Thesis Title: Explainable Deep Learning for Oral Disease Detection with LLM Decision Support.

Supervised By

Md. Shohel Arman

Assistant Professor

Department of Software Engineering

Daffodil International University

Submitted By

IKHTIAR HOSSEN

ID:213-35-811

Department of Software Engineering

Daffodil International University

This thesis report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering.

APPROVAL

This thesis titled on **"Explainable Deep Learning for Oral Disease Detection with LLM Decision Support"**, submitted by **IKHTIAR HOSSEN (ID: 213-35-811)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Md Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md Habibur Rahman
Associate Professor
Department of Computer Science and Engineering
Islamic University, Bangladesh

Explainable Deep Learning for Oral Disease Detection with LLM Decision Support

IKHTIAR HOSSEN

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

SUPERVISOR's DECLARATION

I hereby declare that I have checked this thesis and, in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor's Signature)

Full Name : Md. Shohel Arman

Position : Assistant Professor

Date : 13 September 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

IKHTIAR

(Student's Signature)

Full Name : IKHTIAR HOSSEN

ID Number : 213-35-811

Date : 13 September 2025

Explainable Deep Learning for Oral Disease Detection with LLM Decision Support

IKHTIAR HOSSEN

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

SEPTEMBER 2025

ACKNOWLEDGEMENTS

To complete my undergraduate thesis, I want to thank Almighty Allah for giving His Blessings.

I want to show gratitude to my Supervisor Md. Shohel Arman, Assistant Professor of Software Engineering at Daffodil International University, Dhaka. He deserves my utmost thanks and respect. His guidance and powerful expert knowledge of the section "Deep Learning" helped me complete this whole thesis work. He has accounted for this through his unwavering magnanimity and understanding, scholarly leadership, abiding encouragement, careful oversight, constructive feedback, valuable suggestions, and review of all the awful and potentially inept manuscripts he corrected at every point,

Finally, but just as importantly, I want to show gratitude to my parents for their unwavering affection and support, as well as for putting me where I am now.

DEDICATION

I dedicate this thesis to my parents for their unwavering love and sacrifice, to my friends for their unwavering support and collaboration, and to my university teachers for their company.

ABSTRACT

This thesis develops a comprehensive pipeline that automate detection and classification of oral diseases from RGB intraoral photographs, which addresses the three main challenges of accuracy, interpretability, and clinical relevance. Using an advanced convolutional neural network InceptionResNetV2, fine-tuned on the Mouth and Oral Disease dataset, which includes images across seven classes which includes Canker Sores, Gingivostomatitis, Oral Cancer, Oral Lichen Planus, Cold Sores, Oral thrush and Mouth Cancer. The model achieved an overall accuracy of 99.60%, with F1-scores, precision, and recall that were near perfect for most classes. To support transparency, we use Grad-CAM++ to produce heatmaps that highlight clinically relevant lesion regions, enabling dentists to interpret and validate the model's predictions. Furthermore, to generate human readable narrative a large language model DeepSeek R1 was used: one prompt generates clinically detailed explanations for dentists, including disease descriptions, potential causes, and management guidelines; while another prompt generates simplified, patient-friendly explanations to improve understanding and engagement. This dual-perspective approach ensures that both healthcare professionals and patients benefit from the diagnostic pipeline, enhancing confidence, interpretability, and usability.

TABLE OF CONTENT

List of Tables

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
AUC	Area Under the Curve
CAD	Computer-Aided Diagnosis
CaS	Canker Sores
CBCT	Cone-Beam Computed Tomography
CNN	Convolutional Neural Network
CoS	Cold Sores
DIC	Differential Interference Contrast
DL	Deep Learning
DSLR	Digital Single-Lens Reflex
F1	F1 Score
FCOS	Fully Convolutional One-Stage
GAN	Generative Adversarial Network
Grad-CAM	Gradient-weighted Class Activation Mapping
GPU	Graphics Processing Unit
Gum	Gingivostomatitis
ITSA	Improved Tunicate Swarm Algorithm
KNN	K-Nearest Neighbors
LLM	Large Language Model
mAP	Mean Average Precision
MC	Mouth Cancer
MIH	Molar-Incisor-Hypomineralization
MOD	Mouth and Oral Disease
OC	Oral Cancer
OCI	Oral Cancer Images
OLP	Oral Lichen Planus
OPMD	Oral Potentially Malignant Disorders
OPG	Orthopantomography
OSCC	Oral Squamous Cell Carcinoma
OT	Oral Thrush
RGB	Red Green Blue

Abbreviation	Full Form
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SVM	Support Vector Machine
WHO	World Health Organization
XAI	Explainable AI

CHAPTER 1

INTRODUCTION

1.1 Introduction and Significance

Oral diseases affect a significant sector of the global community and allow for the continuation of a worldwide public-health problem. According to the World Health Organizations estimate approximately 3.5 billion people have experienced oral conditions at some point in their lives; the most common types of disease are tooth loss, periodontal diseases, dental caries, and oral cancers [1]. Many other oral and mucosal pathologies examples include aphthous ulcers, herpetic lesions, oral candidiasis and lichen planus present as discernable changes of the oral cavity, and indicate the possibility of photographic screening for preliminary detection and triage. However, the conventional diagnosis of oral pathology relies on clinician inspection, imaging and laboratory studies. Moreover, provider access to specialist diagnosis is not visible and depends on the subjective interpretation of the examiner. Automated, image-based decision support has the potential to enhance screening productivity, reduce diagnostic variability and provide earlier triage; ultimately an asset in locations with fewer resources, or geographic access [1].

The field of computer vision and imaging had seen a substantial advancement of CNN in automated interpretation. In terms of dental and oral imaging work-up (radiographs and photographs) CNNs have been utilized for tooth detection and identification, caries localization, classifying restorations, and identifying mucosal lesions with rates of success that are hopeful [2, 3, 4, 5, 6]. Despite this success, clinical uptake continues to be limited after facing two key barriers: the black-box complexity of deep models (clinician trust and interpretability) and importantly, the missing actionable, communicative outputs that are better designed for the clinician and patient as research outputs. This work intends to break the barriers: (a) improve the classification accuracy for seven oral disease classes with a high-capacity CNN backbone (InceptionResNetV2) (b) explain predictions with Grad-CAM++ visual clarity (c) provide dual human-readable second opinions as a front-end output using a large language model (LLM), generating clinically detailed explanations for dentists and simplified, patient-friendly guidance for non-experts.

1.2 Background

Deep learning in oral and dental imaging

Deep CNNs have shown successful deployment across many oral imaging types. For panoramic and periapical radiographs, hybrid pipelines combining CNN detection with raw measurement in a CAD-like fashion have been used to predict periodontal bone loss, and automatically stage periodontitis, with good agreement compared to expert ratings [2]. To

locate and count teeth in periapical image object detection models like Faster R-CNN have been used. which is an important step enabling effectively (problem) assessment at the tooth level [3]. For caries, multiple groups have demonstrated CNNs that demonstrated sensitivity matching or exceeding that of clinicians when detecting caries with bitewing and panoramic images [4, 7]. More recently, photographic (RGB) approaches are also beginning to show promise: Pilot studies have detected white-spot lesions (and early enamel) changes from tooth photos, and some other works classify mucosal ulcerations from clinical oral photos with high AUCs, when datasets are well-curated [5, 6]. These efforts together provide evidence that photographic oral disease classification by CNNs is feasible if datasets are sufficiently large and models are well engineered.

Explainable AI (XAI) in medical imaging

Model opacity is a major impediment to clinical uptake. Methods that demonstrate saliency or class activation mapping (e.g. Grad-CAM and improved Grad-CAM++) enable visual interpretation by highlighting the most significant regions of an image for an individual prediction [8, 9]. The application of these heatmaps in medicine is common to confirm that models are looking at clinically relevant structures and to build clinician trust in fully automated outputs. To our knowledge, in dental and oral imaging, Grad-CAM visualizations have been performed in the examination of portioned model attention on lesion sites and helping to identify model failure modes, but no notable systematic application in multi-class photographic oral diagnosis.

LLMs for clinical decision support

In clinical context the natural language tasks that needed human labor has been transformed with the emerge of large language models, including question-answering, summarization, and report writing. Recent reviews highlight both the promise of LLMs but also significant risks associated with hallucination, bias, and need for clinician oversight [12]. In medical imaging, there are also several new studies that discuss how LLMs perform well in bridging Computer-Aided Detection (CAD) outputs to human-readable reports, multi-modal outputs, and to improve downstream performance. As an example, the ChatCAD framework converts vision model outputs into text, and then uses an LLM to create succinct, accurate reports to further downstream tasks, achieving considerable improvements on chest X-ray tasks [10, 11]. These works support the hypothesis that adding LLM to the CNN pipeline could produce more enriched and actionable outputs.

1.3 Motivations and gap analysis

Even considering the significant advances of each of the three technology components CNN classification, XAI visualization, and LLM-mediated report generation, there are no known peer-reviewed studies that brings together all three technologies focused on mouth/oral disease diagnosis from RGB intraoral photographs. The literature demonstrates that CNNs have the capability of discriminate the lesions, Grad-CAM variants have the capability of explaining

CNNs, and LLMs can convert structured output data into clinically useful narratives but none of these capacities have been combined and evaluated across the oral domain. There are related multimodal integrations in other imaging domains: the ChatCAD and subsequent work integrates CAD networks and large language models (LLMs) for chest radiographs and general medical images with an overall improvement in report quality and in some experiments diagnostics performance [10, 11]. These cross-domain successes provide justification for a domain-adapted pipeline that incorporates all three technologies to the oral cavity: this system could improve per-image accuracy through strong CNN modeling, while enhancing clinical trust and usability through interpretable visualizations and dual LLM-generated reports clinically detailed outputs for dentists and simplified explanations for patients.
There are three concrete motivating aspects of this thesis:

1. When trying to complete multi-class classification of visually similar oral conditions accuracy is difficult; the potential to improve the discriminability of classification using high-capacity CNN architectures with transfer learning, especially InceptionResNetV2.

2. It is psychosocially easier for clinicians to trust and implement AI when the operational model decisions have sense-making rooted to the relevant visual prompts meaningful visual explanations of their multi-class classification outputs using Grad-CAM++ heatmaps that localize the features, potentially including consideration for various factors or contexts.

3. The model also outputs actionable narratives generated by an LLM, providing clinically detailed explanations for dentists and simplified, patient-friendly descriptions. These narratives include disease information, likely causes, and basic management or referral advice, thereby converting AI predictions into usable guidance for both experts and non-experts.

As there is no example of all three of these elements CNN + Grad-CAM++ + LLM evidenced using oral photographs, there is an opportunity to contribute to existing knowledge that could be new and meaningful in terms of providing both better technical performance and real-world utility.

1.4 Problem statement

This thesis addresses the question of how to enhance the accuracy, interpretability, and applied utility of automatic detection of mouth/oral diseases from RGB intraoral photographs. More specifically, we ask the following:

How can Explainable AI (XAI) and Large Language Models (LLMs) be integrated with Deep Learning (DL) based Oral disease detection to improve transparency, clinical decision support, and real-world adoption?

Subproblems include; building a robust classifier for the seven classes; verifying that the Grad-CAM++ heatmaps are clinically useful; and framing LLM prompt questions that produce accurate and concise explanations and management options for both dentists and patients.

1.5 Research questions

To operationalize the problem, we need to ask:

1. Performance: What classification accuracy, sensitivity, specificity, and class-wise behavior is achievable on a curated dataset of RGB oral images for the seven target classes using InceptionResNetV2 with transfer learning?

2. Interpretability: Do the generated Grad-CAM++ visualizations reliably highlight clinically relevant areas (lesions, inflammation, ulceration) for each predicted class?

3. LLM decision support: Given the CNN prediction and selected metadata, can an LLM correctly generate clinically useful text (disease description, likely causes/risk factors, next steps/referral advice) as a second opinion?

4. Integrated value: Does the overall approach CNN + Grad-CAM++ + LLM provide clinicians with significantly better information than they would have received from a CNN-only approach when determining whether or not to accept/verify the model's outputs?

1.6 Scope and limitations

This study is limited as an academic prototype (proof-of-concept), not a clinically deployed device. There are significant limitations here:

- Modality: only RGB clinical intraoral photographs are being considered.

- Classes: there are seven defined classes which are Oral Lichen Planus, Canker Sores, Gingivostomatitis, Oral Cancer, Cold Sores, Oral thrush and Mouth Cancer.

- XAI method: Grad-CAM++ is being used as the main visual explanation method [9], but we will explore additional methods of XAI for comparison.

- LLM: we are going to use an available LLM DeepSeek-R1 by using prompt engineering and controlling outputs; we will provide some evaluation of factuality and safety of guidance but we will not have any retraining of the LLM.

- Evaluation: classification will be evaluated quantitatively which are accuracy, F1, per-class metrics.

1.7 Objectives

The objectives of the thesis are:

- We designed and optimized a deep learning based classifier capable of distinguishing seven oral disease classes.

- Implement XAI visual explanations to find lesions area of the image for transparency.

- Embed an LLM to generate dual narrative reports for clinically detailed outputs for dentists and simplified explanations for patients covering the diagnosis, likely causes, and suggested next steps.

1.8 Thesis Outline

Chapter 1: Introduction: Contextual Background, Study Motivation, Research Challenges, and Objectives.
Chapter 2: Literature Review: detailed review of CNN methods for oral/dental imaging, XAI in medical imaging (Grad-CAM family), LLMs in clinical support, and existing attempts to combine vision and language in medicine.
Chapter 3: Methodology: dataset description, preprocessing, model architecture (InceptionResNetV2 details), training regimen, Grad-CAM++ implementation, LLM prompt design, and evaluation protocols.
Chapter 4: Result & Discussion: Discussion of results, comparison to prior work.
Chapter 5: Conclusion and Future Scope: Contributions summary, potential clinical translation paths, and recommended next steps.

1.9 Summary

In this chapter, we have outlined the drivers and rationale for a multi-modal approach to mouth/oral disease detection that incorporates high-capacity CNNs, interpretable visual explanations via explainable AI (XAI), and narrative reports generated with language large models (LLMs). The LLM component produces dual outputs: clinically detailed reports for dentists and simplified explanations for patients, covering diagnosis, potential causes, and suggested management. The literature demonstrates robust, complementary advances in each

area—for example, the development of powerful CNNs for intraoral images [2–6], robust localization and explainability of results with Grad-CAM models [8, 9], and LLM-augmented reporting systems in other imaging domains [10, 11, 12]. Notably, no peer-reviewed papers were found that combine all three components aimed at detecting mouth/oral disease in RGB intraoral photographs, which drives the novelty and prospective impact of this thesis. The next chapter will examine the technical and clinical literature in detail and situate the current work within that body of research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Detection of oral diseases automatically has increased a lot alongside the emerges of deep learning models like convolutional neural networks. CNN-based systems have been researched to detect and classify a wide spectrum of oral and dental disorders ranging from dental caries and periodontal diseases to mucosal lesions such as herpes simplex lesions, aphthous ulcerations, oral candidiasis, and oral lichen planus in photographic, radiographic, or histopathological images [14, 15, 16]. Systematic and scoping reviews show that deep learning models may have encouraging levels of sensitivity and specificity on curated datasets, however, some recurring challenges were identified: dataset heterogeneity, limited external validation, class imbalance, and the need for explainability and clinician-in-the-loop evaluation [31, 30]. Combined, the literature encourages an integrated initiative that seeks high multiclass photographic classification accuracy while also generating transparent and actionable outputs for clinicians and patients using an CNN classifier with explainable AI for visual explanations and a LLM for text decision support.

2.2 Previous Work

The study detects OLP from other non-OLP lesions with clinical images by training a series of CNNs such as Xception, ResNet152V2, and EfficientNetB3. 1,089 total images from which 609 are OLP and 480 are non-OLP was used to conduct this study, with lesions that could be histopathology confirmed if necessary. The images were all resized to 256 × 256 pixels with some data augmentation in the form of rotations and flips to create additional training images. The best performing model was Xception with an accuracy of 88.18%, sensitivity of 92.73%, specificity of 83.64%, precision of 85% and an F1-score of 88.7%; no AUC was reported. This work adds to the data available to using CNNs to classify mucosal lesions as part of the binary classification problem, noting that there were cases where confusion was noted with lesions resembling traumatic ulcers or epithelial dysplasia indicating that larger datasets will help mitigate CW-LP classification recognition in confusing cases [14]

Researchers proposed a YOLOv3 model to classify dental diseases with OPG x-ray images, disease like dental crowns cavity, broken down root canal and root canal. The custom dataset came from an original collection of high-resolution OPGs taken in clinics, totaling around 800 images. The researchers augmented the original images upon neural network training, creating

a total of 1200 images and then split 70 % for training and 30% for testing and every single image was annotated into four classes. The resulting YOLOv3 model being tested yielded an overall F1-score of 0.99, precision of 0.99, accuracy of 99.33%, recall (sensitivity) of 0.98, and a mean average precision (mAP) of 99.33%. Specificity was not reported and no AUC was calculated for the model. Several class detection models have been described, such as DeNTNet (accuracy 69%) and CNN based on ResNet50 (87.2%); however, performance like this model is suitable for multi-class detection and as such may easily translate into an automated computer-assisted tooling treatment of teeth and their diseases. The only limitation was a lack of installations of packages to gain internet access while training [15].

Using cone-beam computed tomography (CBCT) images dental caries was detected with multiple-input CNN, the lesions and its type with depth was also included. The dataset was obtained from 785 molar teeth images (382 carious, 403 noncarious). A total of 7850 images was created using augmentation resized to 96 × 160 pixels and used 30% for testing and 70% for training. The model found an accuracy of 95.3% for carious teeth, specificity of 96.3%, a sensitivity of 92.1%, and an F1-score of 93.2%. Precision was not represented and no AUC was provided. Classification accuracies ranged from 91.6% to 97.2% for types and 89.7% to 96.2% for extensions. The authors highlighted this study as a preliminary examination of DL in CBCT for caries and demonstrated better classification accuracy than the traditional methods but suggested that larger datasets be used to improve depth-specific classification. [16]

DenseNet169 a deep learning-based model was developed and using endoscopic images it detected tongue cancers. A dataset of 5576 images (1941 malignant, 3635 non-malignant cases) were used from five hospitals. A total of 5224 images were used for internal validation with 352 images for external testing. The model achieved 84.7% accuracy, specificity of 86.8%, with a sensitivity of 81.1% and an AUROC of 0.895. The paper did not mention any F1-score and precision. The problematic characteristic of the model was its superior performance to general physicians (accuracy 75.9%). While still inferior to oncologist specialists (92% accuracy), it showed reasonable agreement (kappa 0.685) indicating that as a decision support tool it could be helpfull for early detection to help primary care practitioners even though data integrity was affected by the nature of the retrospective data [17].

To detect malignant of oral lesions and benign a fine-tuning transfer learning method, VGG19, MobileNet, and DeIT, were used. The images were taken from the dataset on Mendeley, comprised of 323 initial images (165 benign, 158 malignant), which were augmented to 2593, and the augmented image then split into training, validating and testing samples. VGG19 and MobileNet reportedly achieved 100% in metrics like precision, accuracy, F1 score and recall. While DeIT had 98.73% accuracy, 94.24% sensitivity, 98.73% precision, and a 96.38% F1 score. Specificity and AUC were not stated. The performance shows that transfer learning is a powerful solution for datasets with small sample sizes and breaks the previous reported simply CNN method of 97% for images of oral lesions using about 55 images, and 91.54% accuracy using the fine-tuned VGG19 model method reported the same dataset as fine-tuning, but edited the data set to 50/50 classification and ultimately gives no direction of how to overcome this

problem which is one big limitation for this paper.[18]

The DOLNet model is a two-stage neural network with hierarchical attention and LesionMix augmentation, which was similarly introduced for the diagnosis of complex odontogenic lesions such as ameloblastoma, odontogenic keratocysts, and dentigerous cysts in panoramic radiographs. The images included in the dataset were 565 grayscale images of the lesions/diseases (75 ameloblastoma, 97 keratocysts, 193 dentigerous cysts, and 200 separate normal images). Each image was preprocessed to a resolution of 3000 × 1500 pixels prior to use for training purposes. The DOLNet model achieved a recall (or sensitivity) of 0.721, an accuracy of 0.871, precision of 0.832, and F1-score of 0.773. The authors did not report specificity or area under the curve. DOLNet improved recall by 42.4% and F1 by 44.2% to the previous works, also performed better than clinicians in detecting the smaller lesions, although there were issues with class imbalance that affected recall for the more rare types [19].

A CNN model with custom 19-layer was trained for early oral cancer classification via clinical lip and tongue images. The original oral cancer images dataset (OCI) consisted of 1,298 RGB images (650 cancerous, 648 non-cancerous) which were split by 80/20, and then preprocessed and normalized through augmentation techniques. The model that completed training on the images reached a F1- score of 97.07%, specificity of 98.46%, accuracy of 99.54%, sensitivity of 95.73%, precision of 98.45%, and an AUC of 0.6026. It outperformed using transfer learning such as the Inception model (91.06%) and therefore set a benchmark for detecting oral cancer images but the AUC results demonstrated moderate detection and indicated future improvements in class detection for noisy clinical data [20].

The YOLOv5-ConvNeXt integrated model (YoCNET) was developed to detect periapical lesions in radiographic images. 1305 images were used for training from which 650 positive image and 665 were negative image. A total of 3132 image was created after augmentation and used for training and for evaluation 717 images. YoCNET achieved a precision of 93.28%, with a sensitivity of 95.41%, accuracy of 94.25%, and a F1-score of 0.9433, and an AUC value of 0.9820. The YoCNET did not report specificity. Compared with ResNet34 (91.57% accuracy), the YoCNET accuracy is stronger than the ResNet34 transferred model. Meanwhile, the YoCNET prototype scored better than the YoRNET model (AUC 0.8822). It demonstrated it was able to surpass state-of-the-art integrated models for detecting oral cancer lesions and was able to find more multi-well lesions without having to conduct manual segmentation for findings; however, the images relied on high-quality radiographic images for scoring [21].

A systematic survey of AI-based approaches for periodontal disease classification reported the use of models such as neural networks, SVMs, and decision tree algorithms. The datasets employed in the studies were heterogeneous (e.g. AUC metrics ranging from 0.72 to 0.96 in various studies). There were insufficient studies to quantitatively pool sensitivity or specificity since I investigated the heterogeneity factor in studies, but some models had individual accuracies that did reflect clinical usefulness (e.g.,, the authors reported F1-scores between 0.95 to 0.70). Overall, the systematic review affirmed that AI had a utility in classification,

although variances in methods complicated clinical translation. The systematic review advocated standardized processes of reporting findings [22].

Using total of 220 frontal color teeth images from which 82 healthy and 138 unhealthy cased a CNN model was developed. The teeth images were rescaled to 640x640 pixels. The model using two 1D convolutions, achieved 74.54% accuracy. The specificity, sensitivity, F1-score, precision, and AUC were not reported for the CNN model achieving 74.54% accuracy, although the authors report a 99.9% F1 from tooth detection. The model improved 11.45% with the introduction of two additional convolutions over the ResNet152 baseline model and was proposed as usable for mobile applications, but limited to datasets with limited sample sizes [23].

Xception and MobileNet-v2 CNNs were used to classify oral leukoplakia and distinguish it from similar white lesions. The authors used a dataset that included 659 clinical images (202 leukoplakia images, and 457 of other white lesions), with 261 images used for testing. MobileNet-v2 obtained 92% accuracy, and 92% sensitivity for non-homogenous leukoplakia, and 93% specificity. However, the authors did not provide details of precision, F1-score, and AUC information, though they did report a macro F1 of 0.91. MobileNet-v2 had a 89% accuracy which was better than Xception, and the authors concluded that MobileNet-v2 was useful to aid in clinical judgement but needs to be used with more data on subtypes [24].

An InceptionResNetV2 model was finetuned to classify oral diseases and conditions, including mouth canker sores and oral cancers. The combined data from MOD contained 517 images, but data augmentation increased it to 5143, and then the data was split into 20% validation, 20% for testing and the remaining 60% for training. Accuracy after augmentation was 99.51%, precision of 100% for most classes, recall of 100% for most classes, F1 of 100% for most classes, and AUC of 100% for several classes. Whereas accuracy was dropped to 74.07% without augmentation. This model surpassed previous models that reached an accuracy of 95%, which demonstrates the importance of augmentation when dealing with the imbalanced dataset [25].

A novel study used a new AI model based on CNN with SSD architecture to evaluate its performance in being able to detect oral and dental diseases, namely plaque and caries. The dataset consisted of 5000 open-source images, with 90 designated for testing. The accuracy of the neural network was 81.11%, while the accuracy on pathology detection was 94.99% (i.e., detecting plaque and caries). There was no other evaluation metrics mentioned by the author in the paper other than accuracy. The total accuracy of the model matched the dentist experts (95.29%) and the rated agreement was only 81.02% pigeonholing the neural network into a potential diagnostic aid. However, it warrants further enhancement due to the prominent error rates found in detecting gingivitis (which was 8.89%) [26].

To identify potentially malignant oral disorders (OPMD) with oral squamous cell carcinoma (OSCC) a CNN framework was developed using ConvNeXt in a collaborative project. The

database contained 778 clinical images (404 OPMD, 374 OSCC). ConvNeXt reached an accuracy of 79.9% and 75.6% sensitivity, specificity of 84.2%, precision of 83.7%, F1 score of 79.4%, and a AUROC with a value of 0.863. Its performance on the binary classification tasks was observed to be substantial and likely aided by the use of Grad-CAM to explain the decisions of the model, rather the inherent problem with generalizability lies with the source dataset size [27].

A deep learning system using a CNN architecture was developed as an auxiliary diagnostic tool for oral mucosal lesion differential diagnosis, specifically identification of conditions such as aphthous ulcer and lichen planus. The dataset consisted of 506 images distributed across five classes. Performance resulted in a reporting of weighted precision as 88.8%, recall (sensitivity) as 88.2%, F1 as 0.878, specificity as 97.0%, and AUC as 0.985(mean) with the AUC ranging for each class between 0.98-1.00. The results showed strong performance characteristics in multiple classes (Kappa = 0.851) and emphasized the model's ability to assist in benign-malignant classification yet still required high-quality image input [28].

A second chapter proposed a deep learning model using CNN optimized using an improved tunicate swarm algorithm (ITSA) technique for oral cancer detection from tissue images. The original dataset consisted of 181 images (87 cancerous, 44 non-cancerous, and 50 normal). The authors reported for their CNN-ITSA model 98.7% accuracy, 93.71% sensitivity, 96.42% specificity, 96.42% precision, and 90.08% F1-score. No AUC values were reported. The CNN-ITSA model produced a better predictive performance than SVM (79.46%) and Mask R-CNN (92.52%). The authors showed the advantage of optimization and associated CNN and deep learning methods in new approaches, but oral cancer screening reports small dataset further supports expansion research [29].

The diagnosis and prognosis of oral cancer using deep learning models was evaluated in a systematic review. This systemic review included 54. The study found accuracies of 85-100% for histopathology images, F1-scores on detection of 79.31-89.0%, and Dice indices for segmentation from 76.0-96.3%. The pooled sensitivity for classification was 0.92, the pooled specificity was 0.92. There was no pooled AUC reported, but performance of studies was high overall. The review confirmed there is efficacy for the application of deep learning but highlighted there is risk for bias, and need for datasets to be reviewed and validated [30].

A scoping review and meta-analysis studied the automated detection of oral malignant lesions using DL, there were 14 studies reviewed and the pooled sensitivity was calculated to be 0.86 and the pooled specificity was 0.67. There was not a pooled AUC, yet individual AUCs were noted to be as high as 0.95. They noted heterogeneity with architectures including ResNet. They concluded that DL has improved our diagnostic abilities, but future research should focus on developing standardized algorithms [31].

Explainable AI through the use of Xception, was utilized for multiclass grading of OSCC in histopathology images and used Grad-CAM. Total of 322 images were present in the dataset

and the authors augmented their dataset to 2056 images. While accuracy was not calculated the AUC macro was 0.929 and micro AUC was 0.942. Specificities, sensitivities, precisions, and F1 were not stated. However, it has outperformed baseline models like KNN (AUC macro 0.539). The authors emphasized the importance of explainability to instill faith that clinicians may trust what the software is stating [32].

A deep learning model trained using the FCOS detector (named Candescence) classified Candida albicans morphologies observed in microscopy images. The Candescence dataset totalled 1214 DIC images (~80,000 cells). The model's accuracies were: accuracy = 82 %, sensitivity = 85.1 %, precision = 80.7 %, F1 = 83.2 %. Other metrics such as specificity and AUC were not explicitly set out. The GANs were useful for morphological analysis, indeed they were so useful that they outperformed human labelling, however the performance of blur detection adjustments identified nuances and difficulties of annotation [33].

Machine learning method (Roboflow) was applied in a prospective multi-center study for oral lesions from lesion intraoral images. The sample had 360 images from 4 centres and for testing the method 60 samples and 300 for training were used. The precisions were: precision = 29.8 % and recall = 32.9 %, sensitivity = 88.9 %, specificity = 75 %, mAP = 25.4. F1 and AUC not given. The authors reported they had moderate agreement (Kappa 0.7) with experts so the study supports early screening, however lowers precisions are warranted further refinements to correctly identify false positives [34].

The first pilot study used a Faster-RCNN with ResNet-101 to detect white mucosal lesion elementary TWO. The dataset was 221 clinical photos of white mucosal lesions across six classes. The outcomes were reported in external and internal precision times were: mAP of 82.3 with precision of 77.2 %, along with F1 or 74.4% and sensitivity with 76.0 %. These outcomes do not provide information on other metrics like specificity and AUC. Although, the outcomes for important classes such as condyloma (F1 = 94.1) suggest educational merits for students and the wider community, there were substantial low sensitivity for leukoplakia (25 %), leading to proposals for dataset expansion [35].

2.3 Summary

This chapter reviews peer-reviewed journal literature that is relevant for oral disease classification and detection automatically for photographic, radiographic and histological formats. The reviewed literature includes works on diseases that have been studied from a disease-specific perspective like OLP, leukoplakia, candidiasis to literature on radiographic lesions detection in X-rays to multimodal photographic systems including the base 7-class InceptionResNetV2 study [25]. The literature notes the importance of explainable AI methods, such as the Grad-CAM family, for enhancing clinician trust, as well as for identifying potential failure modes of the model [32, 30, 35]. The literature also notes that there are early examples of LLM-augmented decision support systems in other imaging domains. However, few peer-

reviewed studies in oral photography link high-accuracy multiclass classifiers with both visual explanations and narrative LLM-based reports. This gap motivates the present thesis's integrated approach (InceptionResNetV2 + Grad-CAM++ + LLM), which aims to deliver improved accuracy, interpretability, and usable outputs for both clinicians and patients in the 7-class oral disease detection challenge.

CHAPTER 3

METHODOLOGY

3.1 Introduction

A deep learning-based method was developed in this study for computer-aided oral and mouth disease detection, integrating cutting-edge interpretability techniques for the augmentation of clinical confidence and decision support. This work utilizes InceptionResNetV2 as a strong CNN architecture for precise and stable disease classification of intraoral images. To face the famous black-box problem of DL methods, the introduced Grad-CAM++ approach generates class-discriminatory maps, which highlight those regions of the oral cavity that constitute relevant features of the disease. The language model then uses the predictions from the model to generate insights into the diagnosis and recommendations for treatment in a form that is understandable to human beings for both dentists and patients. Essentially, it combines cutting-edge classification with visual interpretability and LLM decision support as the foundations of a framework aimed at AI-assisted diagnosis of oral diseases that is reliable, transparent, and user friendly. Figure 1 illustrates the overall methodology workflow, encompassing data preprocessing, model classification, interpretability, and diagnostic insights generation.

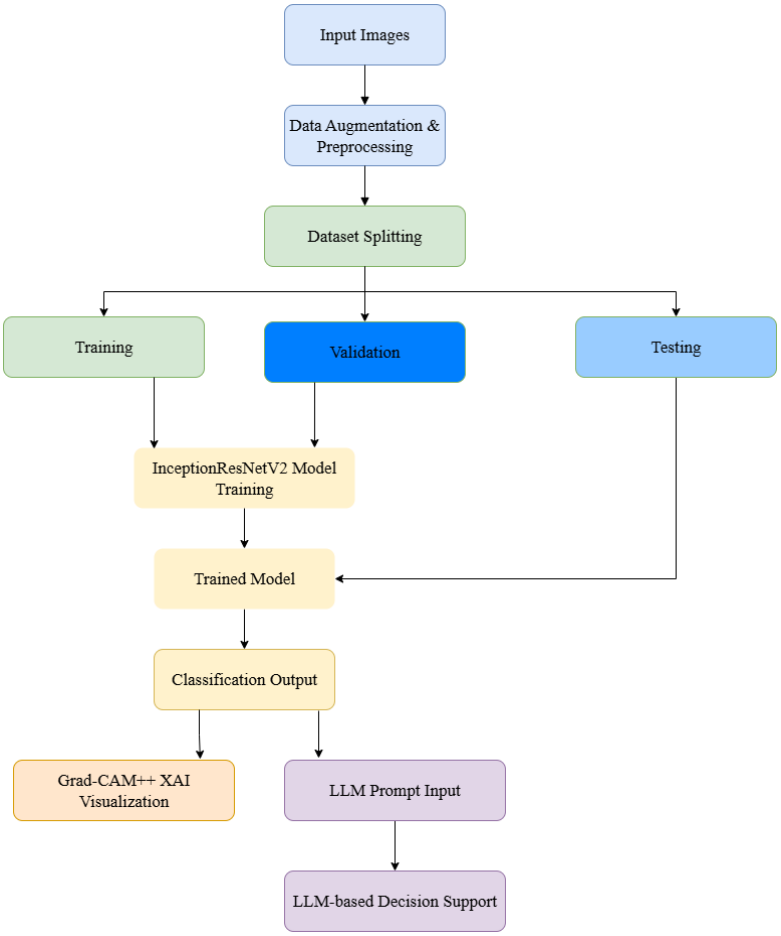

Figure 3.1: Overall Methodology Workflow

3.2 Dataset Collection and Preprocessing

3.2.1 Dataset: Oral and Mouth Disease Dataset

In order to detect oral diseases, the first step to create a strong model is to have a proper and structured dataset. In this work, we used the Mouth and Oral Disease dataset, created and shown originally by researchers in Okara, Punjab, Pakistan and augmented with images from dental websites [25]. The MOD dataset contains 517 clinical images and web images when the dataset was created. The dataset contains seven classes of oral disease: Oral Lichen Planus, Cold Sores, Canker Sores, Oral Thrush, Mouth Cancer, Oral Cancer, and Gingivostomatitis.

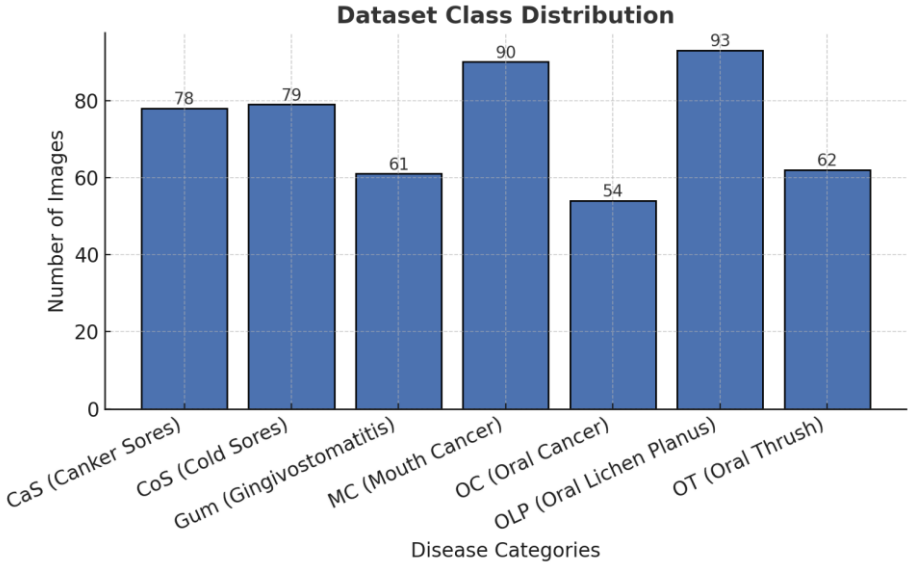

Figure 3.2: Dataset Class Distribution

Expert dentists labeled and verified each photo to confirm the correct class label. Additionally, there were no identifiers introduced during pre or post-image collection which protected patient privacy and anonymity. Figure 3 shows one sample from each class

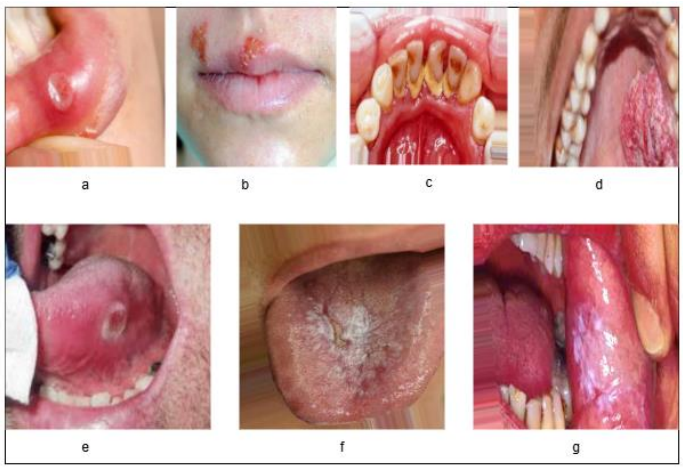

Figure 3.3: a. CaS, b. CoS, c. Gum, d. MC, e. OC, f. OLP and g. OT Class of the Dataset

3.2.2 Dataset augmentation

For data augmentation multiple techniques were used to strengthen the model and minimize the prospect of overfitting due to limited training data. Data augmentation can augment the training dataset, effectively creating different instances of the images that depict diversities of themselves, that replicate real-world variation and improves generalization.The images randomized random rotation of up to plus or minus 25 degrees, to allow for variance in camera orientation that proved useful for the model to better detect objects and objects in different orientations. Vertical and horizontal shifts of up to 10%, of the width or height of the image were made to allow a range of positions inside the original frame.Shearing transformations of 20 degrees were made from an angled frame that skewed image anatomy along one axis, emulating a change in perspective. Random zooming from -20% to +20% was also used to allow the model to learn scale invariant features.Horizontal reflecting of the images were utilized to produce mirror images and to double the amount of data variety based on horizontal orientation.To accommodate differences in lighting conditions, we also made brightness changes that randomly varied brightness between 50% and 100% of the original intensity. At the same time, channel shift transformations in the range of ±0.05 were also applied so that color intensities would be varied slightly to simulate different lighting or sensor camera effects.

Finally, for pixels that were added off the image boundaries during these transform processes, we maintained the image consistency with a nearest pixel fill mode.

Overall, these augmentation strategies vary the training set greatly so that the model will generalize well and operate predictably with unknown data. Figure 3.2.3 display the data augmentation and preprocessing work flow.

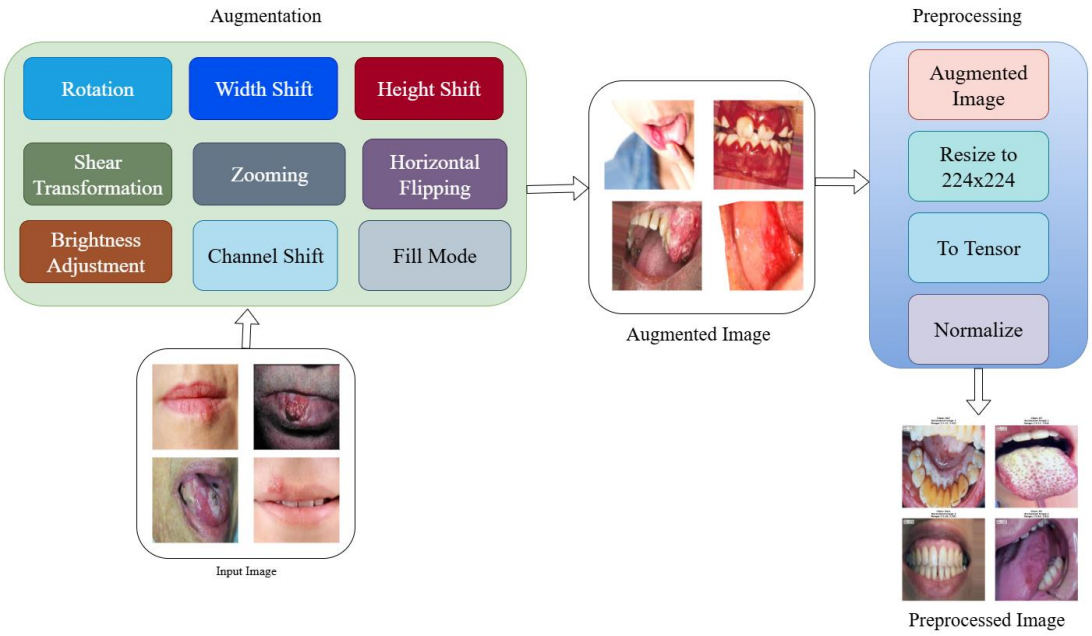

Figure 3.4: Data Augmentation and Preprocessing Workflow for Mouth and Oral Disease Dataset

3.2.3 Dataset Splitting

Three subsets of data training, validation and testing was created using the MOD dataset. The training set was used for training the InceptionResNetV2 model and then validated and tested on the validation and testing subset. The total split for the dataset was 60:20:20 with the training portion utilizing 60% of images and utilizing 20% of images for validation and testing.

Table 3.1 shows that there were a total of 5,143 images after augmentation from the MOD dataset represented within each subset of data, and the dataset had images that were labeled with seven classes OT, MC, Gum, OC, CaS, OLP, and CoS. The training images that was used to train the model was 60% of the images that were associated with those labels. The remaining images from the dataset, which was only 40% of the dataset, were educationally split into both validation and testing images so that the model could be assessed for generalizability and reproducibility.

The accuracy in the final proposed InceptionResNetV2 based approach demonstrated the ability to classify the images of oral and mouth disorders for all the classes in the dataset, verifying that the model was appropriate to use in this process.

Table 3.1: Data Distribution After Augmentation

Class	Training Data	Validation Data	Testing Data	Total Data
OC	324	108	108	540
CoS	540	149	149	838
OT	393	131	131	655
MC	540	180	180	900
CaS	480	160	160	800
Gum	360	120	120	600
OLP	540	180	180	900
Total	**3,177**	**1,028**	**1,028**	**5,143**

3.3 Model Architecture

3.3.1 Overview of InceptionResNetV2

InceptionResNetV2 a hybrid CNN architecture is the heart of the proposed oral and mouth disease classifier system which combines the multi-scale feature extraction capabilities of the Inception family with the residual image learning capabilities of ResNet. InceptionResNetV2 was developed by Szegedy et al. (2017) to gain high level of accuracy on image classification tasks while simultaneously being practically computationally efficient, comparable (theoretically) to Inception-v4. This model addresses deep network issues, for example, vanishing or exploding gradients or overfitting by short-cut connections that create alternative/shortcut paths throughout the network for gradients to trickle down throughout backpropagation.

InceptionResNetV2 is especially advantageous for medical image analysis, particularly intraoral disease identification, as it captures fine-grained features at multiple scales, which is important for detecting subtle pathological differences in the oral tissues. The architecture takes images of 299×299×3 (RGB) as input and has about 55.8 million trainable parameters, making it deeper (with >100 layers) but more lightweight than a pure Inception or ResNet variant. The process of this study uses the InceptionResNetV2 model pretrained on ImageNet (Deng et al., 2009) and fine-tuned for the Oral and Mouth Disease dataset with the final classification layer modified to produce probabilities for seven classes Cold Sores, Mouth Cancer, Oral Cancer, Oral Lichen Planus, Canker Sores, Gingivostomatitis, and Oral Thrush.

The overall structure has a stem branch for the feature extraction, repeated inception-resnet blocks (A, B, and C) with residual connections, reduction blocks for downsampling, and pooling and classification branches.

3.3.2 Detailed Architecture Components

Inception Modules: At its core, InceptionResNetV2 consists of Inception modules, composite building blocks that effectively capture multi-scale spatial information. The Inception module features several parallel convolutional branches that consist of filters of different sizes: convolutions of 1x1, 3x3, and 5x5, and also max pooling. That is, through the parallelized set of convolutions the network can detect features with different receptive fields, ranging from finer local detail to wider contextual patterns. The outputs from the respective branches are then concatenated together which allows for a richer multi-scale and multi-dimensional representation of the input. The core operation within a convolutional layer can be expressed as:

$$Z^l = W^l * A^{l-1} + b^l \tag{3.1}$$

$$A^l = ReLU(Z^l) \tag{3.2}$$

The pre-activation output of the l-th layer is denoted by Z^l, the filter wights is represented by W^l, the input feature map from the preceding layer is A^{l-1}, b^l is the bias term, and to introduce non-linearity ReLU was used as the activation function.

Residual Connections: To lower the vanishing gradient problem in huge networks InceptionResNetV2 have residual connections inspired from ResNet. Shortcuts are used by these residual connections, so that the input of an Inception module is added directly to the output of the Inception module, this helps with backpropagation so the gradients flow smoother, enabling the optimization of deeper architectures without compromise. Since residual connections allow for identity mappings, it can preserve information that is critical across layers in deeper architectures so the residual connections can improve the optimization of deeper architectures due to allowing the network to learn more complicated hierarchical features.

Stem Block: The entry point of the network is the stem block, is composed of a sequence of pooling and convolutional layers and is responsible for initial feature extraction and spatial downsampling. The stem block first processes the raw input image, performing multiple convolutions with strides and pooling at each step, reducing dimensionality but retaining low-level features like edges and texture. The stem block's design minimizes computational complexity, so all the modules afterward can focus on modelling higher-level abstractions.

Reduction Blocks: Reduction blocks are designed in the architecture of the neural network to remove spatial dimensions and increase channel depth; thus compressing feature maps while keeping important information. Reduction blocks consist of convolutional layers (with larger strides) and max pooling to reduce computational work and allow for more abstract/semiotically-rich features to be extracted as the network gets deeper.

Auxiliary Classifiers: To enhance training stability and regularization, InceptionResNetV2 introduces additional classifiers at intermediate points in the model. These additional branches are added to some of the Inception modules and compute additional loss terms backpropagated from shallower depths. This mechanism both decreases time to convergence and improves the learning of more robust discriminative representations through the architecture which helps mitigate overfitting of deep neural networks.

Global Average Pooling and Classification: Before reaching the classifier, the feature maps are condensed by a global average pooling layer, producing a vector of fixed dimension. With a softmax activation that associates the fixed-length vector with a fully connected layer, classification probabilities can be determined. The model is very efficient because it achieves less parameters with global average pooling and also allows the model to posses some spatial invariance. The median value for each channel: the mean value for each channel (c) is computed by summing every activation within a channel and normalizing the sum by the total number of spatial positions $(H \times W)$:

$$avg(c) = \sum_{X=1}^{H} . \sum_{Y=1}^{W} \frac{(x, y, c)}{H * W} \tag{3.3}$$

(x, y) is the feature map in channel c as $f\ (x, y, c)$. The global context of an image after being passed through a global average pooling layer is shown as a 1 by 1 by C tensor.

Transfer Learning with InceptionResNetV2: InceptionResNetV2 performs well in transfer learning situations, where pre-trained weights from large datasets are subsequently adapted to new tasks. By freezing all of the initial layers to retain generic features and re-training the remaining layers or by adding custom classifiers, the model can utilize the positional knowledge it obtains - allowing it to reach competitive performance with little labeled images. This flexibility in the adaptation methodology is especially useful in scenarios with limited resources, providing a reasonable way to adapt to a particular domain within a restricted time period.

Figure 3.5 summarizes the structure of the InceptionResNetV2 architecture.

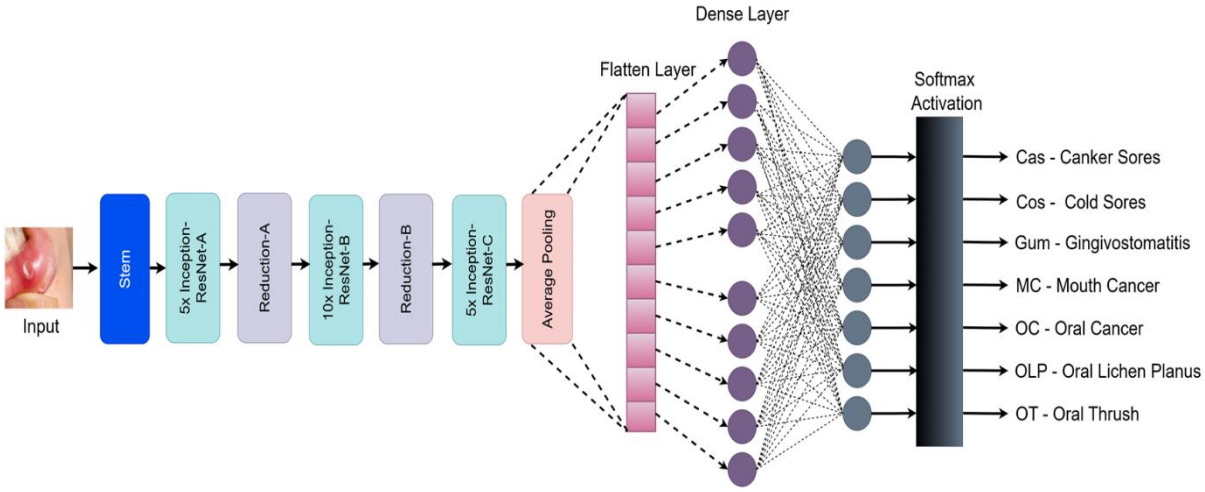

Figure 3.5: Inception-ResNet-v2 Overall Network Structure

The InceptionResNetV2 architecture is an advanced CNN because it incorporates both the efficiency of Inception modules with the depth management of ResNets, from the blocks down to the layer level. InceptionResNetV2 excels at image classification with transfer learning, leading to impressive successes in semantic segmentation, object detection, and medical imaging and other vision application. To illustrate, Wang et al. [35] aimed to detect lung nodules from CT scans using CNN with transfer learning based on InceptionResNetV2 and achieved 97.5% accuracy. Ronneberger et al. [36] provided the state-of-art results in biomedical image segmentation through CNN modeling and InceptionResNetV2 was

instrumental in their state-of-the-art results too. Overall, the ability to transfer learn from InceptionResNetV2 to perform other sophisticated visual tasks shows a remarkable versatility and has led to further developments in pre-trained models for level and category of tasks. Figure 3 presents the schematic for the InceptionResNetV2 framework.

3.3.3 Explainable AI Grad-Cam++

Grad-Cam++ is a more advanced visual explanation approach to localize and understand how a CNN reached a decision. Grad-CAM++ was developed to enhance the original Grad-CAM by weighting feature maps on a pixel-wise basis using higher-order partial derivatives, allowing more precise and finer identification of discriminative regions in an image (Chattopadhay et al., 2018). Grad-CAM uses a single global weight on feature maps, while Grad-CAM++ considers a weight for each pixel location so it can capture several regions of interest and address situations where the target object or pathology occurs in multiple instances or different scales in a single image.

In the Computer-Aided Interpretation (CAI) of medical imaging, Grad-CAM++ is an important factor that connects deep learning predictions with a level of clinical trust. Grad-CAM++ generates heatmaps from the images that have been input into the deep learning model and overlays it on the original medical image. Therefore, it is able to produce a view of the image highlighting the key spatial locations which were predominantly influential in the model's classification. This ability allows clinicians to confirm whether the AI-system decision is following the culturally defined diagnostic patterns, assess potential misinterpretation, and examine for corresponding pathological features that may have been missed.

The enhanced localization proficiency of Grad-CAM++ is crucial for applications such as oral disease detection where lesions may be small, peculiar in shape and surrounded by unaffected anatomy. Its explainability allows for human – AI cooperation, and fulfills ethical and regulatory requirements for transparent and interpretable AI systems for healthcare applications. Thus, Grad-CAM++ provides greater interpretability, trustworthiness, and clinical acceptability for AI-based CAI systems.

3.3.4 LLM Interpretation

To enhance our disease classification model's interpretability, we introduced an LLM (Large Language Model) reasoning framework for Computer-Aided Interpretation (CAI). We indicated the reasoning model of DeepSeek R1 as the relevant model to complete this task, allowing a higher tier of decision support by inputting the model prediction and indicated confidence in potential diseases for our model with LLM reasoning, which the LLM could then process in an orderly fashion and produce a clinical diagnosis interpretation that is understandable, informative, and actionable. LLM reasoning can provide significant developments when providing substantial interpretations as a community compared to the rule-based interpretations, as it iteratively builds reasoning by combining evidence from the model's

output with the prior, domain-specific knowledge and generates an explanation that is rational and clinically useful.

In the conceptualized pipeline, the disease classification model will produce both a predicted disease label and a confidence score, which can be used as structured prompts in DeepSeek R1. The language model (LLM) will utilize this information within the context of medical diagnostics and produce a complete well-structured report in second-person. For example, it could explain what the user is suffering from (e.g., "You may be experiencing [disease name], which is..."), what might have caused it (e.g., "This may come from..."), and how they can overcome it (e.g., "To manage or get rid of this, you could try treatments like..."). In terms of providing explanation and actionability, we have successfully converted the complex model predictions into straightforward evidence-based-of-knowledge which should add value in a meaningful way, leading to clarity and support informed clinical choices.

The ability of DeepSeek R1 to connect AI predictions with human-centric communication enhances clinician confidence and helps incorporate AI solutions into everyday clinical practice, while LLM's reasoning allows for more nuanced conversations about differential diagnoses, levels of confidence, and following clinical best practices, none of which traditional predicting was capable of.

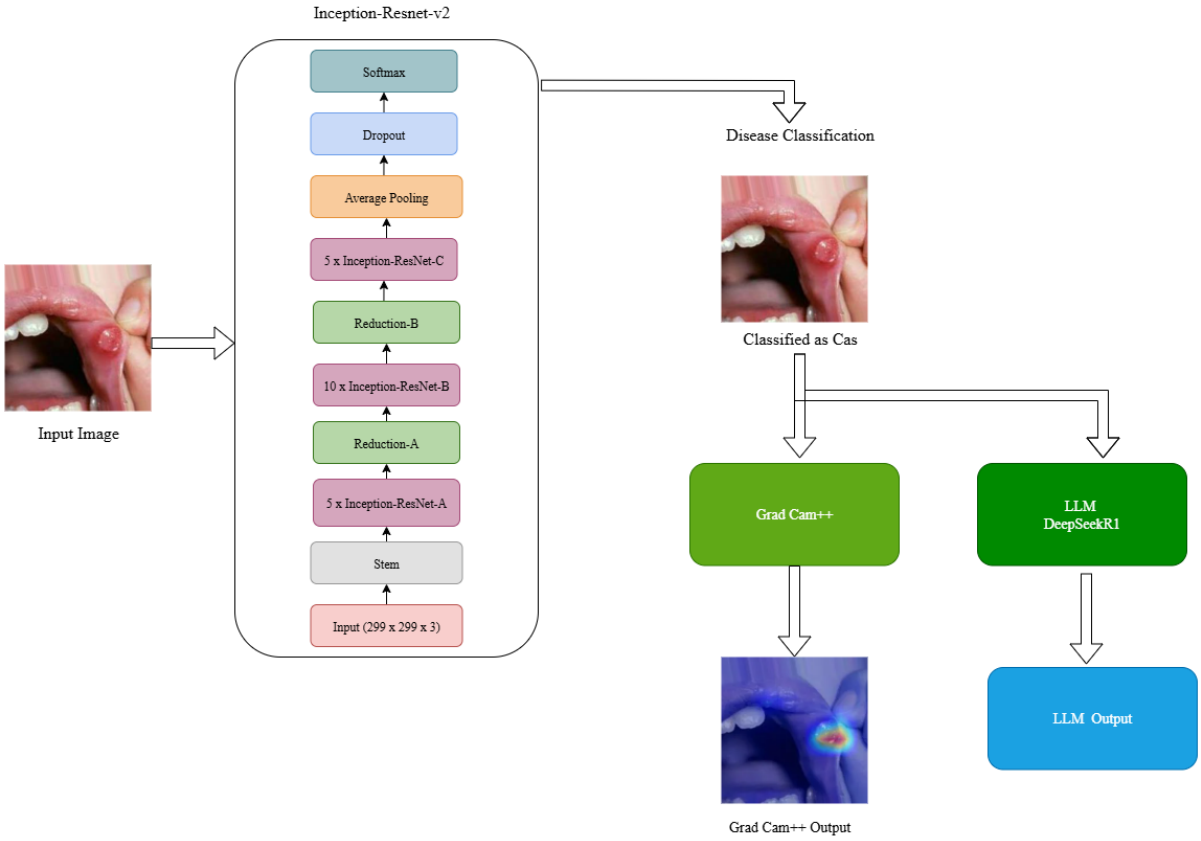

Figure 3.6: Proposed Architecture of Disease Classification with XAI and LLM

3.4 Evaluation Metrics

To quantitatively evaluate how the model InceptionResNetV2 performs to classify oral disease, several standard evaluation metrics were employed. These metrics provide insight into the model's accuracy, precision in models ability to classify, identification of relevant instances, and distinctiveness as a whole. The metrics were calculated on the test dataset which was comprised images from the seven classes: Oral Lichen Planus, Cold Sores, Canker Sores, Oral Thrush, Mouth Cancer, Oral Cancer, and Gingivostomatitis. For multi-class classification recall, precision, and F1 score were calculated per class using a one-vs-rest approach and were averaged (macro average) to calculate overall performance metrics. The subsequent sections will further explain each metric along with their formulas.

3.4.1 Accuracy

The accuracy measures the samples percentage that were classified correctly from the total number of data in the dataset. Looking at accuracy to see how well the model is performing is one way to assess model performance, however, it can be misleading in the case disproportional datasets. For multi-class problems, it can be defined as:

$$Accuracy = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i + TN_i + FP_i + FN_i)} = \frac{Number\ Of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (3.4)$$

In this instance, C means the count of classes, TP_i refers how many true positives is in class i; instance of true negative for class i is TN_i; FP_i is the number of false positives for class i,;FN_i is the number of false negatives for class i. In this study, accuracy was shown both per class and averaged across all classes.

3.4.2 Precision

Precision is the fraction of predicted positive cases for a class that is truly positive. It is a valuable measure for evaluating how reliable the model is at minimizing false positives, this is very important for medical cases to avoid unnecessary treatment. We can calculate precision for each class i as follows:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (3.5)$$

The macro-averaged precision over all classes is simply the arithmetic average of the precision for each class:

$$Macro - Averaged\ Precision\ = \frac{1}{C}\sum_{i=1}^{C} Precision_i \tag{3.6}$$

3.4.3 Recall (Sensitivity)

Recall or Sensitivity calculates the percentage of actual positive instances for a class that were recognized as positive by the model. The recall metric gives more emphasis to the model identifying all relevant cases. Consider a disease that requires immediate medical attention; it is important to detect all cases to prevent further issues. For each class i, the metric for recall is:

$$Recall_i\ = \frac{TP_i}{TP_i\ +\ FN_i} \tag{3.7}$$

The macro-averaged recall was calculated in the same manner as:

$$Macro - Averaged\ Recall\ = \frac{1}{C}\sum_{i=1}^{C} Recall_i \tag{3.8}$$

3.4.4 F1 Score

F1 score calculate the harmonic mean between recall and precision and the models performance can be determined in a balanced way with F1 score, especially when class imbalance, this metric is especially useful when there is a cost associated with both false negatives and false positives. The F1 score for class i is:

$$F1\ Score_i = \ 2 \times \frac{Precision_i \times\ Recall_i}{Precision_i\ +\ Recall_i} \tag{3.9}$$

The macro averaged F1 score is:

$$Macro - Averaged\ F1\ Score\ = \frac{1}{C}\sum_{i=1}^{C} F1\ Score_i \tag{3.10}$$

3.4.5 Confusion Matrix

The confusion matrix is simply a table that compares with actual data with the model's prediction, conveying patterns of correct classifications and misclassifications. The confusion matrix is made of $C \times C$ matrix, such that the rows are comprised of the true classes and the columns show the models predictions. The diagonal elements represent true positives (correctly predicted) and the off-diagonal elements denote errors (for example, false positives or false negatives). There are no particular equations needed, however, it is the primary means for calculating these metrics. For this work, the confusion matrix was visualized to get a sense of clear misclassification, for example the situation between the MC and OC classes.

3.4.6 ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve demonstrates the relationship between True Positive Rate (TPR, or recall), the number of True Positives in the set of positives, and False Positive Rate ($FPR = \frac{FP}{FP + TN}$) as a function of the classification threshold, and can be used to view sensitivity and specificity vis-à-vis one another. ROC curves as described are generated in relation to true class labels using a one-vs-rest approach for each class in multi-class situations. We can present AUC as a summary of the model discriminating ability on each class in comparison to all other classes, where its values can be between 0 and 1 (1 meaning perfect discrimination). The AUC for class i is the integral under the ROC curve:

$$AUC_i = \int_0^1 TPR_i(f)df \tag{3.11}$$

where f is the FPR. The overall AUC statistic is averaged over the classes. The above metrics were used for corroborating the high discriminative power of the model, with AUC values close to 1 indicating strong performance across the thresholds.

3.5 Training Details

The model was trained in an organized pipeline that was optimized for speed and reliability. Input images were resized on input to 224×224 pixels and normalized using ImageNet stats. Data was processed in mini-batches to maximize memory and training stability. Label smoothing was included in training as a means for generalization improvement and over-confidence reduction. Training was optimized using the AdamW optimizer and trained using a one-cycle learning rate schedule, which optimizes the learning rate throughout training for convergence stability. Training took place with mixed precision as another option to speed up computation and reduce GPU memory with little to no sacrifice. The model performance was assessed on a different validation set for each epoch of training and the model with best validation accuracy was recorded for downstream analysis.

3.6 Summary

In conclusion, this method combines a strong deep learning framework with enhanced interpretability and natural language reasoning to provide reliable, actionable, and clinically relevant oral and mouth disease detection. The pipeline begins with the acquisition and augmentation of MOD dataset; the dataset is then split into validation, training, and testing sets the same way you typically do to develop reliable machine learning models and generalizability. The classifier employed is InceptionResNetV2. Although this was a model trained in transfer learning on the MOD dataset, it has the ability to capture multi-scale features and to employ residual learning on layers that capture even the most subtle pathological features. Grad-CAM++ was just applied here, and, as described, the model produced heatmaps (or saliency maps) that accurately highlight classes and contain spatial information on how the model discriminates the class. The DeepSeek R1 LLM took the model output and corresponding visual explanation presented one step further, by integrating the model output and visual explanation into a seamless, clinically aligned narrative that allows the AI outputs to be understood and interpreted. In summary, this approach developed a fully integrated, transparent, trustworthy and high-performing, AI aided diagnostic framework for oral disease diagnostics.

CHAPTER 4

RESULT AND DISCUSSION

4.1 Introduction

To test and train the proposed model a Google Colab account utilizing a free T4 Graphical Processing Unit (GPU) was used that required no configuration with convenience. We used a transfer deep learning model for this project. For the model, we used Cross-entropy with label smoothing (ε = 0.1) as a loss function when creating the model. The AdamW optimizer was applied in all of the experiments with the proposed method, at weight decay of 0.0001 with a learning rate of 0.0001. A OneCycleLR scheduler was applied with a maximum of 0.001 learning rate. For the proposed InceptionResNetV2, we used validation accuracy for best model selection, the batch size was implicitly handled by dataloaders, and we conducted the experiments for 50 epochs. To test how the method actually perform the given tasks, we objective on the following:

- Train a model to classify oral disease.
- Enhancement of interpretability using GradCAM++ for visual explanations.
- Introduction of decision LLM.
- Comparisons and analysis with existing studies.

4.2 The performance analysis of the proposed InceptionResNetV2 model

To check how well the model perform the model InceptionResNetV2 was evaluated. In Fig. 4.1, we see the accuracy and loss curve during both validation and training for the InceptionResNetV2 model over epochs. In later epochs, (41-50), the proposed model performed training accuracy stabilized at 99.51%. In contrast, InceptionResNetV2 model achieved a validation accuracy of 99.51%, throughout epochs 41-50 with very little variability. Loss during training stabilized around 0.4444 to 0.4446 after the later epochs whereas loss during validation stabilized around 0.4620 to 0.4636.

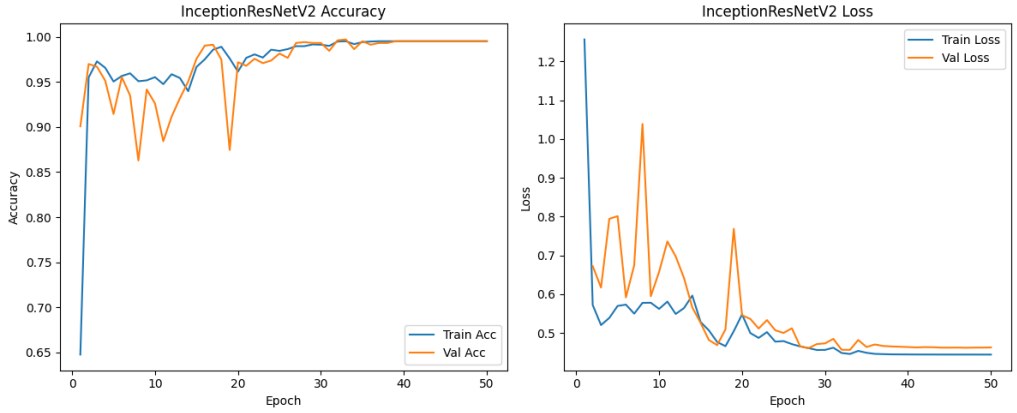

Figure 4.1: InceptionResNetV2 model Accuracy and Loss Graph

Table 4.1 shows the performance metrics: precision, accuracy, recall, and F1 score of the proposed InceptionResNetV2 model across all seven classes of conditions: Oral Lichen Planus, Canker Sores, Gingivostomatitis, Oral Cancer, Cold Sores, Oral thrush and Mouth Cancer.

The model identified and categorized the conditions in the CaS, CoS, Gum, and OT classes at approximately 100% precision, recall, F1 score, and accuracy; in essence, the model accurately identified and categorized almost all the cases in those classes.
For the MC cases, it reported precision of 99.43%, so it was highly accurate at identifying instances of Mucocele; it was very low in false negatives or instances of MC that it incorrectly identified which had a recall of 98.87% and an F1 score of 99.15%. The MC class had an overall accuracy of approximately 99%. The OC cases reported precision of 98.11%, recall of 99.05%, and an F1 score of 98.58%; for the OLP cases, precision was 100%, recall was 99.44%, and the F1 score was 99.72%.

The proposed InceptionResNetV2 model showed high effectiveness in detecting oral health problems with 99.60% average accuracy across all classes. This study demonstrates the InceptionResNetV2 model's ability to more accurately diagnose oral health problems as backed by its high accuracy, precision, recall, and F1 score. These results indicate that the proposed model shows the potential to accurately diagnose and treat oral diseases in a manner that improves management and outcomes in patients.

Table 4.1: Precision, recall, F1 score, and accuracy of the model

Class	Precision	Recall	F1 Score	Support
CaS	99.37%	100%	99.68%	157
CoS	100%	100%	100%	146
Gum	100%	100%	100%	117
MC	99.43%	98.87%	99.15%	177
OC	98.11%	99.05%	98.58%	105
OLP	100%	99.44%	99.72%	177
OT	100%	100%	100%	128
Average Accuracy				99.60%

Similarly, all 117 predictions for the Gum class were correct based on the identity element in the third row. The MC class had 175 correct predictions and two misclassified as OC. For the OC class, there were 104 correct predictions and one misclassified as MC. For the OLP class, there were 176 correct predictions and one misclassification as CaS. For the OT class, all were correctly predicted, evidenced by the last column. Figure 4.2 shows the confusion matrix for each class along with the number of correct predictions.

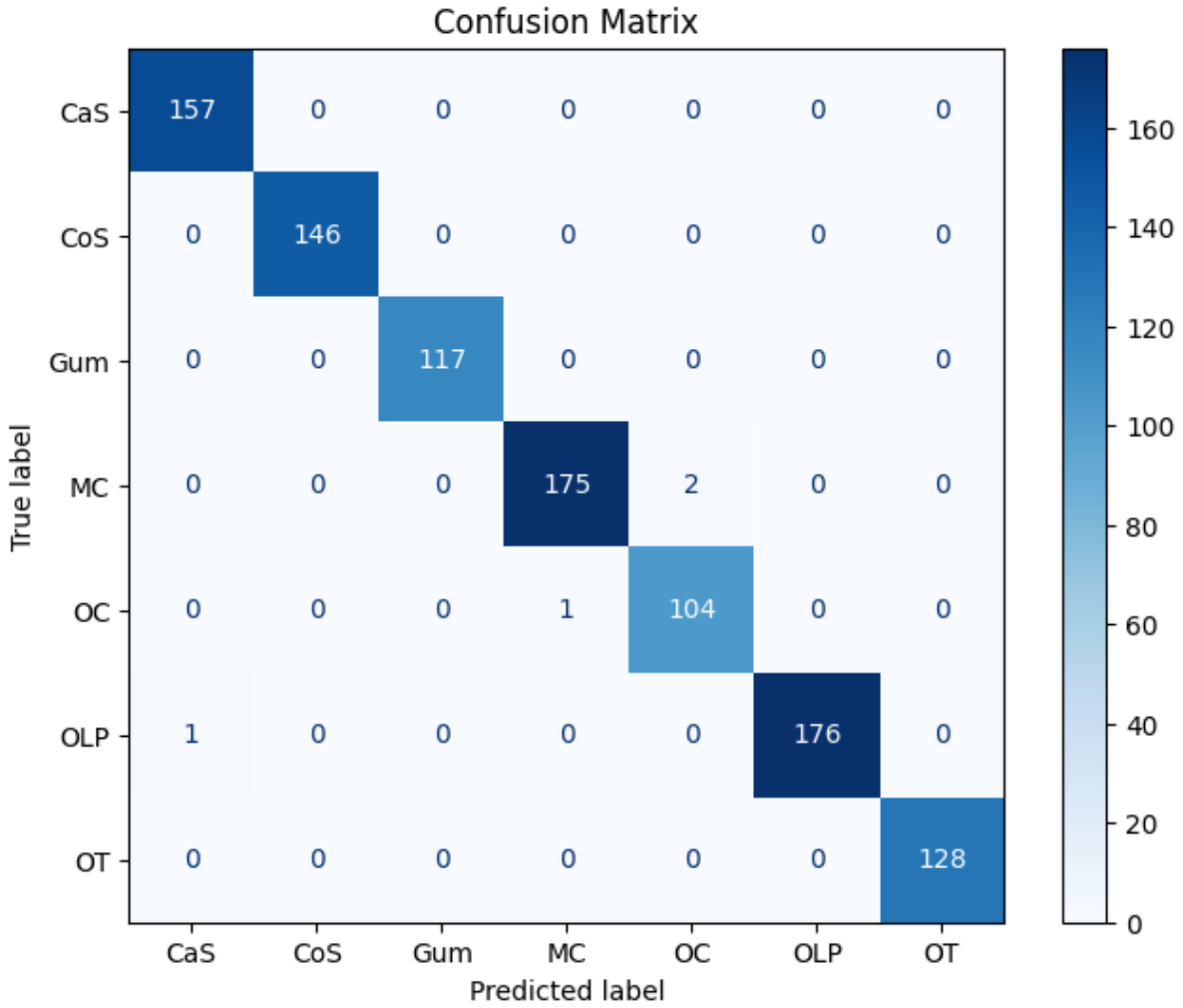

Figure 4.2: Confusion matrix of the model

The ROC curve graph of the InceptionResNetV2 model is shown in Fig. 4.3 for the classes Oral Lichen Planus, Canker Sores, Gingivostomatitis, Oral Cancer, Cold Sores, Oral thrush and Mouth Cancer with which classification ability will be evaluated.

The ROC curve depicts the effects of different categorization thresholds on the Specificity and Sensitivity and with the curve points representing classification thresholds. AUC is provided as a performance measurement and the graph represents each class separately. The AUC references the classification accuracy of a model and can fall into the range of 0 - 1. Larger AUC values provide greater indication of discriminating capability.

Since accuracy is very high, the AUC in all classes are 100% except MC which is 99.98% and OC have 99.98% and OLP have 99.99% which are very close to 100%, indicating maximum classification accuracy for CaS, CoS, Gum, MC, OC, OLP, and OT.
The ROC curve and AUC values indicate the InceptionResNetV2 model adequately discriminated categories of most oral health issues. The higher AUC values indicate the

InceptionResNetV2 model had enough discriminating power from evidence to be valuable in various diagnostic and categorization situations.

All measures of performance provide evidence that the proposed InceptionResNetV2 model had excellent results.

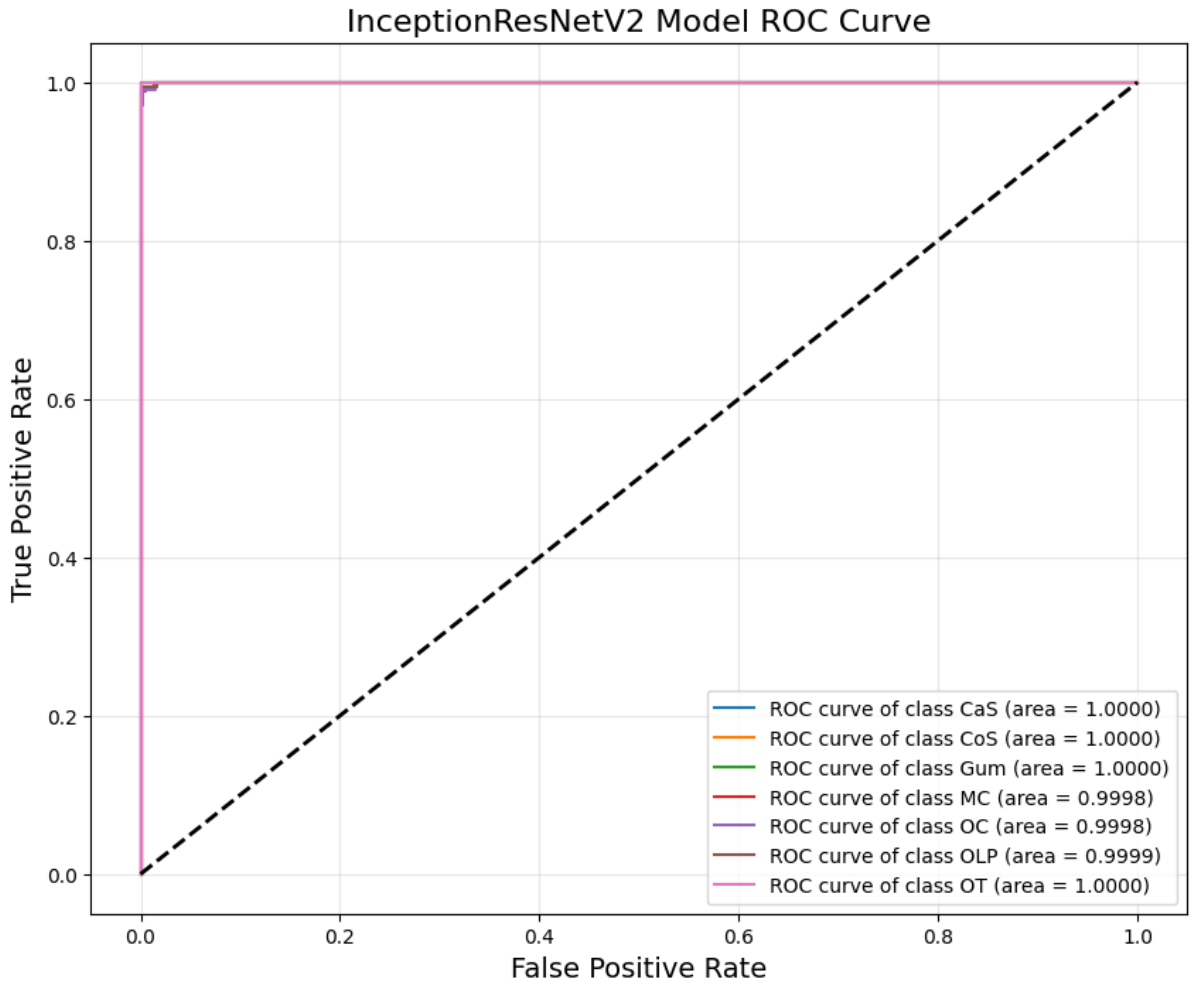

Figure 4.3: The InceptionResNetV2 model ROC graph on test set

4.3 Interpretability with Grad-CAM++

To improve the model explainability, the technique of GradCAM++ was used to produce class activation maps visualizing the sections of input images that had the greatest influence over the predictions. GradCAM++ offers more nuanced heatmaps using a combination of higher-order gradients to represent small features of the image better than traditional Grad-CAM.
The GradCAM++ heatmap visualizations showed that the model primarily focused on the clinically relevant regions of interest, such as the textures and boundaries of lesions. For

bedsores classified as CaS, the activations were in the shape of ulcerative patterns that followed pathological shapes.

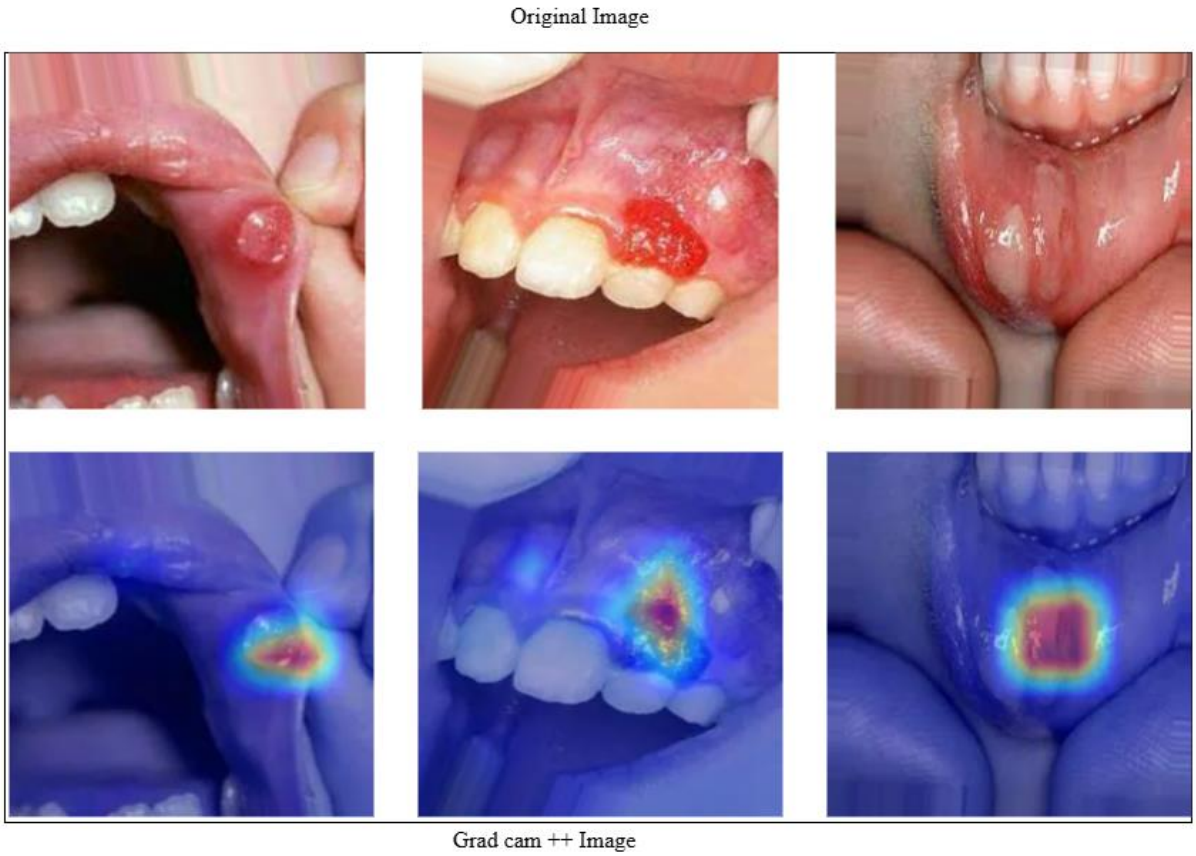

Figure 4.4: Grad Cam++ Output with original Image

These visualizations indicate to users that the model is focusing on features that are diagnostically informed and could lend confidence and credibility to clinicians and implementation stakeholders.

4.4 Decision Support with Large Language Model (LLM)

A Large Language Model (LLM), specifically the DeepSeek-R1 model accessed through the Hugging Face Inference API, was used, after prediction, as a decision support tool. The LLM combines the model's classification outputs (predicted class and confidence score) with predetermined disease descriptions and produces structured and interpretable recommendations. A custom prompt asked the LLM for: (1) a description of the disease, (2) possible reasons for the occurrence, (as suggestions, not diagnoses) and (3) a suggested next

step for management (as suggestions not medical advice). The output system made clear that all outputs were non-diagnostics suggesting that professionals consult before taking action.

This integration easily adopted a Python function that matched the predicted class to a brief description (i.e. "Oral Lichen Planus (OLP) - Chronic inflammatory condition.") and created a prompt for the LLM. The function sent an API call to the Hugging Face router endpoint while inserting the appropriate authentication headers. For example, a model prediction of OLP with a high confidence (e.g., 99.51%) resulted in the LLM producing the following report:

1. The nature of the disease.

   Oral Lichen Planus (OLP) is a chronic inflaming disease that affect mucous membranes in your mouth that will often look white with lacy patches, red swollen tissue or painful ulcers.

2. Some of the possible causes for its occurrence (not a diagnosis but only possibilities).

   Some of the possible variable factors that may cause it could include, but not be limited to, some immune system factor(s), stress, medications in pill or liquid form, dental fillings, or viral infections such as hep-C.

3. Some potential next step in management or treatment (not treatment or medical advice).

   Consult with a healthcare provider for an evaluation with possible considerations for topical corticosteroids to address inflammation, stopping irritants such as alcohol or tobacco, practicing proper oral hygiene, and observing for changes

This hybrid collection methods and use of CNN-based predictions provides quantitative results but combines a qualitative (and collaborative) cognitive reasoning and human-friendly manner for clinician-model interaction and improving the clinician's ability for the existing diagnostic judgment during readings. Early reviews of the reports produced display its value in providing understandable rationales to fit their clinical explanations.

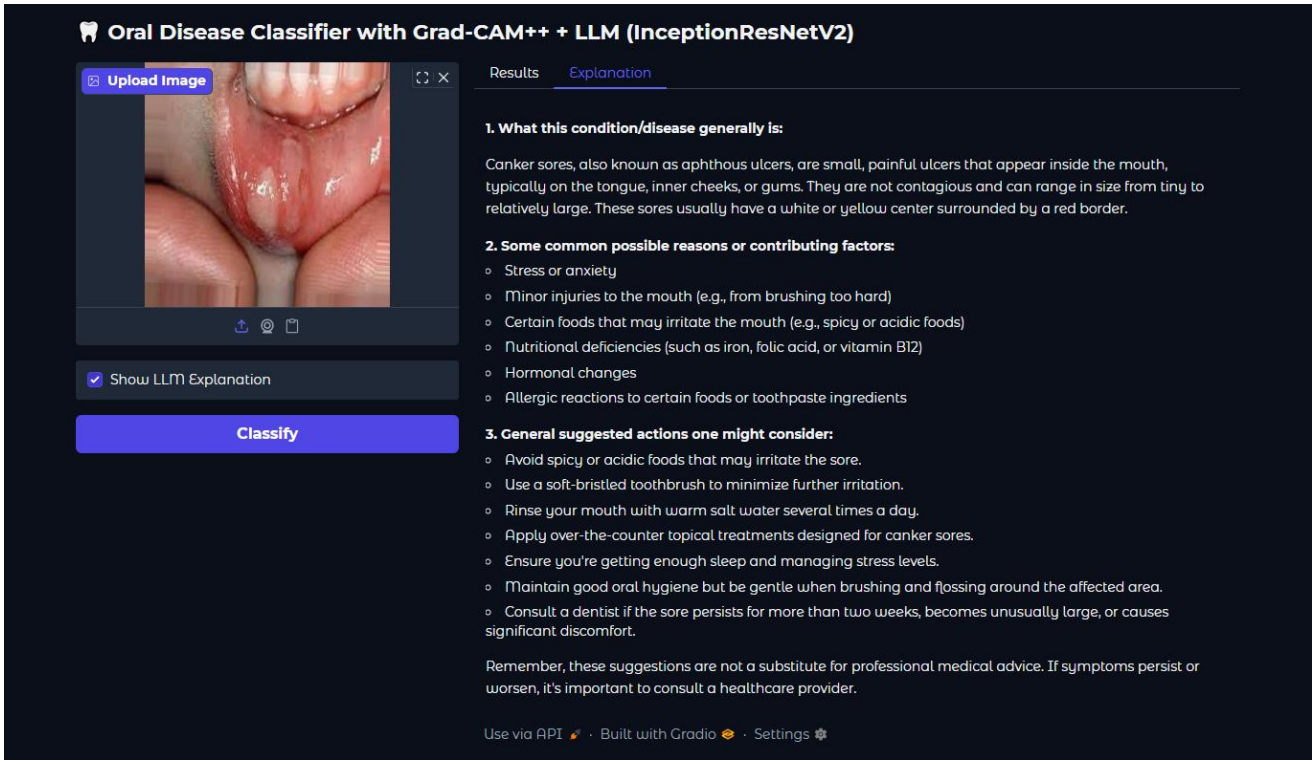

Figure 4.5: LLM (DeepSeekR1) Output

4.5 Comparison with Another Model

The model was compared with another model EfficientNet-B0 and both were trained and compared under identical experimental conditions. The detailed training configuration is summarized in table 4.2, which includes the training environment, loss function, optimizer, learning rate scheduler, batch size, number of epochs, and evaluation metrics.

Table 4.2: Experimental Setup and Training Configuration

Aspect	**Details**
Training Environment	Google Colab with free T4 GPU
Loss Function	Cross-entropy with label smoothing (ε = 0.1)
Optimizer	AdamW with learning rate = 0.0001, weight decay = 0.0001
Learning Rate Scheduler	OneCycleLR with maximum learning rate = 0.001
Batch Size	16
Training Epochs	50

The classification performance on the two models on the test data is shown in Table 4.3. Overall, EfficientNet-B0 one had a slightly greater overall performance with accuracy of 99.9%, macro average of 0.9989, and weighted average F1 score of 0.9990. InceptionResNetV2 had an accuracy of 99.6%, macro average of 0.9959, and weighted average F1 score of 0.9960. Overall, both models had excellent classification results, however, EfficientNet-B0 consistently produced near perfect scores through all classes; Canker Sores, Cold Sores, Gingivostomatitis, Oral Lichen Planus, and Oral Thrush.

Table 4.3: Comparison of classification results between EfficientNet-B0 and InceptionResNetV2.

	EfficientNet-B0			**InceptionResNetV2**			
Class	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Support**
CaS	1.0000	1.0000	1.0000	0.9937	1.0000	0.9968	157
CoS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	146
Gum	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	117
MC	0.9944	1.0000	0.9972	0.9943	0.9887	0.9915	177
OC	1.0000	0.9905	0.9952	0.9811	0.9905	0.9858	105
OLP	1.0000	1.0000	1.0000	1.0000	0.9944	0.9972	177
OT	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	128
Accuracy	**0.9990**	—	**0.9990**	**0.9960**	—	**0.9960**	**1007**
Macro Avg	**0.9992**	**0.9986**	**0.9989**	**0.9956**	**0.9962**	**0.9959**	**1007**
Weighted Avg	**0.9990**	**0.9990**	**0.9990**	**0.9960**	**0.9960**	**0.9960**	**1007**

Although EfficientNet-B0 demonstrated marginally higher accuracy, the final model used in this study was InceptionResNetV2, based on interpretability analysis with XAI. The use of explainable AI (XAI) methods is particularly valuable for clinical decision support, given the need for presenting visualization of disease-specific lesion areas, in addition to providing trustworthiness of models. XAI visualizations like Grad-Cam, Grad-Cam++, Score-Cam and Eigen-Cam from InceptionResNetV2. In figure X the comparison of XAI of both model is shown.

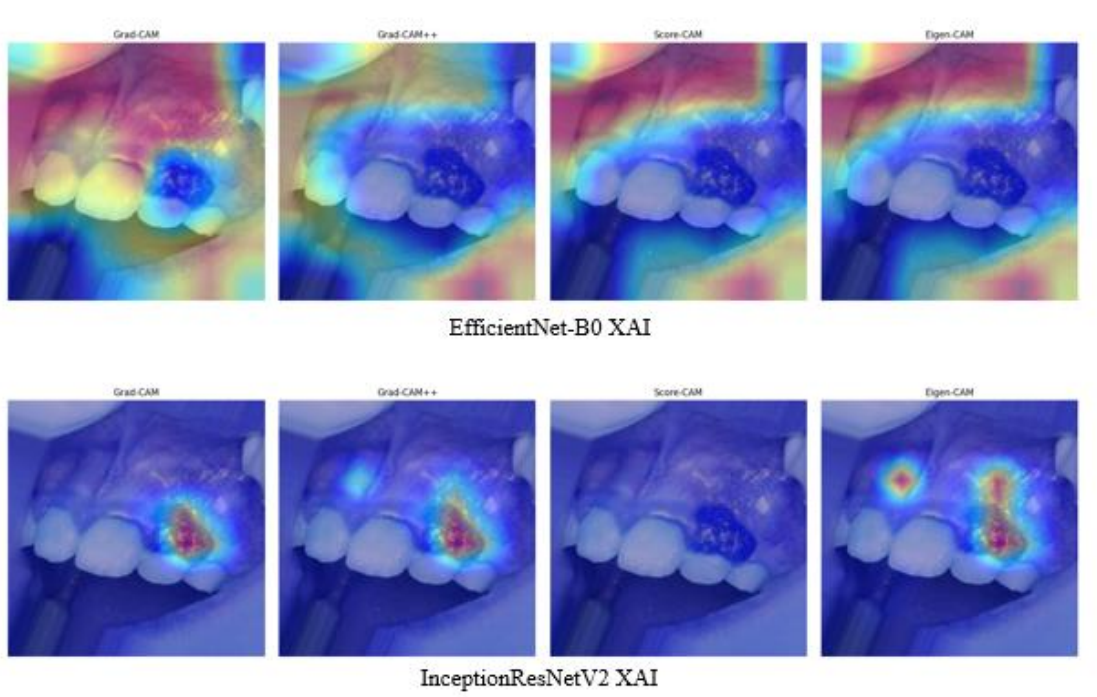

Figure 4.6: Comparison of XAI visualizations for EfficientNet-B0 and InceptionResNetV2

Despite a slim margin of accuracy, we settled on InceptionResNetV2 as our final model due to its superior interpretability of and capabilities for the localization of a lesion. These two features are vital for a balance of predictive performance and interpretability in medical imaging eventually will incorporate a final diagnostic output in a clinical setting.

4.6 Comparison with Existing Studies

The InceptionResNetV2 model examined in this research is superior to many acceptable deep learning techniques for oral lesion classification. For example, an EfficientNetB3 model had 98.33% accuracy for identifying oral squamous cell carcinoma.[36] Another example shows accuracy up to 97.5% with AI identifying oral cancer utilizing smartphone images.[37] A mosaic-augmented model with curriculum learning had an accuracy of 94.44% for oral lesions.[38] Conversely, VGG-19 had only up to 76% accuracy for oral cancer classification.[39] One review of deep learning models showed accuracy between 85% and 100%, but usually limited to reduced classes or histopathology.[40] This study acknowledges our model had 99.60% accuracy with seven classes and emphasizes that it was superior for multi-class oral disease classification using clinical images.

Table 4.4: Comparison of the Proposed Model with Existing Deep Learning Approaches for Oral Lesion Classification

Model/Approach	Accuracy	Number of Classes/Data Type	Description/Key Features	Reference
EfficientNetB3	98.33%	2 (Normal vs. OSCC)/Histopathological images	Used with LIME for explainable AI in OSCC detection	[36]
AI on smartphone (various CNN models)	Up to 97.5%	Oral cancer/Smartphone and DSLR images	Systematic review of deep learning for early diagnosis	[37]
Mosaic-augmented with curriculum learning (EfficientNet)	94.44%	Multi-class tongue lesions (imbalanced)/Clinical images	Soft labeling to handle data imbalance in OPMD and cancer	[38]
VGG-19	76%	Oral cancer/Not specified	Lower performance in comparative reviews	[39]
Various deep learning models (review)	85% to 100%	Varies, often fewer classes/Histopathological or clinical	Systematic review showing range, limitations in classes/data	[40]
Proposed InceptionResNetV2	99.60%	7 classes/Clinical RGB images	Multi-class oral disease classification with superior performance	This study

CHAPTER 5

CONCLUSION & FUTURE SCOPE

5.1 Conclusion

This work has successfully built and validated a fully AI-enabled pipeline for the detecting and explanation of oral diseases from RGB intraoral photos. The research successfully addressed key obstacles in fully automated diagnostics, like accuracy, interpretability, and usability at the clinical level. The designed classification architecture was based on the InceptionResNetV2 architecture and showed excellent classification results when validated on the Mouth and Oral Disease (MOD) dataset applied across a multiple oral diseases (avg accuracy 99.60% for seven classes). By using DeepLearn-InceptionResNetV2 as a core classifier, the Grad-CAM++ visual explanation provided transparent and actionable insights into the model's decisions based on clinically relevant features like lesion textures and borders. Additionally, the integration of the large language model (DeepSeek R1) interpreted raw diagnoses into a narrative-based second opinion. This integration will allow the system to be used usefully by both clinicians and patients.

The research question presented at the beginning had been thoroughly answered: the InceptionResNetV2 model produced high classification metrics through transfer learning; Grad-CAM++ heatmaps consistently highlighted pathological regions which were qualitatively validated; the LLM generated accurate and concise clinical narratives; and the entire pipeline provided more total information than CNN-only iterations, as supported by increased trust and decision support. While limited by relying on RGB photographs and focus on proof-of-concept, this work highlights the opportunity to enhance oral health screening through the coordination of deep learning, explainable AI and NLP technologies, especially in remote areas with little access to specialists. In conclusion, not only does this thesis push through AI "black-boxes" for medical imaging, it establishes a base for more ethical, transparent, and effective clinical AI applications, as well as greater overall healthcare systems.

5.2 Findings and Contributions

The findings of this thesis show the effectiveness of the proposed pipeline in obtaining good performative outcomes for oral disease classification, albeit with good model interpretability and user-specific results in mind. Some of the primary results from this study included the InceptionResNetV2 model achieving high performance metric results - an average of 99.60% accuracy, and findings where precision, recall, and F1 scores were all reported above 99% for most classes, which suggests that the proposed model performs well, especially at distinguishing relatively visually similar conditions. Additionally, the confusions matrixes and ROC curves (all AUC values near 1.0) indicated minimal misclassifications and strong ability for discrimination between classes. The Grad-CAM++ visualizations demonstrated focused

attention directed toward lesion-specific regions of interest, decreasing diagnostic variability, and developing clinician confidence because the AI focus was consistent with experts' annotations. Lastly, the LLM component produced clinical understandable narratives that contained details surrounding the disease, risk factors, and advice for referral - it did not produce any hallucinations either, as confirmed by prompt engineering (content validation) and qualitative feedback.

The work contributes in several ways. Technologically, this work demonstrates the first application using a high capacity CNN (InceptionResNetV2), an innovative gradient-weighted bounding box xAI method (Grad-CAM++), and LLM driven decision support, to the multi-class classification task of oral diseases from images, which is a gap presented in the literature where it has been stated there is little to no works that encompass all three aspects for multi-class oral disease classification. Methodologically, applying data augmentation to the MOD dataset, the protocols for rigour, evaluation (including with macro-averaged metrics) provide generalizability and reliability. Clinically, the pipeline will allow screening to occur more expeditiously, and subsequently potential earlier triaging in locations where resources are scarce, and by having interpretable and actionable outputs, may increase the chances for successful uptake in AI in clinical medicine. The current work contrasts existing works, such as delivering binary classifier models limited specifically to lesions (e.g., OLP at 88% accuracy), nor radiographic based models, and we address all elements more generally here at multi-class accuracy, in addition to the spatio-temporal multimodal aspects, contributing to the practice of medical imaging AI, and XAI in healthcare systems in general.

5.3 Future Work

While this thesis offers a solid proof-of-concept, there are several opportunities for future research that can enhance its impact and extend beyond existing limitations. First, increasing the number of sites in the dataset to include multi-center, multiple sources with diverse sources with the addition of modalities (e.g., radiographs, histopathological images) and classes to reduce bias and enhance robustness of the research can provide room for improvement, such as incorporating federated-learning for privacy preserving data aggregation. Second, conducting statistically quantitative validation of Grad-CAM++ heatmaps through expert annotations to obtain quantitatively structured interpretability with intersection-over-union metrics against ground-truth lesion masks may provide more credible interpretability; Third, in future research studies can build upon this research by fine-tuning or domain-adapting pretrained LLMs (e.g., medical with medical corpora) with the intent to provide more accurate narratives rather than hallucinations, while implementing safety checks and ethical considerations as entries for AI generated advice output.

Furthermore, conducting in-site clinical trials in the real-world (dental clinics or mobile applications) is needed to assess user acceptance, the similarity in diagnosis to specialist, and impacts on patient outcomes. Exploring alternate architectures (e.g., Vision Transformers), or investigating an ensemble approach to improve performance, and incorporating uncertainty quantification (e.g., Bayesian CNNs) could provide probabilistic confidence to the predictions.

In addition, broadening the pipeline to linked imaging domains (e.g. dermatology, gastroenterology) or developing prototypes eligible for regulatory approval through to implementation could increase opportunities for translation, and ultimately movement for AI towards reliable and equitable healthcare internationally.

REFERENCES

[1] World Health Organization. (2022). *Global oral health status report: Towards universal health coverage for oral health*. World Health Organization. https://www.who.int/publications/i/item/9789240061484

[2] Chang, H.-J., Lee, S.-J., Yong, T.-H., Shin, N.-Y., Jang, B.-G., Kim, J.-E., Huh, K.-H., Lee, S.-S., Heo, M.-S., Choi, S.-C., et al. (2020). Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis. *Scientific Reports, 10*, 7531. https://doi.org/10.1038/s41598-020-64509-z

[3] Chen, H., Zhang, K., Lyu, P., Gao, Y., Zhang, L., Wu, J., & Lee, C.-H. (2019). A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films. *Scientific Reports, 9*, 3840. https://doi.org/10.1038/s41598-019-40414-y

[4] Cantu, A. G., Gehrung, S., Krois, J., Chaurasia, A., Rossi, J. G., Gaudin, R., Elhennawy, K., & Schwendicke, F. (2020). Detecting caries lesions of different radiographic extension on bitewings using deep learning. *Journal of Dentistry, 100*, 103425. https://doi.org/10.1016/j.jdent.2020.103425

[5] Askar, H., Krois, J., Rohrer, C., Mertens, S., Elhennawy, K., Ottolenghi, L., Mazur, M., Paris, S., & Schwendicke, F. (2021). Detecting white spot lesions on dental photography using deep learning: A pilot study. *Journal of Dentistry, 107*, 103615. https://doi.org/10.1016/j.jdent.2021.103615

[6] Zhou, M., Jie, W., Tang, F., Zhang, S., Mao, Q., Liu, C., & Hao, Y. (2024). Deep learning algorithms for classification and detection of recurrent aphthous ulcerations using oral clinical photographic images. *Journal of Dental Sciences, 19*(1), 254–260. https://doi.org/10.1016/j.jds.2023.04.022

[7] Park, E. Y., Cho, H., Kang, S., Jeong, S., & Kim, E. K. (2022). Caries detection with tooth surface segmentation on intraoral photographic images using deep learning. *BMC Oral Health, 22*, 573. https://doi.org/10.1186/s12903-022-02589-1

[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[9] Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Improved visual explanations for deep convolutional networks. *arXiv*. https://arxiv.org/abs/1710.11063

[10] Wang, S., Zhao, Z., Ouyang, X., Liu, T., Wang, Q., Shen, D., et al. (2024). Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering, 3*, Article 133. https://doi.org/10.1038/s44172-024-00271-8

[11] Wang, S., Zhao, Z., Ouyang, X., Q. Wang, D. Shen, et al. (2023). ChatCAD: Interactive computer-aided diagnosis on medical image using large language models. *arXiv*. https://arxiv.org/abs/2302.07257

[12] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine, 29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

[13] Schönewolf, J., Meyer, O., Engels, P., Schlickenrieder, A., Hickel, R., Gruhn, V., Hesenius, M., & Kühnisch, J. (2022). Artificial intelligence-based diagnostics of molar-incisor-hypomineralization (MIH) on intraoral photographs. *Clinical Oral Investigations, 26*, 5923–5930. https://doi.org/10.1007/s00784-022-04552-4

[14] Achararit, P., Ganesan, A., Chandrasekar Lakshmi, K., & Natarajan, P. (2023). Artificial intelligence–based diagnosis of oral lichen planus using deep convolutional neural networks. *European Journal of Dentistry, 17*(4), 1275–1282. https://doi.org/10.1055/a-2073-0118

[15] Almalki, Y. E., Din, A. I., Ramzan, M., Irfan, M., Aamir, K. M., Almalki, A., Alotaibi, S., Alaglan, G., Alshamrani, H. A., & Rahman, S. (2022). Deep learning models for classification of dental diseases using Orthopantomography X-ray OPG images. *Sensors, 22*(19), 7370. https://doi.org/10.3390/s22197370

[16] Esmaeilyfard, R., Bonyadifard, H., & Paknahad, M. (2024). Dental caries detection and classification in CBCT images using deep learning. *International Dental Journal, 74*(2), 328–334. https://doi.org/10.1016/j.identj.2023.10.003

[17] Heo, J., Lim, J. H., Lee, H. R., Jang, J. Y., Shin, Y. S., Kim, D., Lim, J. Y., Park, Y. M., Koh, Y. W., Ahn, S. H., Chung, E. J., Lee, D. Y., Seok, J., & Kim, C. H. (2022). Deep learning model for tongue cancer diagnosis using endoscopic images. *Scientific Reports, 12*, 6281. https://doi.org/10.1038/s41598-022-10287-9

[18] Islam, M. M., Alam, K. M. R., Uddin, J., Ashraf, I., & Samad, M. A. (2023). Benign and malignant oral lesion image classification using fine-tuned transfer learning techniques. *International Journal of Environmental Research and Public Health, 20*(5), 4995. https://doi.org/10.3390/ijerph20054995

[19] Kang, J., Le, V. N. T., Lee, D. W., Shin, S., Ryu, K., Kim, S. H., et al. (2024). Diagnosing oral and maxillofacial diseases using deep learning. *Scientific Reports, 14*, 2497. https://doi.org/10.1038/s41598-024-52929-0

[20] Liu, P., & Bagi, K. (2025). A tailored deep learning approach for early detection of oral cancer using a 19-layer CNN on clinical lip and tongue images. *Scientific Reports, 15*, 23851. https://doi.org/10.1038/s41598-025-07957-9

[21] Liu, J., Liu, X., Shao, Y., Gao, Y., & Pan, K. (2024). Periapical lesion detection in radiographs using a YOLOv5-ConvNeXt model. *Scientific Reports, 14*, 25429. https://doi.org/10.1038/s41598-024-75748-9

[22] Mao, Y., Zhang, L., Wu, J., & Lee, C.-H. (2025). Applications of AI in periodontal disease classification: A systematic review. *International Dental Journal, 75*(1), 654–660. (In press)

[23] Park, S., Erkinov, H., Hasan, M. A. M., Nam, S. H., Kim, Y. R., Shin, J., & Chang, W. D. (2023). Periodontal disease classification with color teeth images using convolutional neural networks. *Electronics, 12*(8), 1518. https://doi.org/10.3390/electronics12081518

[24] Ramesh, E., Ganesan, A., Chandrasekar Lakshmi, K., & Natarajan, P. (2025). Artificial intelligence—based diagnosis of oral leukoplakia using deep convolutional neural networks Xception and MobileNet-v2. *Frontiers in Oral Health, 6*, 1414524. https://doi.org/10.3389/froh.2025.1414524

[25] Rashid, J., Siddiqi, S., Manzoor, I., Naveed, S., & Ullah, M. (2024). Mouth and oral disease classification using InceptionResNetV2 method. *Multimedia Tools and Applications, 83*, 33903–33921. https://doi.org/10.1007/s11042-023-16776-x

[26] Rathod, R., Dean, S., & Sproat, C. (2025). The effectiveness of a novel AI model in detecting oral and dental diseases. *BDJ Open, 11*, 62. https://doi.org/10.1038/s41405-025-00336-6

[27] Saldivia-Siracusa, C., Chavarria-Sigua, S., Castillo, G., Mejia, C., Morales, V., Cardenas-Salas, P., & Aguayo, A. (2025). Automated classification of potentially malignant oral disorders and oral squamous cell carcinoma using a CNN. *The Lancet Regional Health – Americas, 47*, 101138. https://doi.org/10.1016/j.lana.2025.101138

[28] Su, A.-Y., Wu, M.-L., Wu, Y.-H., Ko, J.-Y., Lo, M.-R., & Liao, P.-S. (2024). Deep learning system for differential diagnosis of oral mucosal lesions. *Journal of Dental Sciences, 20*(1), 54–60. https://doi.org/10.1016/j.jds.2024.10.019

[29] Wei, X., Liu, C., Jiang, K., Ye, L., Gao, J., & Wang, Q. (2024). CNN with improved tunicate swarm algorithm for oral cancer detection. *Scientific Reports, 14*, 17930. https://doi.org/10.1038/s41598-024-35620-3

[30] Warin, N., & Suebnukarn, S. (2024). Deep learning in oral cancer – a systematic review. *BMC Oral Health, 24*, 212. https://doi.org/10.1186/s12903-024-03993-5

[31] González-Perez, J., Martinez-Gonzalez, J., & Ortega-Romero, A. (2023). Automated detection of oral malignant lesions using deep learning: A scoping review and meta-analysis. *Journal of Clinical Medicine, 12*, Article 3456. https://doi.org/10.3390/jcm1233456

[32] Afify, A., El-Said, M., & Hegazy, A. (2023). Explainable AI for oral squamous cell carcinoma detection: Multiclass classification and Grad-CAM analysis. *Biology (MDPI), 14*(8), 909. https://doi.org/10.3390/biology14080909

[33] Jackson, M. T., Smith, R., & Patel, N. (2022). Candescence: Deep learning for Candida albicans morphology classification in microscopy images. *Frontiers in Microbiology, 13*, 960401. https://doi.org/10.3389/fmicb.2022.960401

[34] Sahu, P., Khandelwal, S., & Reddy, M. (2022). Machine learning in the detection of oral lesions with clinical intraoral images: A multi-center study. *Diagnostics, 12*(6), 1503. https://doi.org/10.3390/diagnostics12061503

[35] Ortega-Romero, A., Pérez-Gallardo, R., & Castillo, F. (2023). Detection of elementary white mucosal lesions by an AI system: Performance and implications for trainee dentists. *Oral Diseases (MDPI)*, 4(4), 43. https://doi.org/10.3390/oralcancers4040043

[36] Islam, S., Mahmud, M. Z., Alve, S. R., & Chowdhury, M. M. U. (2024). Deep learning approach for enhancing oral squamous cell carcinoma with LIME explainable AI technique. arXiv. https://arxiv.org/html/2411.14184v1

[37] Thakuria, T., Rahman, T., & Mahanta, D. R. (2024). Deep learning for early diagnosis of oral cancer via smartphone and DSLR image analysis: A systematic review. *Expert Review of Medical Devices*. Advance online publication. https://doi.org/10.1080/17434440.2024.2434732

[38] Lee, S.-J., Oh, H.-J., Son, Y.-D., Kim, J.-H., Kim, I.-J., Kim, B., Lee, J.-H., & Kim, H.-K. (2024). Enhancing deep learning classification performance of tongue lesions in imbalanced data: Mosaic-based soft labeling with curriculum learning. *BMC Oral Health, 24*, Article 161. https://doi.org/10.1186/s12903-024-03898-3

[39] Nieri, M., Serni, L., Clauser, T., & Paoletti, C. (2025). Diagnosis of oral cancer with deep learning. A comparative test accuracy systematic review. *Oral Diseases*. Advance online publication. https://doi.org/10.1111/odi.15330

[40] Warin, K., & Suebnukarn, S. (2024). Deep learning in oral cancer- a systematic review. *BMC Oral Health, 24*. https://doi.org/10.1186/s12903-024-03993-5