was pre-trained on ImageNet dataset [21], which contained 1.2 million images with 1,000 classes (e.g. leopard, mushroom, go-kart). Our dataset was small and different to ImageNet, in this scenario the consensus to train a model was to freeze the initial layers and fine-tune the rest. To be thorough, we explored fine-tuning different extents of the network, from just the heads to all layers. The best model was achieved when freezing the layers prior to conv4_1 of ResNet-101 and then fine-tuning the rest of the system with our oral lesion dataset. Hence, this kept the low-level features and mid-level features unchanged.

Three separate image classification models were built to explore the task at varying levels of difficultly (detailed below). For each model, the number of neurons in the softmax classification layer of ResNet-101 was selected based on the number of classes, outputting class confidence scores. An outline of the multi-class model is provided in Fig. 6.

- Binary image classification of 'lesion' vs. 'no lesion'. i.e. 'no lesion' vs. the remaining four classes in Table 4 combined to produce the 'lesion' class.
- Binary image classification of 'referral' vs. 'non-referral'. i.e. 'no lesion' and 'no referral needed' combined to produce the 'non-referral' class vs. the remaining three classes of Table 4 combined to produce the 'referral' class.
- Multi-class image classification with five classes as detailed in Table 4.

1) IMPLEMENTATION DETAILS

Training: We used backpropagation and stochastic gradient descent (SGD) with momentum. A single scale was used for the images of 224 × 224 pixels. Horizontal and vertical flipping, scaling (80% to 120% both axes), translation (-20% to +20% per axis) and rotation were used to augment the training data.

Each SGD mini-batch had 128 images. Due to their class imbalance (see Table 4), the loss contributed from each class was weighted. The model was initialized with pre-trained weights and the model was fine-tuned from conv4_1 and up as explained earlier in this section. We used a learning rate of 0.001 for 100 epochs, learning rate decay presented no improvements. We used a momentum of 0.9 and a weight decay of 0.005. The model was built on the training set and hyperparameters were derived from performance on the validation set. The hyperparameters stated here were those for the multi-class model; the three models were explored separately.

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

Inference: The same image resizing from training was used.

C. OBJECT DETECTION

Object detection refers to a process in computer vision that determines where objects are located in a given image (normally with a bounding box) and which class each object belongs to [37]. We made use of deep learning based object detection, which combined classification and bounding box regression into a multi-task manner, specifically the Faster R-CNN [27].

The Faster R-CNN was a two-stage approach. The first stage was the region proposal network (RPN) which generated a sparse set of object/region proposals each with an objectness score. The second stage is known as the detection network which classified the region proposals into object classes and background. Both networks shared a common set of convolution layers. These common layers form the backbone/base of the framework which was a CNN (can be referred to as the base CNN), whose output from an intermediate convolutional layer provided rich hierarchical features for the input image.

To elaborate further, the RPN can be effectively considered as passing an image through the base CNN to produce a feature map. Followed by sliding a small network over every location of the feature map taking in an input of spatial size 3 × 3. This same process was applied at every location

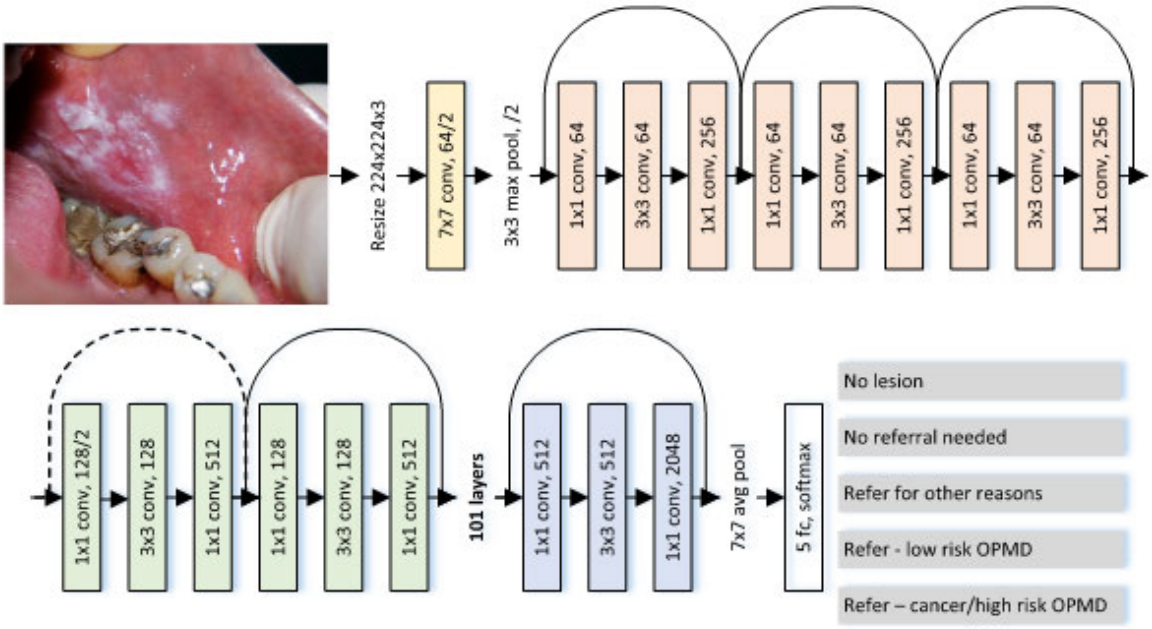


FIGURE 6. **Outline of ResNet-101 applied to multi-class image classification of oral images. Full details of the ResNet-101 architecture can be found in the original article [33].**

for a set of different anchors, which were fixed bounding boxes of various scales and aspect ratios (which made reference to the original image). For each anchor the small network would output refined bounding box coordinates using a regression layer and an objectness score using a classification layer that performed binary classification on whether it's an object or not an object. The RPN implemented this all efficiently in a fully convolutional manner. This resulted in large number of region proposals across a regular grid of the image. NMS suppression followed and further to this only the top ranked region proposals were sent through to the detection network. For each of the remaining region proposals the corresponding region on the feature map from the base CNN was converted to small fixed size using a pooling layer known as RoIPool [26]. This was followed by the detection network providing for each region proposal an output of further refined bounding box coordinates using a regression layer and the object class with a confidence score using a softmax classification layer. Class based NMS provided the final detections. Further details can be found in the original article [27].

Following on from updates from the original Faster R-CNN [27] paper, our model used ResNet-101 [33] with the feature pyramid network [38] as the base CNN. Also, RoIPool layer was replaced with the more effective RoIAlign layer introduced by the Mask R-CNN [28]. Transfer learning was applied, the Faster R-CNN model was pre-trained on the COCO dataset [24], which contained 328,000 images with 80 classes. Prior to this the base CNN (ResNet-101) was pre-trained on ImageNet dataset [21]. The best model was achieved when freezing the layers prior to conv5_1 of the base CNN and then fine-tuning the rest of the system with our oral lesion dataset.

Three separate object detection models were built to explore the task at varying levels of difficultly (detailed below). Our models output the bounding boxes and the class with confidence score for each detection. For each model, the number of neurons in the softmax classification layer of the detection network was selected based on the number of classes. An outline of the multi-class model is provided in Fig. 7.

- One object class representing all lesions.
- Two object classes for the lesions of 'referral' vs. 'no referral needed'. i.e. 'no referral needed' class as detailed in Table 3 vs. the remaining three classes combined to produce the 'referral' class.
- Four object classes with the four referral decision classes for lesions as detailed in Table 3.

1) IMPLEMENTATION DETAILS

Training: As the RPN and detection network shared convolutional layers of the base CNN, end-to-end joint training was used [39]. We used backpropagation and stochastic gradient descent (SGD) with momentum. A single-scale was used for the images [26], such that the shorter side was 800 pixels but ensuring that the scaling did not make the longer side > 1024 pixels, followed by zero padding to make them 1024 x 1024 pixels. Horizontal and vertical flipping, scaling (80% to 120% both axes) and translation (-20% to +20% per axis) were used to augment the training data.

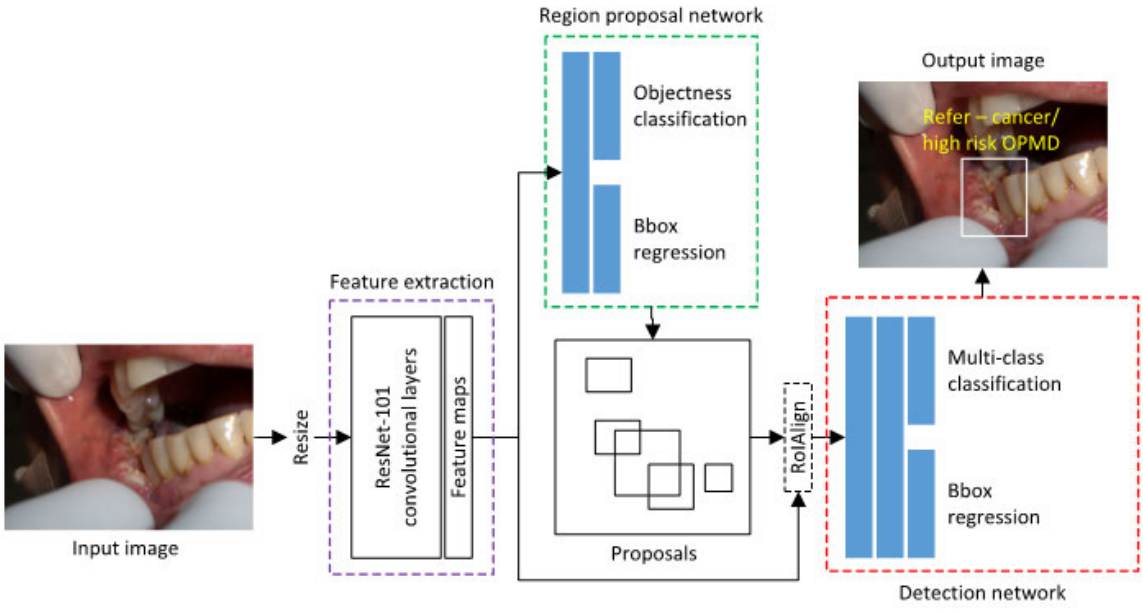


FIGURE 7. Outline of the faster R-CNN object detection framework applied to four-class oral lesion detection. Bbox = bounding box.

Each SGD mini-batch had 2 images. Each image had 64 anchors sampled to minimize the loss associated with the RPN. Sampled positive and negative anchors had a ratio of 1:1, defined as positives having an IoU ≥ 0.5 and negatives having an IoU < 0.3 with the composite annotation bounding boxes. NMS with an IoU threshold of 0.7 was applied to the output of RPN to leave 2000 region proposals per image and each image had 128 region proposals randomly sampled to minimize the loss associated with the detection network. Sampled negative and positive region proposals had a ratio of 3:1, defined as positives having an IoU ≥ 0.5 and negatives having an IoU in the range of 0.1 to 0.5 with the composite annotation bounding boxes (hard negative mining). The positive region proposal samples were made up from the 4 referral decision classes from Table 2. Due to their class imbalance (see Table 3), the loss contributed from each class in the classification head of the detection network was weighted (did not apply to the negative/background class).

The model was initialized with pre-trained weights and the model was fine-tuned from conv5_1 and up as explained earlier in this section. We used a learning rate of 0.001 for 100 epochs, learning rate decay presented no improvements. We used a momentum of 0.9 and a weight decay of 0.005. The model was built on the training set and hyperparameters were derived from performance on the validation set. The hyperparameters stated here were those for the four-class model; the three models were explored separately.

A Nvidia GeForce RTX 2080 Ti graphics card with 11GB memory was used for training. We used an open-source implementation of the Mask R-CNN by Matterport [40] and detached the mask head to derive the Faster R-CNN. This implementation used Keras and TensorFlow.

Inference: The same image resizing from training was used. Following on from the NMS applied to the output of RPN to leave 2000 region proposals (mentioned above), only the top 300 ranked of these were used for inference/detection. NMS with an IoU threshold of 0.3 and a specified class confidence score threshold was applied to the output of the detection network to produce the final detections.

IV. EXPERIMENTAL EVALUATION

The performance measures are detailed in Section IV.A, followed by the results of image classification in Section IV.B and the results of object detection in Section IV.C.

A. PERFORMANCE MEASURES

For image classification, the predicted class was compared to the expected class derived from composite annotation. Binary image classification had the outcomes of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) as detailed in Table 5. Precision, recall and the F_1 score (harmonic mean of precision and recall) could then be calculated as defined in Table 6, preferred to sensitivity, specificity and accuracy which could be misleading when the class distribution was imbalanced. These outcomes were based on selecting a confidence score threshold that produced the best operating point defined by the F_1 score.

For multi-class image classification, the outcomes were calculated per class (one vs. all approach), along with precision, recall and the F_1 score. With respect to the confidence score threshold, images with a score below the threshold were given the 'no lesion' class and the best operating point was defined by the macro-average F_1 score. This could be calculated using the definition of F_1 score in Table 6, but instead

TABLE 5. Outcomes of binary image classification.

Outcome	Description
True Positive (TP)	Correctly predicted the positive class
False Positive (FP)	Incorrectly predicted the positive class
True Negative (TN)	Correctly predicted the negative class
False Negative (FN)	Incorrectly predicted the negative class

TABLE 6. Performance measures.

Measures	Description
Precision	TP/(TP+FP)
Recall	TP/(TP+FN)
F_1 score	(2 x Precision x Recall)/(Precision + Recall)

TABLE 7. Outcomes of object detection, calculated per class (inspect only the predictions and composite annotations of the specified class).

Outcome	Description
True Positive (TP)	Correct detection. Predicted bounding box's IoU ≥ 0.5 with the composite annotation bounding box. If multiple correct detections of the same composite annotation bounding box existed, only one counted as correct and the rest were considered as wrong (FP).
False Positive (FP)	Wrong detection. Predicted bounding box's IoU < 0.5 with the composite annotation bounding box.
False Negative (FN)	Missed detection. Composite annotation bounding box not detected.

using the macro-average precision and macro-average recall, which were simply the precision and recall calculated for each class and then averaged. We prefer the macro-average over the micro-average for this multi-class tasks as the latter could be misleading when the class distribution was imbalanced.

For object detection, a detection was considered correct if the predicted bounding box's IoU ≥ 0.5 with the composite annotation bounding box and if the class was predicted correctly. The definition of the outcomes for object detection are detailed in Table 7. True negatives (TN) were not practical to define in object detection tasks (also they were not required). Outcomes along with precision, recall and the F_1 score were calculated per class. Based on a confidence score threshold that produced the best operating point defined by the macro-average F_1 score (or simply the F_1 score for the one object class model).

B. IMAGE CLASSIFICATION RESULTS

Evaluation was performed on the test set and the results of the three image classifications models is reported in Tables 8-10. The identification of images that contained lesions achieved a precision of 84.77%, a recall of 89.51% and an F_1 score of 87.07% as detailed in Table 8. The identification of images that required referral achieved a precision of 67.15%, a recall of 93.88% and an F_1 score of 78.30% as detailed in Table 9. Multi-class classification, which provided the type of referral decision, achieved a macro-average precision of 52.13%, a macro-average recall of 49.11% and a macro-average F_1 score of 50.57% as detailed in Table 10. Examples of outputs from the multi-class image classification model are provided in Fig. 8.

C. OBJECT DETECTION RESULTS

The three object detection models were evaluated on the test set and the results are reported in Tables 11-13. The detection of lesions achieved a precision of 46.61%, a recall of 37.16% and an F_1 score 41.35% as detailed in Table 11. The detection of lesions that required referral achieved a precision of 32.94%, a recall of 54.90% and an F_1 score of 41.18% as detailed in Table 12. The detection of lesions according to the type of referral decision achieved a macro-average precision of 17.71%, a macro-average recall of 39.74% and a macro-average F_1 score of 24.50% as detailed in Table 13. Examples of outputs from the four-class object detection model are provided in Fig. 9.

V. DISCUSSION

In this paper, we combined annotations from multiple clinicians using data provided from the first phase of collection. We then demonstrated the performances of deep learning based image classification and deep learning based object detection frameworks for the use on oral lesion detection and classification for the early detection of oral cancer. The use of deep learning means that complex patterns could be derived for tacking this difficult task. For image classification, ResNet-101 was used to classify the entire image. It achieved an F_1 score of 87.07% for identification of images that contained lesions, an F_1 score of 78.30% for the identification of images that required referral and a macro-average F_1 of score 50.57% for classifying images according to the type of referral decision. For object detection, the faster R-CNN (with ResNet-101 as the base CNN) was used to locate and classify oral lesions. Object detection achieved an F_1 of score 41.35% for the detection of lesions, an F_1 score of 41.18% for the detection of lesions that required referral and a macro-average F_1 of 24.50% for the detection of lesions according to the type of referral decision.

As part of the MeMoSA® project, the MeMoSA® Annotate tool is currently being used by our clinical collaborators to develop a library of well-annotated images of oral lesions. With time, this has the potential to become a very large and powerful resource, helping to better understand the disease and being crucial in providing a solution for the early detection of oral cancer. For this we have proposed a novel strategy to combine bounding box annotations from multiple clinicians to produce composite annotations, applicable to oral lesions in images as well as potentially to other similar medical tasks.

TABLE 8. Binary image classification results. 'lesion' vs. 'no lesion', the former was the positive class and the latter was the negative class.

Image class	TP	FP	TN	FN	Precision (%)	Recall (%)	F_1 score (%)
Lesion	128	23	38	15	84.77	89.51	87.07

TABLE 9. Binary image classification results. 'referral' vs. 'non-referral', the former was the positive class and the latter was the negative class.

Image class	TP	FP	TN	FN	Precision (%)	Recall (%)	F_1 score (%)
Referral	92	45	61	6	67.15	93.88	78.30

TABLE 10. Multi-class image classification results. Five classes, each separately evaluated using the one vs. all approach.

Image class	TP	FP	TN	FN	Precision (%)	Recall (%)	F_1 score (%)
No lesion	48	35	108	13	57.83	78.69	66.67
No referral needed	18	19	140	27	48.65	40.00	43.90
Refer for other reasons	14	18	151	21	43.75	40.00	41.79
Refer - low risk OPMD	16	19	146	23	45.71	41.03	43.24
Refer - cancer/high risk OPMD	11	6	174	13	64.71	45.83	53.66
Macro-averaged	-	-	-	-	52.13	49.11	50.57

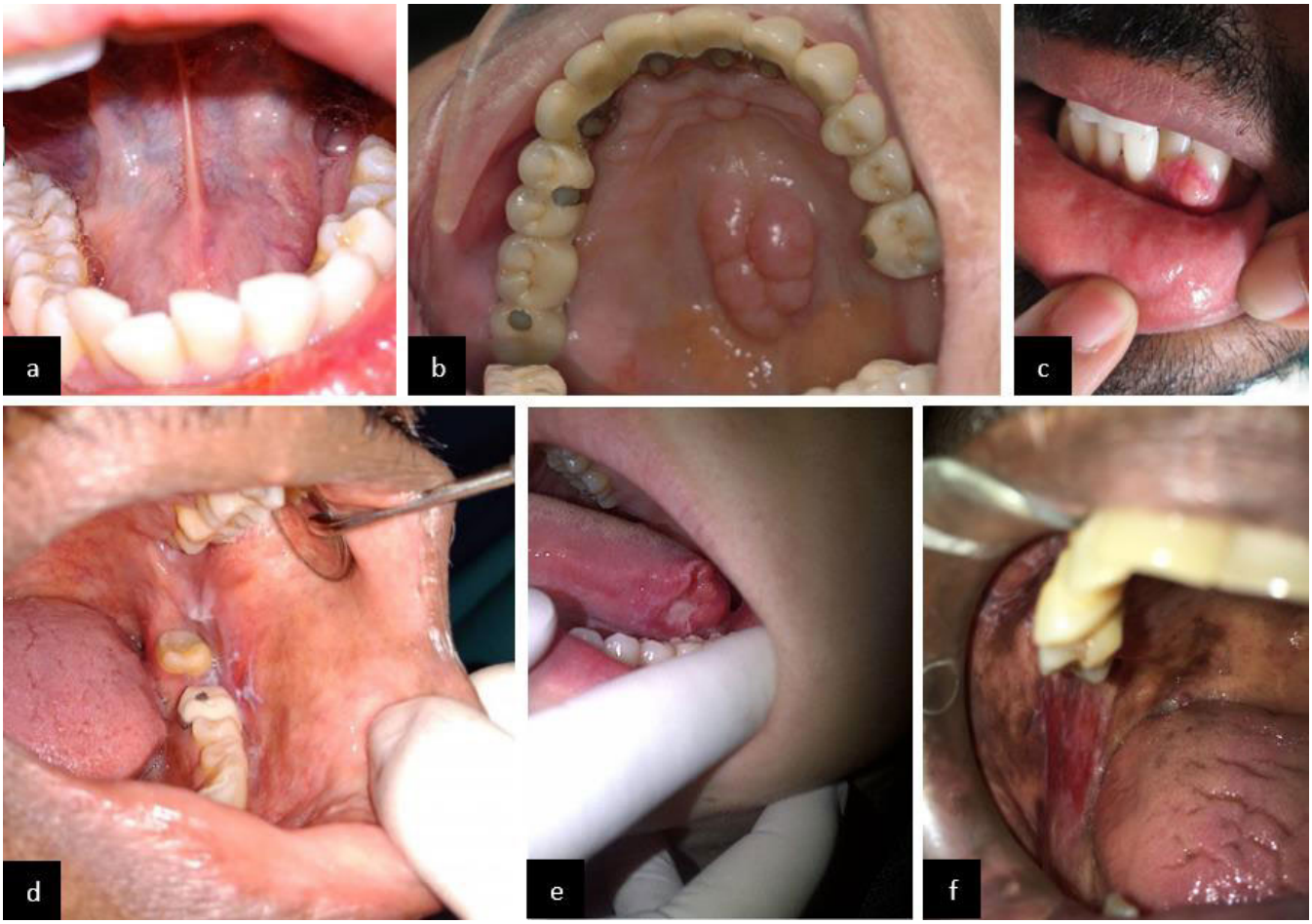


FIGURE 8. Results of multi-class image classification. (a)-(e) Correct classifications, (f) incorrect classification. Expected class derived from composite annotation. (a) Expected and predicted class = 'No lesion'. (b) Expected and predicted class = 'no referral needed'. (c) Expected and predicted class = 'refer for other reasons'. (d) Expected and predicted class = 'refer - low risk OPMD'. (e) Expected and predicted class = 'refer - cancer/high risk OPMD'. (f) Expected class = 'refer - low risk OPMD' and predicted class = 'refer - cancer/high risk OPMD.'

Another contribution of this paper is the novel application of deep learning based object detection to tackle oral lesion detection and classification for the early detection of oral cancer. Whilst a similar framework has been applied to detect cold sores and canker sores [20], we attempt the much more challenging task of detecting OPMDs and oral cancer. Object

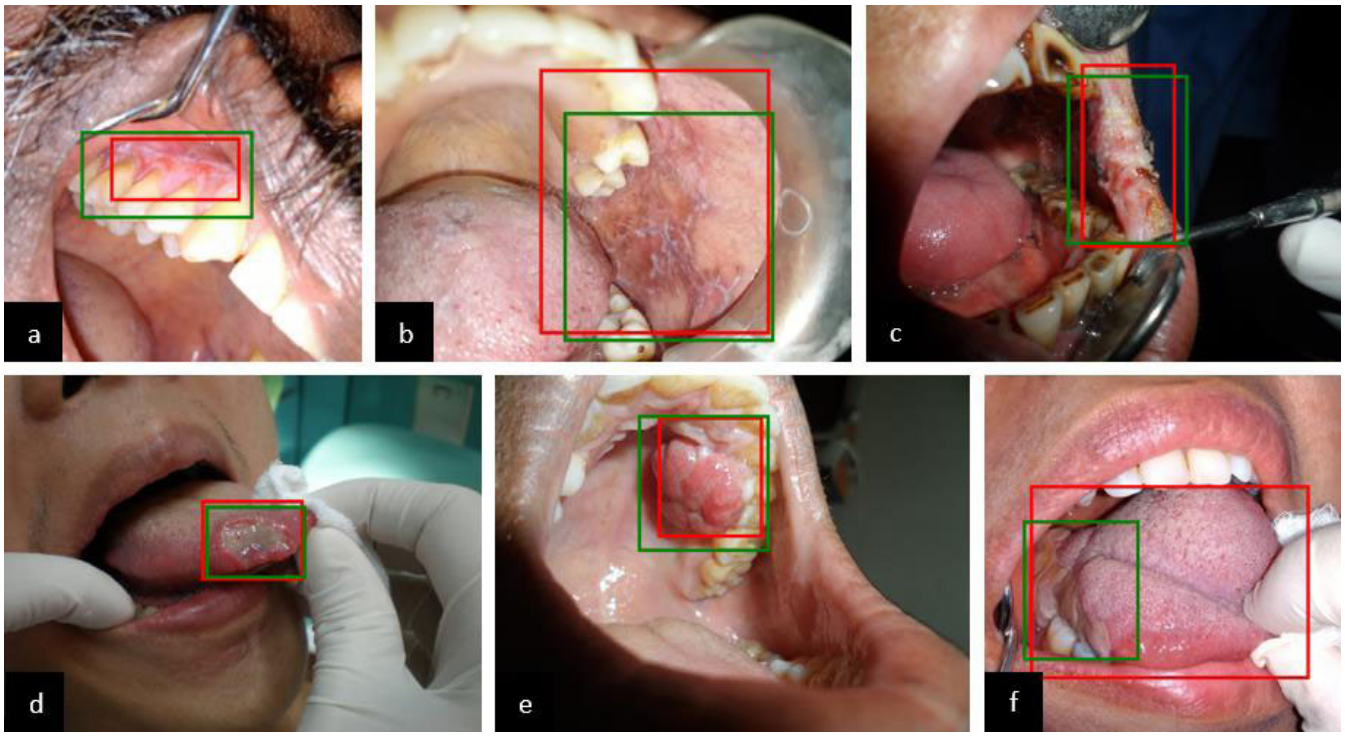


FIGURE 9. Results of four-class object detection. (a)-(d) Correction detections, (e)-(f) wrong and missed detections. The composite annotation bounding boxes are green and the predicted bounding boxes are red. (a) Green and red = 'refer - low risk OPMD', IoU = 0.52. (b) Green and red = 'refer - low risk OPMD', IoU = 0.71. (c) Green and red = 'refer - cancer/high risk OPMD', IoU = 0.72. (d) Green and red = 'refer - cancer/high risk OPMD', IoU = 0.83. (e) Green = 'refer for other reasons', red = 'refer - cancer/high risk OPMD', IoU = 0.70. (f) Green and red = 'no referral needed', IoU = 0.27.

TABLE 11. One-class object detection results.

Object class	TP	FP	FN	Precision (%)	Recall (%)	F_1 score (%)
Lesion	55	63	93	46.61	37.16	41.35

TABLE 12. Two-class object detection results.

Object class	TP	FP	FN	Precision (%)	Recall (%)	F_1 score (%)
No referral needed	8	23	38	25.81	17.39	20.78
Referral	56	114	46	32.94	54.90	41.18
Macro-averaged	-	-	-	29.37	36.15	32.41

TABLE 13. Four-class object detection results.

Object class	TP	FP	FN	Precision (%)	Recall (%)	F_1 score (%)
No referral needed	8	75	38	9.64	17.39	12.40
Refer for other reasons	15	60	21	20.00	41.67	27.03
Refer - low risk OPMD	18	50	23	26.47	43.90	33.03
Refer - cancer/high risk OPMD	14	81	11	14.74	56.00	23.33
Macro-averaged	-	-	-	17.71	39.74	24.50

detection is considered a much more challenging task than image classification, as locations of multiple objects have to be accurately attained. Failing to do so can heavily penalize performance scores, as is evident in Fig. 9(f). Our motivation for using object detection was that identifying the location allows the classification to be much more targeted to a specific region (just the lesion) and hence avoids redundancies that may be present in the image as a whole. Hence, this equates to a type of attention mechanism (terminology made popular in [41]), with the RPN telling the detection network

where to look. Although, we should not rule out the fact that global information provided from the whole image could be important to consider. Also, the final bounding boxes provides an insight into the model's decision making process, as opposed to just having an image based label from image classification.

Uthoff *et al.* [16] used a VGG CNN [42] and reports a sensitivity of 85.00% and a specificity of 88.75% for classifying pairs of autofluorescence and white light images as suspicious and not suspicious. Aubreville *et al.* [10] classified laserendomicroscopy images as clinically normal and carcinogenic, achieving a sensitivity of 86.6% and a specificity of 90.0%. Anantharaman *et al.* [20] used the dice coefficient to report performance of image segmentation of canker and cold scores and achieved a score of 0.744. These studies report good performances; however, they cannot be directly compared to the performance stated of this paper. Predominantly because different datasets present vastly different challenges. Our particular dataset was built by clinicians to represent the challenge in its true nature, demonstrating the variation of the oral disease presentations. The results of this paper do not currently offer a solution, but they are encouraging when we consider the scale of the problem. Image classification, which achieved F_1 scores of 87.07% and 78.30%, offers a more viable approach than object detection. The requirement from object detection to attain the accurate localization of lesions currently presents difficulties. Although both approaches warrant further exploration using a larger dataset.

With this study being performed using the first phase of data collection, we acknowledge that there were a number of limitations. The size of the dataset may be larger than the majority of cases presented in Table 1; however, it was still relatively small in the context of deep learning. Whilst transfer learning can quite successfully be applied to small datasets. Our dataset is currently problematic as it's extremely varied, not just because of the varied disease types, but also the varied presentation of each disease type. Further exemplified by the train, validation and test sets almost appearing to present different distributions despite being from the same distribution of data. This was clearly not practical when building a system. Large datasets are key to deep learning, so more data will improve results significantly, allowing complex patterns to be found whilst being generalizable. Another limitation relates to the approach we took to boost the size of the dataset. In addition to the 800 images annotated by multiple clinicians, a further 1355 images were annotated by a single clinician. Thus, this part of the data had not benefited from composite annotation and also may differ in its characteristics to the rest. This had the impact of making the data more unstable for training and also making the test data more difficult to perform well on. A final limitation to state was that the dataset contained some images with poor resolution. As the dataset grows, we will put constraints on what is acceptable in terms of image resolution; therefore, promoting high quality data.

In addition to working with a larger dataset, the future plan is to make use of the metadata to be used as input alongside the images. We also intend our models to output several of the other labels that the clinicians assigned, to gain the potential benefits in performance provided by multi-task learning. With the baseline models now in place, we will adapt the model architecture to better suit our task. We will also explore image classification with attention [43], [44]. The developed algorithm once incorporated into the MeMoSA® app will either use phone based or cloud based deployment. We intend to explore light-weight models to enable the former case, although real-time analysis is not necessary.

VI. CONCLUSION

This paper has discussed the collection and annotation of images from the oral cavity and demonstrated results for automating the early detection of oral cancer. The contribution of this paper is a novel strategy to combine bounding box annotations from multiple clinicians; followed by the assessment of two different deep learning based approaches to provide a solution to automation. Our promising initial results demonstrate the effectiveness of deep learning and suggest it has the potential to tackle this challenging task. Performances are set to increase as the dataset grows and this will have a significant impact in low- and middle-income countries where health resources are limited.

ACKNOWLEDGMENT

The authors are grateful to Dr. Kohgulakuhan Yogalingam for technical assistance at the initial stage of MeMoSA® Annotate development. They would also like to thank Matterport, Inc., for allowing access their source code.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[2] H. Gelband, P. Jha, R. Sankaranarayanan, and S. Horton, *Disease Control Priorities: Cancer*, vol. 3. Washington, DC, USA: World Bank, 2015.

[3] J. Rimal, A. Shrestha, I. K. Maharjan, S. Shrestha, and P. Shah, "Risk assessment of smokeless tobacco among oral precancer and cancer patients in eastern developmental region of nepal," *Asian Pacific J. Cancer Prevention*, vol. 20, no. 2, pp. 411–415, Feb. 2019.

[4] J. G. Doss, W. M. Thomson, B. K. Drummond, and R. J. R. Latifah, "Validity of the FACT-H&N (v 4.0) among Malaysian oral cancer patients," *Oral Oncol.*, vol. 47, no. 7, pp. 648–652, 2011.

[5] H. Amarasinghe, R. D. Jayasinghe, D. Dharmagunawardene, M. Attygalla, P. A. Scuffham, N. Johnson, and S. Kularatna, "Economic burden of managing oral cancer patients in sri lanka: A cross-sectional hospital - based costing study," *BMJ Open*, vol. 9, no. 7, Jul. 2019, Art. no. e027661.

[6] R. D. Jayasinghe, L. P. G. Sherminie, H. Amarasinghe, and M. A. Sitheeque, "Level of awareness of oral cancer and oral potentially malignant disorders among medical and dental undergraduates," *Ceylon Med. J.*, vol. 61, no. 2, p. 77, Jun. 2016.

[7] O. Kujan, A.-M. Glenny, R. Oliver, N. Thakker, and P. Sloan, "Screening programmes for the early detection and prevention of oral cancer," *Austral. Dental J.*, vol. 54, no. 2, pp. 170–172, Jun. 2009.

[8] N. Haron, R. B. Zain, W. M. Nabillah, A. Saleh, T. G. Kallarakkal, A. Ramanathan, S. H. M. Sinon, I. A. Razak, and S. C. Cheong, "Mobile phone imaging in low resource settings for early detection of oral cancer and concordance with clinical oral examination," *Telemed. e-Health*, vol. 23, no. 3, pp. 192–199, Mar. 2017.

[9] M. M. R. Krishnan, V. Venkatraghavan, U. R. Acharya, M. Pal, R. R. Paul, L. C. Min, A. K. Ray, J. Chatterjee, and C. Chakraborty, "Automated oral cancer identification using histopathological images: A hybrid feature extraction paradigm," *Micron*, vol. 43, nos. 2–3, pp. 352–364, Feb. 2012.

[10] M. Aubreville, C. Knipfer, N. Oetter, C. Jaremenko, E. Rodner, J. Denzler, C. Bohr, H. Neumann, F. Stelzle, and A. Maier, "Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning," *Sci. Rep.*, vol. 7, no. 1, p. 11979, Dec. 2017.

[11] J. Folmsbee, X. Liu, M. Brandwein-Weber, and S. Doyle, "Active deep learning: Improved training efficiency of convolutional neural networks for tissue classification in oral cavity cancer," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 770–773.

[12] R. K. Gupta, M. Kaur, and J. Manhas, "Tissue level based deep learning framework for early detection of dysplasia in oral squamous epithelium," *J. Multimedia Inf. Syst.*, vol. 6, no. 2, pp. 81–86, Jun. 2019.

[13] P. R. Jeyaraj and E. R. Samuel Nadar, "Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm," *J. Cancer Res. Clin. Oncol.*, vol. 145, no. 4, pp. 829–837, Apr. 2019.

[14] S. Xu, Y. Liu, W. Hu, C. Zhang, C. Liu, Y. Zong, S. Chen, Y. Lu, L. Yang, E. Y. K. Ng, Y. Wang, and Y. Wang, "An early diagnosis of oral cancer based on three-dimensional convolutional neural networks," *IEEE Access*, vol. 7, pp. 158603–158611, 2019.

[15] B. Song, S. Sunny, R. D. Uthoff, S. Patrick, A. Suresh, T. Kolur, G. Keerthi, A. Anbarani, P. Wilder-Smith, M. A. Kuriakose, P. Birur, J. J. Rodriguez, and R. Liang, "Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning," *Biomed. Opt. Express*, vol. 9, no. 11, pp. 5318–5329, 2018.

[16] R. D. Uthoff, B. Song, S. Sunny, S. Patrick, A. Suresh, T. Kolur, G. Keerthi, O. Spires, A. Anbarani, P. Wilder-Smith, M. A. Kuriakose, P. Birur, and R. Liang, "Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device with neural network classification for low-resource communities," *PLoS ONE*, vol. 13, no. 12, Dec. 2018, Art. no. e0207493.

[17] A. Rana, G. Yauney, L. C. Wong, O. Gupta, A. Muftu, and P. Shah, "Automated segmentation of gingival diseases from oral images," in *Proc. IEEE Healthcare Innov. Point Care Technol. (HI-POCT)*, Nov. 2017, pp. 144–147.

[18] B. Thomas, V. Kumar, and S. Saini, "Texture analysis based segmentation and classification of oral cancer lesions in color images using ANN," in *Proc. IEEE Int. Conf. Signal Process., Comput. Control (ISPCC)*, Sep. 2013, pp. 1–5.

[19] R. Anantharaman, V. Anantharaman, and Y. Lee, "Oro vision: Deep learning for classifying orofacial diseases," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Aug. 2017, pp. 39–45.

[20] R. Anantharaman, M. Velazquez, and Y. Lee, "Utilizing mask R-CNN for detection and segmentation of oral diseases," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 2197–2204.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2012, pp. 1097–1105.

[23] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[26] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[30] W. Liu, D. Anguelov, D. Erhan, and C. Szegedy, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[31] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, "Automatic colon polyp detection using region based deep CNN and post learning approaches," *IEEE Access*, vol. 6, pp. 40950–40962, 2018.

[32] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Sci. Rep.*, vol. 8, no. 1, p. 4165, Dec. 2018.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] N. Haron, R. B. Zain, and A. Ramanathan, "m-Health for early detection of oral cancer in low-and middle-income countries," *Telemed. e-Health*, vol.26, no. 3, pp. 278–285, 2019.

[35] V. C. Raykar, S. Yu, L. H. Zhao, and G. H. Valadez, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, Apr. 2010.

[36] X. Li, B. Aldridge, J. Rees, and R. Fisher, "Estimating the ground truth from multiple individual segmentations with application to skin lesion segmentation," in *Proc. Med. Image Understand. Anal. Conf.*, vol. 1, London, U.K., 2010, pp. 101–106.

[37] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[38] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[39] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/Accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7310–7311.

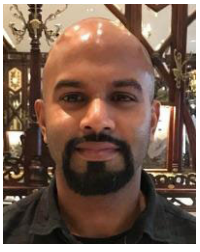
[40] W. Abdulla, "Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow," Matterport, Inc., Sunnyvale, CA, USA, Tech. Rep., 2017.

[41] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[43] K. Xu, J. Ba, R. Kiros, K. Cho, and A. Courville, "Show, attend and tell: Neural image caption generation with visual attention," *Comput. Sci.*, vol. 2015, pp. 2048–2057, Feb. 2015.

[44] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.



ROSHAN ALEX WELIKALA received the B.Sc. degree in imaging science from the University of Westminster, U.K., the M.Sc. degree in medical image computing from the University College London, U.K., and the Ph.D. degree in medical image analysis from Kingston University, U.K. He is currently a Postdoctoral Research Associate in computer vision with Kingston University. His research interests include machine learning and medical imaging.

PAOLO REMAGNINO (Senior Member, IEEE) received the Laurea (M.Res.) degree in electronic engineering from the University of Genoa, Italy, and the Ph.D. degree in computer vision from the University of Surrey, U.K. He is currently a Professor of computer vision and leads the Robot Vision Team (ROVIT), Kingston University, U.K. His research interests include image and video understanding, computational botany, and distributed intelligence.

JIAN HAN LIM (Graduate Student Member, IEEE) received the B.Sc. degree in software engineering from the Tunku Abdul Rahman University College, Malaysia. He is currently pursuing the Ph.D. degree in artificial intelligence with the University of Malaya, Malaysia. His research interests include computer vision and deep learning with a focus on image captioning.

CHEE SENG CHAN (Senior Member, IEEE) received the B.Eng. degree in electronics engineering from Multimedia University, Malaysia, and the M.Sc. degree in communications systems engineering and the Ph.D. degree in artificial intelligence from the University of Portsmouth, U.K. He is currently an Associate Professor of artificial intelligence with the University of Malaya, Malaysia. His research interests include computer vision and machine learning with a focus on scene understanding.

SENTHILMANI RAJENDRAN received the B.D.S. degree in dental surgery from Vinayaka Mission's Research Foundation University, India, and the M.Sc. degree in health informatics from the University of Leeds, U.K. She is currently a Research Associate with the mHealth for Oral Cancer, Cancer Research Malaysia. Her research interests include digital health and health systems research.

THOMAS GEORGE KALLARAKKAL received the B.D.S. degree in dental surgery from The Tamil Nadu Dr. M. G. R. Medical University, India, and the M.D.S. degree in oral pathology from the University of Kerala, India. He is currently an Associate Professor with the University of Malaya, Malaysia. His main research interest includes oral cancer.

ROSNAH BINTI ZAIN received the B.D.S. degree in dental surgery from The University of Queensland, Australia, and the M.Sc. degree in oral pathology and diagnosis from the University of Michigan, USA. She is currently a Professor of oral pathology and oral medicine with MAHSA University, Malaysia, and an Honorary Professor with the University of Malaya, Malaysia. Her research interests include oral mucosal lesions and oral cancer. She received the Fellowship from the American Academy of Oral Pathology.

RUWAN DUMINDA JAYASINGHE received the B.D.S. degree in dental surgery from the University of Peradeniya, Sri Lanka, and the M.S. degree in oral surgery from the University of Colombo, Sri Lanka. He is currently a Professor of oral medicine and periodontology with the University of Peradeniya. His research interests include oral potentially malignant disorders, oral cancer, and smokeless tobacco and areca nut cessation.

JYOTSNA RIMAL received the B.D.S. degree in dental surgery from Manipal University, India, and the M.D.S. degree from Kathmandu University, Nepal. She is currently a Professor and the Head of the Department of Oral Medicine and Radiology, B. P. Koirala Institute of Health Sciences, Nepal. Her research interests include oral cancer, oral potentially malignant disorders, orofacial pain, digital dentistry, and dental education. She received the Fellowship from the Foundation for Advancement of International Medical Education and Research (FAIMER).

ALEXANDER ROSS KERR received the B.Sc. degree in life sciences from Queen's University at Kingston, Canada, the M.S.D. degree in dentistry from the University of Washington, Seattle, WA, USA, and the D.D.S. degree in dental surgery from McGill University, Canada. He is currently a Clinical Professor with the New York University College of Dentistry, USA, where he is the Director of the Oral Mucosal Disease Service. His research interests include oral cancer and premalignant disorders.

RAHMI AMTHA received the M.D.S. degree in oral medicine from the University of Malaya, Malaysia, and the Ph.D. degree from the Cancer Research Malaysia, University of Malaya. She is currently an Oral Medicine Specialist with Trisakti University, Indonesia. Her main research interest includes oral cancer.

KARTHIKEYA PATIL is currently a Professor and the Head of the Oral Medicine and Radiology Department, Jagadguru Sri Shivarathreeshwara University, India. His main research interest includes oral cancer. He is a member of the Indian Academy of Oral Medicine and Radiology and the Indian Dental Association.

WANNINAYAKE MUDIYANSELAGE TILAKARATNE received the B.D.S. degree in dental surgery from the Medical School Peradeniya, Sri Lanka, the M.S. degree from the University of Colombo, Sri Lanka, and the Ph.D. degree from Niigata University, Japan. He is currently the Dean and a Professor of oral pathology with the University of Peradeniya, Sri Lanka. His research interests include oral cancer and oral submucous fibrosis. He received the FDSRCS from the Royal College of Surgeons of England, U.K., and the FRCPath from the Royal College of Pathologists, U.K.

JOHN GIBSON graduated in medicine and dentistry from the University of Glasgow, U.K. He received the Ph.D. degree from the University of Glasgow, on the condition known as Orofacial Granulomatosis and undertook his specialist clinical training in oral medicine. He is currently a Professor of oral medicine and the Director of dentistry with the Institute of Dentistry, University of Aberdeen, U.K. His main research interest includes oral cancer.

SOK CHING CHEONG received the B.Sc. degree in biochemistry and molecular biology and the Ph.D. degree in molecular biology from The National University of Malaysia. She is currently a Senior Group Leader with the Cancer Research Malaysia and an Adjunct Professor with the University of Malaya, Malaysia. Her research interests include oral cancer and oral potentially malignant disorders.

SARAH ANN BARMAN received the B.Sc. degree in physics from the University of Essex, U.K., the M.Sc. degree in applied optics from Imperial College London, U.K., and the Ph.D. degree in optical physics from King's College London, U.K. She is currently a Professor of computer vision with Kingston University, U.K. Her research interest includes the development of computer vision techniques applied to medical images.