DETECTION OF ORAL CANCER FROM CLINICAL IMAGES USING DEEP LEARNING

A Thesis
Submitted to
the Temple University Graduate Board

In Partial Fulfillment of the Requirements for the Degree
MASTER OF SCIENCE

by
Anusha Solanki
May 2024

Examining Committee members:
Chukwuebuka Ogwo, Thesis advisor, Oral Health Sciences
Riya Kuklani, Clinical Pathology and Laboratory medicine
Louis DiPede, Restorative Dentistry
ABSTRACT

**Objectives:** To detect and distinguish oral malignant and non-malignant lesions from clinical photographs using YOLO v8 deep learning algorithm.

**Methods:** This is a diagnostic study conducted using clinical images of oral cavity lesions. The 427 clinical images of the oral cavity were extracted from a publicly available dataset repository specifically Kaggle and Mendeley data repositories. The datasets obtained were then categorized into normal, abnormal (non-malignant), and malignant oral lesions by two independent oral pathologists using Roboflow Annotation Software. The images collected were first set to a resolution of 640 x 640 pixels and then randomly split into 3 sets: training, validation, and testing – 70:20:10, respectively. Finally, the image classification analysis was performed using the YOLO V8 classification algorithm at 20 epochs to classify and distinguish between malignant lesions, non-malignant lesions, and normal tissue. The performance of the algorithm was assessed using the following parameters accuracy, precision, sensitivity, and specificity.

**Results:** After training and validation with 20 epochs, our oral cancer image classification algorithm showed maximum performance at 15 epochs. Based on the generated normalized confusion matrix, the sensitivity of our algorithm in classifying normal images, non-malignant images, and malignant images was 71%, 47%, and 54%, respectively. The specificity of our algorithm in classifying normal images, non-malignant, and malignant images were 86%, 65%, and 72%. The precision of our algorithm in classifying normal images, non-malignant images, and malignant images was 73%, 62%, and 35%, respectively. The overall accuracy of our oral cancer image classification algorithm was
55%. On a test set, our algorithm gave an overall 96% accuracy in detecting malignant lesions.

**Conclusion:** Our object classification algorithm showed a promising application in distinguishing between malignant, non-malignant, and normal tissue. Further studies and continued research will observe increasing emphasis on the use of artificial intelligence to enhance understanding of early detection of oral cancer and pre-cancerous lesions.

**Keywords:** Normal, Non-malignant, Malignant lesions, Image classification, Roboflow annotation software, YOLO v8 object/image classification algorithm.
ACKNOWLEDGEMENTS

I would like to thank my main committee members Drs. Ogwo, Kuklani, DiPede and additionally Drs. Gangwani and Barnts for helping me all the way during all stages of this thesis. Thank you for your guidance, dedication, and time. I would also like to thank my family and close friends for your steady support and encouragement throughout the course of my project.
# TABLE OF CONTENTS

ABSTRACT ........................................................................................................................ ii

ACKNOWLEDGMENTS ................................................................................................. iv

LIST OF TABLES ............................................................................................................ vii

LIST OF FIGURES ......................................................................................................... viii

CHAPTER

1. INTRODUCTION .........................................................................................................1

2. REVIEW OF LITERATURE ........................................................................................4

   2.1 Introduction .......................................................................................................4

   2.2 Section of Individual articles .............................................................................5

   2.3 Integration of methods and results ...................................................................11

   2.4 Discussion ........................................................................................................16

   2.5 Conclusion .......................................................................................................19

   2.6 Gaps in future directions and needed future directions  ..................................19

3. METHODS ..................................................................................................................21

   3.1 Recruitment method .......................................................................................21

   3.2 Study procedures and data analysis ................................................................22

4. RESULTS ......................................................................................................................24

5. DISCUSSION ................................................................................................................29

   5.1 Data source and quality ..................................................................................29
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Eligibility criteria</td>
<td>22</td>
</tr>
<tr>
<td>2. Image classification algorithm outputs for each category</td>
<td>27</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A train/loss function graph</td>
<td>24</td>
</tr>
<tr>
<td>2. A validation/loss function graph</td>
<td>25</td>
</tr>
<tr>
<td>3. A metrics/accuracy_top1 function graph</td>
<td>25</td>
</tr>
<tr>
<td>4. A metrics/accuracy_top 5 function graph</td>
<td>26</td>
</tr>
<tr>
<td>5. Confusion matrix normalized</td>
<td>28</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

Cancers of the oral cavity and pharynx account for 3% of cancers diagnosed in the United States each year. Cancers at these sites can differ anatomically and histologically and might have different causal factors, such as tobacco use, alcohol use, and infection with human papillomavirus (HPV). U.S. Cancer Statistics data were analyzed to examine trends in the incidence of cancers of the oral cavity and pharynx by anatomic site, sex, race/ethnicity, and age group. During 2007–2016, incidence rates increased for cancers of the oral cavity and pharynx combined, base of the tongue, anterior tongue, gum, tonsil, oropharynx, and other oral cavity and pharynx. Incidence rates declined for cancers of the lip, floor of the mouth, soft palate and uvula, hard palate, hypopharynx, and nasopharynx, and were stable for cancers of the cheek and other mouth and salivary gland. The American Cancer Society’s most recent estimates for oral cavity and oropharyngeal cancers in the United States for 2023 were about 54,540 new cases of oral cavity or oropharyngeal cancer and around 11,580 deaths from oral cavity or oropharyngeal cancer. Ongoing implementation of proven population-based strategies to prevent tobacco use initiation, promote smoking cessation, reduce excessive alcohol use, and increase HPV vaccination rates might help prevent cancers of the oral cavity and pharynx. (Ellington et al 2020)

Warnakulasuriya and Johnson (1993) stated that 95% of oral cancers are squamous cell carcinoma. To reduce the occurrence of oral squamous cell carcinoma at an early stage and improve the quality of life for the patients, it becomes necessary to perform early screenings for such patients. Also, premalignant lesions have a 1% chance of turning into
malignancy, so, premalignant lesions are also harmful, if not diagnosed correctly. In the USA, more than half of the people are diagnosed with oral cancer at late stages resulting in a 5-year survival rate of about 52%. For early detection of oral cancer, screening of stages of OSCC leads to improved survival rates. With advances in early detection, nearly half of the patients were detected with either stage 1 or stage 2 cancer. There is a chance of around 4% of stage 1 and 2 cancers to develop with second primary cancers yearly. (Hung M et al 2020)

Artificial intelligence (AI) and machine learning (ML) are terms that are often used in research and are interchangeable even though they have different meanings. John McCarthy called the father of artificial intelligence, coined the term ‘artificial intelligence’ to describe machines with the potential to perform actions that were considered intelligent without any human intervention. Machine learning (ML) is a subset of artificial intelligence. Simon Cowell coined this term in 1959 (Patil et al, 2022). ML predicts the outcome based on the dataset provided to it using algorithms, such as artificial neural networks (ANN). These networks mimic the human brain and have interconnected artificial neurons that receive and analyze data signals. Convolutional neural network (CNN) or deep learning (DL) is an approach in ML introduced in 2006 by Hinton et al. It utilizes multi-layer neural networks to compute data. Deep learning algorithms have the potential to analyze patterns based on the data and improve the outcome. Recently, artificial intelligence in dentistry alone has created immense attention in specialties such as orthodontics, endodontics, prosthodontics, restorative dentistry, periodontics, and oral and maxillofacial surgery. Research reveals promising results, although most applications are in the developmental phase. It becomes a necessity that dentists need to understand the
foundational concepts and applications of AI in dentistry to adapt to a changing healthcare landscape. (Patil et al, 2022)

The development of object detection algorithms in recent years has introduced computer-aided diagnosis systems to allow clinicians to interpret and analyze information more comprehensively in a shorter amount of time. With this, we can also help patients who are still having difficulty getting the dental care that they need with more accessible services, like telemedicine and teledentistry.
CHAPTER 2
LITERATURE REVIEW

2.1 Introduction:

Oral cancer is the 6th most common cancer worldwide with high mortality rates because most of oral cancer cases are detected at a very late stage. According to the International Agency for Research and Cancer, there were an estimated 377,000 new cases of lip and oral cavity cancers in 2020, with nearly 177,000 deaths worldwide. The overall 5-year survival rate of diagnosed oral cancer patients is around 50% and has varied by race and area. The survival rates have been as high as 65% in developed countries and as low as 15% in rural areas (Lin et al, 2021). Oral squamous cell carcinoma (OSCC) which makes up to 90% of oral cancer cases, is often preceded by oral potentially malignant disorders (OPMDs), such as leukoplakia, lichen planus, and erythroplakia. The detection of OPMD has the risk of malignant transformation and is very important to detect in order to reduce the morbidity and mortality cases of oral cancer. Besides, early-stage OSCC lesions and OPMD are typically asymptomatic and may appear as small, harmless lesions, leading to late presentation of patients and ultimately leading to further diagnostic delay. Advances in the fields of computer vision and deep learning offer powerful methods to develop adjunctive technologies that can perform an automated screening of the oral cavity and provide feedback to healthcare professionals during patient examinations as well as to individuals with the examination (Tanriver et al, 2021). Machine learning in artificial intelligence shows significant development and acceptance among researchers (Dinesh Y et al, 2023). Studies also show that deep learning algorithms can surpass
the performance of human experts in many disease recognition scenarios (Lin et al., 2021). This section reviews three articles on artificial intelligence (Lin et al., Dinesh et al and Tanriver et al) for early detection of oral cancer using intraoral clinical images.

2.2 Section of individual articles:

Lin et al (2021) conducted a retrospective study to determine oral cancer detection using a smartphone-based white light inspection method. It was approved by the Medical Research Ethics Committee of the First Affiliated Hospital, College of Medicine, Zhejiang University, China (Lin et al., 2021).

Image collection was carried out in daily outpatient clinics and inpatients in the Hospital (Lin et al., 2021). Four different smartphones were used which were iPhone 11, iPhone 12, IQOOUI, 360N7Pro to retrospectively collect oral images from subjects over 18 years of age, including healthy people, aphthous ulcer patients, OPMD, and oral cancer. Loss of homogeneity is a key visual feature indicating potential carcinogenesis, low-risk (homogeneous) and high-risk (non-homogeneous) were categorized based on clinical manifestations. All the oral images were divided into 5 categories namely normal, aphthous ulcers, low-risk OPMD, high-risk OPMD, and oral cancer. The images were annotated by three oral medicine specialists and all the controversial cases were excluded from the dataset. The final oral dataset consisted of 688 oral lesions and 760 normal mucosa images. The 688 oral lesions included images of aphthous ulcers (251), low-risk OPMD (231), high-risk OPMD (141), and cancer (65). The dataset was divided into two subsets: training set (993 cases) and testing set (455 cases). The test set consisted of 288 normal (30%), 76 aphthous ulcers (30%), 69 low-risk OPMDs (30%), 52 high-risk OPMDs (37%), and
30 cancers (46%). Data processing was done by CNN based system which took one image at a time to output the probability of oral disease. The processing steps included: 1. Cropping, 2. Centering, 3. Image normalization, 4. Resizing the image to 512 x 512 pixels (Lin et al, 2021).

Image classification was done using high-resolution learning network (HRNet), which was pre-trained on ImageNet. HR-Net W18 is a Pytorch-based algorithm running on the Ubuntu 18.04 operating system. HR-Net was trained under the supervised learning method that minimizes the error between predicted probability and true class labels. For speeding up the training and improving training performance, preinitialization from the Image Net database was used. For multilabel classification, macro averages of sensitivity, specificity, precision, and F1 metrics were used to evaluate the performance of the models. Neural network architectures like VGG16, Res Net50, Dense Net169, and HR-Net W18 were compared with each other on multiclass classification. The HR-Net was fine-tuned to analyze the images and output a relative diagnosis for the five categories mentioned. Image rotation and localizing the image center method was performed to get a training sample of 1596 normals, 1575 aphthous ulcers, 1458 low-risk OPMDs, 1335 high-risk OPMDs and 1575 cancers (Lin et al, 2021).

Lin et al (2021) derived the results by using HR-Net W18 CNN-based network. A comparison was made between the images with centered lesions and the images without any prerequired centered lesions. The performance of the proposed method (center positioning + resampling) achieved a sensitivity of 83%, specificity of 96.6%, precision of 84.3%, and F1 of 83.6% on the 455 cases of test images. The F1 score
for the method was about 8% higher than the random positioning method. Also, the resampling method had an additional 6% performance improvement. Various CNN-based networks like VGG-16, ResNet 50, Dense Net 169, and HR-Net W18 were compared. HR-Net W18 performed slightly better than other models with a sensitivity of 83%, specificity of 96.6%, precision of 84.3%, F1 score of 83.6%. Lin et al (2021) showed some failure cases with high-risk OPMD cases wrongly identified as an ulcer with a 1.9% of the rate. Also, high-risk OPMD was misdiagnosed as cancer in many cases.

In conclusion, Lin et al (2021) concluded that automatic identification/detection of oral diseases was effective using a smartphone-based image approach.

Dinesh Y et al (2023), study was conducted in Saveetha Dental College and Hospitals, Chennai, India. The photographic intraoral images were collected by using a mobile camera or iPad with a resolution ranging from 4MP to 16 MP. The data utilized were clinical intra-oral images (n=360) of OSCC, OPMDs, and non-pathological oral images. For training n=300 images and for testing n=60 images were used. All the images used in the study were histopathologically confirmed for clinically normal, OPMDs and OSCCs. The study focused on analyzing images of OSCCs and OPMDs presented as white and red lesions in the form of patches, ulcers, striae, plaques, etc. by clinicians and machine learning. The test set n=60 was blinded by two expert oral and maxillofacial pathologists; they analyzed the images to categorize them as normal mucosa and suspected lesions. Roboflow software was used to classify and annotate the images with multiclass annotation and object
detection models. Mean average precision (mAP), precision, recall, sensitivity, and specificity were calculated.

Dinesh Y et al (2023) derived the results by using Kappa Statistics with SPSS software v23.0 (IBM Corp, Armonk, USA) with the test set of n=60 images. The images were analyzed by two experts using the images and compared with the results of Roboflow software. The Mean Average precision (mAP) of 25.4%, precision of 29.8%, and recall of 32.9%. When a comparison was made between the subject experts and Roboflow software, the sensitivity of Roboflow software was 88.9% and the specificity of 75%. The true positive results of clinically healthy mucosa by subject experts were 100% and by Roboflow was 80%. The true positive results of suspected lesions (OPMDs and OSCCs) by the subject expert were 90% and the Roboflow model was at 85%. The Statistical analysis derived a 0.7 value showing a moderate agreement between the blinded subject experts and the Roboflow model being almost perfect.

In conclusion, Dinesh Y et al (2023) concluded that there was a potential for automatic detection of OPMDs and OSCC patients using machine learning algorithms like Roboflow software.

Tanriver et al (2021), this retrospective study was conducted in collaboration with the Oncology Institute at Istanbul University and approved by the Ethics Committee at Istanbul University. Photographic images of oral lesions were collected from the Department of Tumour Pathology. The rest of the images were retrieved from Google and Yandex on 27th August 2020. Each lesion was classified as benign, OPMD, and carcinoma. Lesion instances were annotated by an expert oral pathologist using the
VGG Image Annotator (VIA) tool. The final dataset contained 652 images which were split into approximately 80% training, 10% validation, and 10% testing set. A two-stage model was proposed to detect oral lesions with a detector network and classify the detected regions into three categories (benign, OPMD, OSCC) and a second-stage classifier network. A semantic segmentation experiment was used as an image detection network. Instance segmentation, and object detection experiments did both image detection and classifier functions. Classification experiments were performed for the lesion classification task.

Tanriver et al (2021) used U-net architecture, with Qubvel’s segmentation library, built on the PyTorch framework (v1.7). Semantic segmentation experiments were performed. Data augmentation and transfer learning were utilized for implementation to overcome the overfitting of the models. Test-time augmentation (TTA) was also implemented to improve predictions at test time. The performance of the model was measured using a dice coefficient score (F1 measure). Instance segmentation experiments were performed to differentiate between different lesion instances in an image. Mask R-CNN, developed by Facebook AI Research (Facebook Detectron2 library), a well-known instance segmentation framework was used. It predicts a bounding box, a class label, and a pixel-level mask for each object instance separately. Average Precision (AP) was calculated. Object detection experiments were performed in order to combine localization and classification in a single network. YOLO v5 (PyTorch implementation of YOLOv4) was available with four different versions YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x. All four models were evaluated for the lesion detection task and were initialized with pre-trained weights.
on the dataset to prevent overfitting. Mosaic data augmentation was utilized by combining up to four images into one during training to improve the detection of small-scale objects. Average Precision (AP) was calculated. Classification experiments were performed with various CNN-based architectures like ResNet152, DEnseNet161, Inceptionv4, and EfficientNetb4 models were compared. Performance metrics like precision, recall, and F1 score were calculated.

Tanriver et al (2021) derived the results for a semantic segmentation experiment with various backbones like Efficient Netb3, DenseNet-161, Inception-v4, EfficientNetb7, ResNeXt-101_32x8d evaluated dice score with and without TTA. Among them, the Efficient Net b7 model achieved the highest dice score with 92.6% and a dice score with TTA of 92.9%. For instance, segmentation experiments, ResNet-50, ResNet-101, and ResNeXt-101 were compared with and without TTA. Among them, the ResNeXt-101 model achieved an AP score of 43.90% for box detection and 37.85% for mask detection respectively. For the object detection experiment, different versions of YOLOv5 were evaluated in one class lesion detection task. YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5l+TTA, YOLOv5x, YOLOv5x+TTA, YOLOv5s & 5m ensemble were compared. Among them, the YOLOv5s & 5m ensemble achieved the best performance with an AP of 63.7%. For the classification experiment, various CNN models like efficient Net-b4, Inception-v4, DenseNet-161, Ensemble, and ResNet-152 were compared. Among them, Efficient Net-b4 performed well with the precision of 86.9%, Recall of 85.5%, and F1 score of 85.8%. According to different classes like benign. For OPMD and carcinoma, the precision for the carcinoma class was 100%, the recall was 82% and
F1 score was 86% respectively. Tanriver et al (2021) concluded that segmentation and/or object detection of lesion areas is an essential step for the identification of oral lesions.

In conclusion, Tanriver et al (2021) concluded that there was a potential for artificial intelligence in the segmentation, detection, and multiclass classification of oral lesions with benign, OPMD, and OSCC categories.

2.3 Integration of methods and results:

All three studies assessed oral disease detection from the lesion images. The sources and the number of images varied substantially. These studies were generally similar because all were cross-sectional retrospective studies used to train the machine learning algorithm. Lin et al (2021) used a smartphone-based image approach which was approved by the Medical Research Ethics Committee of the First Affiliated Hospital, College of Medicine, Zhejiang University, China. Dinesh Y et al (2023) collected retrospective photographic intraoral images from the Department of Oral Pathology, Saveetha Dental College taken by using a mobile camera or iPad with resolution ranging from 4MP to 16MP. The data was retrieved from Dental Information Archived Software (DIAS) from January 2021 to March 2023. And Tanriver et al (2021) used a dataset from the Hospitals and Oncology Institute at Istanbul University from the Department of Tumour Pathology. Also, Tanriver et al (2021) collected the rest of the images from publicly available search engines like Google and Yandex (retrieved on 27th August 2020).

Lin et al (2021) categorized the collected intraoral images into 5 different categories namely normal, aphthous ulcers, low-risk OPMD, high-risk OPMD, and
oral cancer. Whereas Dinesh Y et al (2023) categorized the collected images into 3 different categories namely normal mucosa, OPMD, and OSCC. And Tanriver et al (2021) categorized the collected images into 3 categories namely benign, OPMD and carcinoma.

Lin et al (2021) evaluated the intraoral images by inviting three oral medicine specialists to manually annotating the data. Whereas Dinesh Y et al (2023) annotated the collected intraoral images with the help of two oral and maxillofacial pathologists. Tanriver et al (2021) annotated the collected intraoral images by an expert oral pathologist using the VGG Image Annotator (VIA) tool.

Lin et al (2021) had a final oral dataset consisting of 688 oral lesion images and 760 normal mucosa images. The 688 lesions images included images of aphthous ulcers (251), low-risk OPMD (231), high-risk OPMD (141), and cancer (655). The dataset was divided into two subsets: training set (993 cases) and testing set (455 cases). The test set consisted of 288 normals, 76 aphthous ulcers, 69 low-risk OPMD, 52 high-risk OPMD, and 30 cancers. Whereas Dinesh Y et al (2023) the data utilized were clinical intra-oral images (n=360) of OSCC, OPMDs, and non-pathological oral images from different areas of the oral cavity. About (n=300) of the images were used for training and (n=60) for testing purposes. And Tanriver et al (2021) had the final dataset for the object detection experiments consisting of 652 images, which were split into approximately 80% for training, 10% for validation, and 10% for testing. This was divided into a training set of 552 images, a validation set of 63 images, and a testing set of 69 images.
Lin et al (2021) performed the data processing by CNN-based system which took one image at a time to output the probability of oral disease. The processing steps included: 1. Cropping, 2. Centering, 3. Image normalization, 4. Resizing the image to 512 x 512 pixels. Whereas Dinesh Y et al (2023) did not transform or edit the collected intraoral images. The images in the study varied in resolution with the largest image measuring 4501 x 2986 pixels and the smallest image measuring 1120 x 821 pixels. Tanriver et al (2021) cropped and resized the images from 546 by 397 pixels to 512 by 512 pixels.

Lin et al (2021) also performed image resampling to alleviate the effect of variability in images. Image rotation and capturing the lesion center methods were used to relocate and the generated samples obtained were 1596 normals, 1575 aphthous ulcers, 1458 low-risk OPMDs, 1335 high-risk OPMDs, and 1575 cancers. Whereas Dinesh Y et al (2023) and Tanriver et al (2021) did not perform image resampling methods.

Lin et al (2021) treated the oral disease diagnosis as a multiclass classification. When compared with other networks like ResNet50, VGG16, DenseNet169, HR-Net W18 performed slightly better. So, the HR-Net W18-based network was used for further analysis. For multilabel classification, the macro averages of sensitivity, specificity, precision, and F1 metrics were calculated to achieve the best-performing model. Whereas Dinesh Y et al (2023) did not use any CNN-based network for multiclass classification. Tanriver et al (2021) performed classification experiments with various Convolutional Neural Network (CNN) architectures like ResNet-152, Dense-Net 161, Inception-v4, and EfficientNet-b4 and a comparison was made.
Performance metrics like precision, recall, and F-1 scores were calculated. Confusion matrices and related metrics were calculated using the scikit-learn library.

Lin et al (2021) demonstrated the effectiveness of the image-capturing method by comparing four different methods random positioning, center positioning, center positioning + over-sampling, and center positioning + resampling method. Among them, the center positioning + resampling method performed the best. The results showed that the centered ruled dataset achieved an 8% higher F1 score than that trained on the random positioning dataset. For the center positioning + resampling method the macroaverages were sensitivity of 83%, specificity of 96.6%, precision of 84.3%, and F1 score of 83.6% respectively. Various CNN networks like VGG-16, ResNet 50, and DenseNet-169 were compared in which HR-Net W18 performed the best. HR-Net W18 had values like the sensitivity of 83%, specificity of 96.6%, the precision of 84.3%, and F1 score of 83.6% respectively. A confusion matrix was generated for the HR-Net W 18 model to the test set and the results showed that the F1 score was 95.0% on the normal class when compared with the lower F1 scores for other classes, this concluded that classes with scarce samples are generally more problematic to classify. For various classes the AUC and F1 scores were for normal class AUC 94.9%, and F1 score of 95.1% respectively. Dinesh Y et al (2023) did not perform any classification network-based model detection. Tanriver et al (2021) performed a semantic segmentation experiment for image recognition tasks that deals with assigning each pixel of an image to a particular class including the background. Various backbones from U-net architecture were used. Various backbones like Efficient Netb3, DenseNet-161, Inception-v4, EfficientNetb7, and ResNeXt-
101_32x8d evaluated dice scores with and without TTA. Among them, the Efficient Net b7 model achieved the highest dice score with 92.6% and a dice score with a TTA of 92.9%. Also, classification experiments were performed with various CNN models like efficient Net-b4, Inception-v4, DenseNet-161, Ensemble, and ResNet-152 were compared. Among them, Efficient Net-b4 performed well with a precision of 86.9%, Recall of 85.5%, and F1 score of 85.8%. According to different classes like benign. For OPMD and carcinoma, the precision for the carcinoma class was 100%, the recall was 82% and the F1 score was 86% respectively.

Lin et al (2021) also introduced a class activation mapping technique on HR-Net W18 for prediction visualization. First, the last layer was removed and then upsampled to the original image size. The diagnostic multiclass performance of the method for all the five categories was compared. Normal achieved the best performance with an AUC of 94.9% and an F1 score of 95.1% respectively. Dinesh Y et al (2023) performed the statistical analyses were done using Kappa statistics with SPSS software v23.0 (IBM Corp, Armonk, USA). The specificity and sensitivity of the Roboflow Machine learning model were 75% and 88.9% respectively. The true positive results of clinically healthy oral mucosa by the subject experts were 100% and for Roboflow was 80%. The true positive results for suspected lesions (OPMDs and OSCCs) by the subject experts was 90% and the Roboflow model was 85%. Based on Kappa statistical analysis, there was a 0.7 value showing moderate agreement between the blinded subject experts and the Roboflow model was almost perfect. Tanriver et al (2021) performed instance segmentation experiments to predict a bounding box, a class label, and a pixel-level mask for each object instance.
separately. Various CNN-based networks like ResNet-50, ResNet-101, and ResNeXt-101 were compared with and without TTA. Among them, the ResNeXt-101 model achieved an AP score of 43.90% for box detection and 37.85% for mask detection respectively. For the object detection experiment, different versions of YOLOv5 were evaluated in one class lesion detection task. YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5l+TTA, YOLOv5x, YOLOv5x+TTA, YOLOv5s & 5m ensemble were compared. Among them, the YOLOv5s & 5m ensemble achieved the best performance with an AP of 63.7%.

2.4 Discussion:

Lin et al (2021) demonstrated the effectiveness of the image-capturing method by comparing four different methods random positioning, center positioning, center positioning + over-sampling, and center positioning + resampling method. Among them, the center positioning + resampling method performed the best. The results showed that the centered ruled dataset achieved an 8% higher F1 score than that trained on the random positioning dataset. For the center positioning + resampling method the macro averages were sensitivity, specificity, precision, and F1 score were calculated. Various CNN networks like VGG-16, ResNet-50, and DenseNet-169 were compared in which HR-Net W18 performed the best. HR-Net W18 achieved the highest sensitivity, specificity, precision, and F1 score. A confusion matrix was generated for the HR-Net W18 model to the test set and the results showed that the F1 score was higher in the normal class when compared with the lower F1 scores for other classes, this concluded that classes with scarce samples are generally more problematic to classify. For various classes, AUC and F1 scores were calculated and
for normal class achieved the highest performance score. Whereas Dinesh Y et al (2023) did not perform any classification network-based model detection. Tanriver et al (2021) performed a semantic segmentation experiment in which image recognition tasks were performed that dealt with assigning each pixel of an image to a particular class including background. Various backbones from U-net architecture were used. Various backbones like Efficient Netb3, DenseNet-161, Inception-v4, EfficientNetb7, and ResNeXt-101_32x8d evaluated dice scores with and without TTA. Among them, the Efficient Net b7 model achieved the highest dice score without TTA and dice score with TTA. Also, classification experiments were performed with various CNN models like efficient Net-b4, Inception-v4, DenseNet-161, Ensemble, and ResNet-152 were compared. Among them, Efficient Net-b4 performed well with values like precision, Recall, and F1 score. According to different classes like benign, OPMD, and carcinoma the precision for carcinoma class, recall and F1 score achieved the highest performance.

Lin et al (2021) also introduced a class activation mapping technique on HR-Net W18 for prediction visualization. First, the last layer was removed and then upsampled to the original image size. The diagnostic multiclass performance of the method for all the five categories was compared. Normal achieved the best performance with AUC and F1 scores. Whereas Dinesh Y et al (2023) performed the statistical analyses were done using Kappa statistics with SPSS software v23.0 (IBM Corp, Armonk, USA). The specificity and sensitivity of the Roboflow Machine learning model were calculated. The true positive results of clinically healthy oral mucosa by the subject experts and for Roboflow were compared and Roboflow
achieved 90%. The true positive results for suspected lesions (OPMDs and OSCCs) by the subject experts and the Roboflow model were compared and Roboflow achieved 85%. Based on Kappa statistical analysis, showed that there was a moderate agreement between the blinded subject experts and the Roboflow model being almost perfect. Tanriver et al (2021) performed instance segmentation experiments to predict a bounding box, a class label, and a pixel-level mask for each object instance separately. Various CNN-based networks like ResNet-50, ResNet-101, and ResNeXt-101 were compared with and without TTA. Among them, the ResNeXt-101 model achieved the highest AP score for box detection and mask detection. For the object detection experiment, different versions of YOLOv5 were evaluated in one class lesion detection task. YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5l+TTA, YOLOv5x, YOLOv5x+TTA, YOLOv5s & 5m ensemble were compared. Among them, the YOLOv5s & 5m ensemble achieved the best performance with an AP score.

For Lin et al (2021), the limitations were that 1) The images of oral diseases were only obtained from one hospital, with the patients coming from surrounding cities. More data from additional hospitals and provinces are required to test the generalizability of our approach to other patient populations. (2) The images in the experiments were mainly collected from four different smartphone cameras. (3) There were only five categories for evaluation in the study and there were many other oral disease types not included. Whereas for Dinesh Y et al (2023), the limitations were the collection of images of varying resolutions with mobile cameras or iPads. Even physiologically altered conditions like Fordyce granules were identified as suspected lesions. The algorithm could not make definite predictions for specific OPMDs and
their potential for malignant transformation which was like subject experts unless specific parameters were assessed. Also, Tanriver et al (2021) had a limitation where the study did not have a lot of categories for lesion classification. Only three lesion classes were included to detect the lesion images, there should be more specific categories for classification of the lesion images.

2.5 Conclusion:

In conclusion, all three studies were appropriate for the early prediction of oral cancer detection using intraoral clinical images which can assist dentists, patients, and the community. With early detection of oral cancer, management and diagnosis of oral cancer and OPMDs will be easier. These studies are the stepping stones for using artificial intelligence (ML models) for managing oral cancer patients and it is also helpful in reducing other disparities across the world.

2.6 Gaps in literature review and needed future directions:

Lin al (2021), had some limitations. The images of oral diseases were only obtained from one hospital, with the patients coming from surrounding cities. The images in the experiments were mainly collected from four different smartphone cameras with different resolutions. There were only five categories for evaluation in the study and there were many other oral disease types not included. Whereas Dinesh Y et al (2023), had the limitations were the collection of images of varying resolutions with mobile cameras or iPads. Even physiologically altered conditions like Fordyce granules were identified as suspected lesions. The algorithm was not able to make definite predictions for specific OPMDs and their potential for malignant transformation which was like subject experts unless specific parameters
were assessed. And Tanriver et al (2021) had a limitation where the study did not have a lot of categories for lesion classification. There should be more specific categories for the classification of the lesion images.

To train machine learning as quickly and accurately as possible, a larger dataset with more examples of challenging lesions should be used to attain significant accuracy for the models. More focus should be given to it in future studies. There should be more categories in classifying the lesion images for attaining a higher accuracy and precision. Also, different sizes of images with different pixels, sharpness, and contrast should be used to make the algorithm more apt. In the future, this can help the dentists and community by functioning as an oral cancer screening tool and smartphone mobile applications.

Therefore the aim of the study is to detect and distinguish normal tissue, oral malignant and non-malignant lesions from clinical photographs using YOLO v8 deep learning algorithm.
CHAPTER 3
MATERIAL AND METHODS

This was a diagnostic study conducted using clinical images of oral cavity lesions. There were 455 clinical images of the oral cavity which were extracted from a publicly available dataset repository specifically the Kaggle and Mendeley data repositories. [(Chandrashekar et al (2021), Shivam et al (2020)].

3.1 Recruitment Methods

The data was from patients who suffered from any ulcerative, red and white lesions from different areas of the oral cavity such as buccal mucosa, tongue, upper/lower alveolar ridge, floor of the mouth, retromolar trigone, and lip from publicly available datasets like Kaggle and Mendeley data repositories [(Chandrashekhar et al (2021), Shivam et al (2021)]. The dataset included a diverse range of lesions resulting from various oral diseases and anatomical regions. Based on some previous studies and domain expertise, we estimated a sample size of 500 intraoral photos. To achieve this, we obtained about 455 clinical images from data repositories. After removing the blurry and duplicate images from the dataset, we got 427 clinical images for this study.
Table 1. Eligibility Criteria

<table>
<thead>
<tr>
<th><strong>Population</strong></th>
<th>An individual with the presence of ulcerative, red and white lesions of the oral cavity.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Setting</strong></td>
<td>Publicly available data repositories</td>
</tr>
<tr>
<td><strong>Materials Collected</strong></td>
<td>Clinical intraoral photograph:</td>
</tr>
<tr>
<td></td>
<td>- Presence of ulcerative, red and white lesions from different areas of the oral cavity like buccal mucosa, tongue, upper/lower alveolar ridge, floor of the mouth, retromolar trigone, and lip.</td>
</tr>
<tr>
<td><strong>Image Quality</strong></td>
<td>Blurry and duplicate images were eliminated</td>
</tr>
</tbody>
</table>

3.2 Study Procedures and Data Analysis

The intraoral images were extracted directly from Kaggle and Mendeley data repositories [(Chandrashekhar et al (2021), Shivam et al (2020)] and were stored in a password-protected computer. The data collected included clinical images of red and white lesions. No extraoral photos or videos were collected. All the images utilized in the study were divided into three categories namely normal, non-malignant, and malignant lesions. The dataset was annotated by two well-experienced oral pathologists, one oral surgeon, one prosthodontist, and one general dentist using Roboflow annotation software. (Dwyer et al, 2024)

The images collected were first set to a resolution of 640 x 640 pixels and then randomly split into 3 sets: training, validation, and testing – 70:20:10, respectively. The images were divided into n=299 images as training set, n= 85 images as validation set and
n= 43 images as testing set. Images in the training and validation set were annotated by two independent oral pathologists, one oral surgeon, one prosthodontist, and one general dentist using Roboflow annotation software (Dwyer et al, 2024). The data was then analyzed using the YOLO v8 object classification algorithm (Jocher et al, 2023). Ninety percent (90%) of the dataset was used in the training and validation sets which ran at a batch size of 1 with 20 epochs for the YOLO v8 algorithm (Jocher et al, 2023). And the remaining 10 percent (10%) was used to test the performance of the algorithm in the detection of oral cancer, by running it through our already trained and validated algorithm. The performance of the algorithm was assessed using the following parameters – accuracy, sensitivity, specificity, and precision.
CHAPTER 4

RESULTS

After training and validation with 20 epochs, the result of our study showed an overall accuracy of 55\%, sensitivity of 56.33\%, a specificity of 43.33\%, and precision of 57.33\%. The training loss function graph - plot showing loss functions for training set across different epochs (Figure 1 A). The validation loss function graph - plot showing loss functions for validation set across different epochs (Figure 1 B). For metrics/accuracy top 1 function graph the plot shows accuracy top 1 across different epochs (Figure 1 C). For metrics/accuracy top 5 function graph - the plot showing accuracy top 5 function across different epochs. (Figure 1 D). Our oral cancer image classification algorithm showed maximum performance at 15 epochs.

Figure 1. A train/loss function graph
Figure 2. A validation/loss function graph

Figure 3. A metrics/accuracy_top1 function graph
Based on the generated normalized confusion matrix (Figure 2), the accuracy of our algorithm in classifying normal images, non-malignant images, and malignant images was 81%, 64%, and 68% respectively. The sensitivity of our algorithm in classifying normal images, non-malignant images, and malignant images was 71%, 47%, and 54%, respectively. The specificity of our algorithm for normal images, non-malignant images, and malignant images was 86%, 65%, and 72% respectively. The precision of our algorithm in classifying normal images, non-malignant images, and malignant images was 73%, 62%, and 35%, respectively. The overall accuracy of our oral cancer image classification algorithm was 55%.
### Table 2. Image classification algorithm outputs for each category

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>81%</td>
<td>71%</td>
<td>86%</td>
<td>73%</td>
</tr>
<tr>
<td>Non-malignant lesions</td>
<td>64%</td>
<td>47%</td>
<td>65%</td>
<td>62%</td>
</tr>
<tr>
<td>Malignant lesions</td>
<td>68%</td>
<td>54%</td>
<td>72%</td>
<td>35%</td>
</tr>
</tbody>
</table>

As shown in Figure 2, the generated confusion matrix showed true positives, true negatives, false positives and false negatives for the normal class was 0.73, 1.69, 0.27, and 0.3 respectively. The true positives, true negatives, false positives, and false negatives for the non-malignant class were 0.61, 1.3, 0.38, and 0.7 respectively. The true positives, true negatives, false positives, and false negatives for the malignant class were 0.35, 1.69, 0.65, and 0.3 respectively.
Figure 5. Confusion matrix normalized
CHAPTER 5

DISCUSSION

Oral cancer is a disease that has recently increased worldwide but is still not fully understood. OSCC accounts for more than 90% of oral cancers. The cause of OSCC is multifactorial, including extrinsic factors such as tobacco products and alcohol and intrinsic factors such as malnutrition and iron-deficiency anemia (Tanriver et al, 2021). Several novel methods are evolving in the field of diagnosis of cancers such as machine learning, exosomes, and paper-based biosensors. Oncology is being revolutionized by cutting-edge machine learning algorithms which provide rapid and cutting-edge methods for the diagnosis of potential lesions with an accuracy comparable to that of medical specialists. Despite the easy accessibility of the buccal cavity during regular examinations, numerous malignancies remain undetected until they reach advanced stages. Consequently, employing AI holds promise in addressing the high mortality rates linked to oral cancer. The increasing availability of medical digital data makes the progress of AI-driven image analysis at a rapid rate. This advancement shows the potential to enhance the quality of life of patients (Dinesh Y et al, 2023).

Exploiting the advancements in deep learning, our study classified the detected regions into normal, non-malignant, and malignant lesions based on the risk of malignant transformation. This study can enable the detection of normal, non-malignant, and malignant lesions and presents an opportunity for the development of a vision based oral cancer screening tool.
5.1 Data Source and Quality

In terms of data source, our study assessed oral cancer detection using retrospective publicly available image data from Kaggle and Mendeley data repositories [(Chandrashekhar et al (2021), Shivam et al (2020)]. Similarly, Dinesh et al (2023) used retrospective photographic intraoral images from their institution’s electronic health record while Tanriver et al (2021) used retrospective photographic images from their institution’s electronic health record and publicly available data. In contrast, Lin et al (2021) used smartphone-based images collected prospectively from patients. Our study consisted of a dataset of 427 intraoral images after deleting the duplicate images and blurry images which is higher than the 360-sample size used in the Dinesh et al (2023)’s study but much lower than the 1448 sample size used in Lin et al (2021) and Tanriver et al (2021) had a final dataset for object detection experiments consisting of 652 images.

In summary, despite the limited amount of data, our dataset provides a good representation of real-world data given the use of multiple publicly available datasets. This study also highlights the scarcity and difficulty in obtaining clinical photographs for building and deployment of image-based AI models in dentistry.

5.2 Algorithm and Performance of Deep Learning Model

Deep learning models can use three types of object detection depending upon the operator’s purpose, 1) object classification, 2) object detection, and 3) object segmentation. Object classification refers to a type of labeling images, with the goal of answering the question, “What is in this image?” on the other hand, object detection is a computer vision technique that deals with distinguishing between objects in an image. It is
more specific in what it identifies, applying classification to distinct objects in an image and using bounding boxes for precise detection. Object segmentation is a type of labeling where each pixel in an image is labeled with given concepts, making it easier to analyze.

In terms of the use of the algorithm, our study used the YOLO V8 object detection algorithm (Jocher et al, 2023) for annotating and classifying the clinical images obtained. Whereas Tanriver et al (2021) and Lin et al (2021) used various CNN-based models for object classification experiments. In contrast to it, Dinesh Y et al (2023) and Tanriver et al (2021) used Roboflow and YOLO v5 for object detection experiments. In addition, Tanriver et al (2021) also performed object segmentation experiments by using U-Net-based models.

In terms of algorithm performance, our study analyzed the images using the YOLO v8 object detection algorithm (Jocher et al, 2023) and calculated sensitivity, specificity, accuracy, and precision. Our accuracy was relatively high in detecting normal, non-malignant, and malignant categories. We found no other studies that calculated accuracy in their studies (Lin et al (2021); Tanriver et al (2021); Dinesh Y et al., (2023)). Our study found a higher precision in detecting normal tissue compared to the findings by Dinesh et al’s (2023) study but had lower precision compared to the findings from Tanriver et al (2021). Our study also compared sensitivity and specificity for all three categories which was relatively low compared to that of Lin et al (2021) which had sensitivity and specificity.

In summary, though our study had lower percentages of precision, sensitivity, and specificity, it highlighted higher percentages of accuracy in detecting oral cancer and precancerous lesions.
5.3 Strengths and Limitations:

The strengths of our study include (1) the use of multiple experts in the oral pathology domain for annotating the available dataset. These were two oral pathologists, one oral surgeon, one prosthodontist, and one general dentist. Senior oral pathologist was considered as “gold standard” in our study. All data were keenly observed by all the data annotators and finally reviewed by the oral pathologist to confirm the final diagnosis of each annotator. (2) Categorization of the lesions into three classes - normal, malignant, and non-malignant lesions. Previous studies had categorized the lesions into only two classes. These categories are important for detecting oral cancer and pre-cancerous lesions present in the oral cavity.

Our study has some limitations. One major limitation is that (1) It is a retrospective study, using finite oral images. Because of this, the algorithm was not able to get a higher percentage of precision, sensitivity, and specificity. (2) The images of our dataset were only obtained from publicly available data repositories, especially Kaggle and Mendeley. After the search, we found that there is a limited amount of data available on search engines for research studies. (3) Validation of the authenticity and lack of control of image quality and accuracy of the labels. The study was based on clinically available data, so the image quality and quantity were extremely affected. (4) There was no clinical diagnostic evidence to back our labeling of lesions. Because it has not been diagnosed by a certified institute. (5) There were issues for misclassification which included not properly accounting for different cancer types. Oral diseases are divided into various red and white lesion categories, for precise diagnosis of the clinical images, more categories and defined classification becomes necessary. (6) Our study only performed resizing and
noise reduction for the given images but did not perform any other data preprocessing such as grayscaling, normalization, binarization and contrast enhancement which would have further enhanced the performance of our algorithm. Our study had only performed object classification for the given clinical images. For precise diagnosis of a lesion, object detection and object segmentation have the higher performance of the algorithm.

5.4 Future directions:

In order to train machine learning as quickly and accurately as possible, a larger data set with more examples of challenging lesions should be used to attain a significant accuracy for the models. More focus should be given to it in future studies. A possible prospective study backed by labs or histological evidence of cancer should be conducted. External validity of our algorithm by comparing our algorithm with a clinician (general dentist) with a histological result and oral pathologist as the gold standard. There should be more categories in classifying the lesion images for attaining a higher accuracy and precision. Also, detailed image preprocessing should with different pixels, contrast and sharpeness should be performed to refine the algorithm. Image preprocessing is the process of manipulating raw image data into a usable and meaningful format. It eliminates unwanted distortions and enhance specific qualities essential for computer vision applications. There are several techniques used in image preprocessing:
(1) Resizing: Resizing images to a uniform size is important for machine learning algorithms to function properly. (2) Grayscaling: Converting color images to grayscale can simplify the image data and reduce computational needs for some algorithms. (3) Noise reduction: Smoothing, blurring, and filtering techniques can be applied to remove
unwanted noise from images. (4) Normalization: Normalization adjusts the intensity values of pixels to a desired range, often between 0 to 1. (5) Binarization: Binarization converts grayscale images to black and white by thresholding. (6) Contrast enhancement: The contrast of images can be adjusted using histogram equalization. With the right combination of these techniques, a significant improvement can be seen in image data and computer vision applications. In the future, this can help the dentists and community by functioning as an oral cancer screening tool and smartphone mobile applications.

However, we believe the generalization performance of the model can be guaranteed after using other kinds of oral cavity images from multiple users for network training, as the appearance of disease variants can be learned by the network. In addition, further validation on other oral disease types (such as oral thrush) is still required to fully establish the performance characteristics of our AI diagnosis system. In summary, our algorithm may be helpful in places where the doctor has less experience or for patients’ self-prediagnosis but is unlikely to become a tool for clinical specialists.
CHAPTER 6

CONCLUSION

We conclude that there is a potential in machine learning algorithms for the automatic identification of normal, malignant, and non-malignant lesions. Based on the results obtained, we propose the given model for an oral cancer screening tool with low computational costs. This model can work as a low-cost, non-invasive, easy-to-use tool for the early detection of malignancy and pre-malignant lesions. Future studies will focus more on object detection and object segmentation training of the algorithm and obtain a large dataset to train the model with more examples and challenging lesions.
REFERENCES


Kaggle datasets download -d shivam17299/oral-cancer-lips-and-tongue-images


APPENDIX A

NORMAL CLINICAL IMAGE
APPENDIX B

NON-MALIGNANT CLINICAL IMAGE
APPENDIX C

MALIGNANT CLINICAL IMAGE