AI Assisted Screening of Oral Potentially Malignant Disorders Using Smartphone Photographic Images - An Indian Cohort Study

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

Vivek Talwar
2020701034
vivek.talwar@research.iiit.ac.in

International Institute of Information Technology Hyderabad
500 032, India

June 2025

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis proposal titled ***AI Assisted Screening of Oral Potentially Malignant Disorders Using Smartphone Photographic Images - An Indian Cohort Study*** by *Vivek Talwar* has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. P.K. Vinod

_____ _____

Date Advisor: Prof. C.V. Jawahar

Acknowledgment

I would like to begin by expressing my sincere gratitude to my advisor, Prof. C.V. Jawahar. His guidance has paved the way for me to develop a research-oriented mindset and has imparted invaluable lessons on the importance of commitment and perseverance in research—lessons I am confident will profoundly shape my future pursuits. I also extend my thanks to my co-advisor, Prof. Vinod P.K., for his insightful suggestions and for encouraging me to tackle problem statements with significant medical relevance; his contributions have greatly enhanced the quality of my work.

I feel incredibly fortunate to have the unwavering support and motivation of my friend and colleague, Dr. Pragya Singh. Her constant encouragement and expertise in the data aspects of the project have made this experience truly enriching. As an oral cancer specialist, she was my primary resource for resolving data queries, and her enthusiasm infused my master's and research journey with excitement and zeal. I am equally grateful to my friend Kushal Borkar, with whom I began my master's studies; his support, motivation, and late-night discussions inspired me to question untested ideas and improve my experimental results. My journey also included many engaging discussions with members of the Robotics Research Lab—Laksh, Avneesh, and Sanket—whose camaraderie and banter created countless fond memories. I likewise cherish the friendships I formed with junior bachelor's students who look to me as a mentor; I am grateful for the opportunity to offer them valuable insights into their career trajectories.

I extend my sincere thanks to the seniors at the CVIT Lab for their thoughtful reviews and for sharing their wealth of knowledge and experience. I also appreciate my other lab mates and colleagues at CVIT, with whom I have enjoyed invaluable discussions across technical, philosophical, and personal domains. Finally, I acknowledge the unwavering support and love of my family, without which this work would not have been possible. I am especially thankful to my father and all my family members for encouraging every decision I have made, beginning with my choice to pursue my master's at IIIT Hyderabad.

To Papa

Abstract

The escalating prevalence of Oral Potentially Malignant Disorders (OPMDs) and oral cancer in low- and middle-income countries presents a critical challenge, exacerbated by limited resources that hinder population screening in remote areas. The study evaluates the efficacy of artificial intelligence (AI) and digital imaging diagnostics as tools for OPMD detection in the Indian population, utilizing smartphone-captured oral cavity images. Trained front-line healthcare workers (FHWs) contributed a dataset comprising 1,120 suspicious and 1,058 non-suspicious images. Various deep-learning models, including DenseNets and Swin Transformers, were assessed for image-classification performance. The best-performing model was then tested on an independent external set of 440 images collected by untrained FHWs. DenseNet201 and Swin Transformer (base) models exhibited high classification accuracy on the internal test set, achieving F1-scores of 0.84 (CI 0.79–0.89) and 0.83 (CI 0.78–0.88). However, performance declined on the external set—characterized by significant variation in image quality—with DenseNet201 yielding the highest F1-score of 0.73 (CI 0.67–0.78). The AI model demonstrates potential for identifying suspicious versus non-suspicious oral lesions via photographic images. This image-based solution holds promise for facilitating early screening, detection, and timely referral for OPMDs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, the prevalence of oral potentially malignant disorders (OPMDs) [1] has emerged as a significant public health concern, particularly in low- and middle-income countries (LMICs). These conditions, characterized by cellular abnormalities in the oral mucosa that can progress to cancer, pose a substantial burden on healthcare systems worldwide. The escalating incidence of OPMDs [2] underscores not only the urgency of effective preventive and diagnostic strategies but also the disparities in healthcare access, especially in remote and resource-constrained regions.

OPMDs serve as crucial precursors to oral cancer, and early detection and intervention are pivotal in mitigating their progression to malignancy [3]. Despite significant advancements in medical technology, persistent challenges remain in implementing comprehensive, standardized, and timely screening programs, especially in regions where healthcare infrastructure and resources are limited. Consequently, there is a pressing need for innovative, cost-effective, and accessible approaches to facilitate the early identification of OPMDs, enabling timely intervention and improving patient outcomes.

1.1 Cancer Diagnosis

Cancer [4] is characterized by the abnormal, uncontrolled proliferation of cells originating from specific tissues and organs. Although cancers differ by cell type and origin, they all arise from unregulated growth of abnormal cells. Furthermore, cancer cells can promote angiogenesis to establish their own blood supply, detach from their tissue of origin, invade surrounding tissues, migrate through the bloodstream or lymphatic system, and disseminate to distant organs. Morbidity and mortality result from organ dysfunction caused by both local tumor expansion and metastatic spread. The development of cancer involves numerous genetic and epigenetic alterations influenced by multiple factors, including lifestyle choices such as tobacco use and diet, genetic inheritance through familial mutations, and exposure to environmental carcinogens. In some cases, the precise cause of cancer remains unclear.

1.1.1 Incidence Rates

Cancer incidence rates [5] can vary significantly across countries, influenced by diverse factors such as population demographics, lifestyle, genetics, and healthcare infrastructure. In the United States [6],

cancer remains a major public health concern, with high overall incidence rates and robust screening programs. Common cancers include breast, lung, colorectal, and prostate malignancies. The prevalence of specific cancer types often correlates with an aging population, lifestyle behaviors such as tobacco use and diet, and advanced healthcare systems that facilitate early detection and diagnosis.

By contrast, India exhibits a distinct cancer profile with a different mix of prevalent tumors. Although it bears a substantial burden of breast, cervical, and colorectal cancers, India also faces a higher incidence of malignancies linked to infectious agents such as oral and skin cancers. Contributing factors include socio-economic disparities, infectious disease prevalence, and lifestyle practices.

These contrasts underscore the importance of considering regional variations and unique risk factors when addressing cancer globally. Tailored prevention, screening, and treatment strategies are essential to effectively combat the diverse challenges that cancers pose across countries.

1.1.2 Cancer Prevalence in India

Cancer prevalence in India reveals a multifaceted landscape, with a pronounced focus on oral and skin malignancies, especially in remote and resource-constrained regions. Oral cancer—which encompasses cancers of the mouth, tongue, and throat—is highly prevalent in India, accounting for a significant proportion of the country's overall cancer burden. This high incidence is often linked to widespread tobacco and betel-nut use, combined with low awareness, socioeconomic disparities, and limited access to preventive healthcare. Similarly, skin cancer—particularly non-melanoma types—is on the rise in India, driven by factors such as prolonged ultraviolet (UV) exposure and shifting lifestyle habits. The predominance of outdoor occupations and inadequate sun-protection practices further elevate skin-cancer risk. Addressing these challenges demands robust, comprehensive public-health campaigns to educate the population, primary-care providers, and healthcare workers about risk factors, promote early detection, and implement effective, cost-effective prevention, screening, and treatment measures, thereby reducing the impact of oral and skin cancers nationwide.

1.1.3 Procedure for Diagnosis

When patients present with symptoms suggestive of cancer or when screening test results raise concern, medical professionals typically delve deeper to ascertain whether the condition is truly malignant. This investigative phase commonly includes a review of the patient's family medical history, and doctors may order laboratory tests, imaging procedures (such as scans), or other diagnostic assessments. Following these initial steps, a biopsy is often required as the only crucial method to definitively confirm malignancy. The end-to-end detailed procedure includes the following:

1.1.3.1 Imaging Tests

Medical professionals utilize various non-invasive radiological imaging techniques to examine internal body areas and identify the presence of tumors. These imaging modalities include CT scans, PET

scans, X-rays, bone scans, and MRI scans. [7] In some cases, individuals may be required to ingest or receive contrast materials intravenously before the procedure, enhancing the visibility of tumors in the resulting images. This aids doctors in obtaining detailed and accurate pictures to assess and diagnose potential tumors within the body.

1.1.3.2 Biopsy

The essential step in cancer diagnosis often involves a biopsy. During this procedure, a medical professional extracts a small tissue sample for examination. Subsequently, a pathologist meticulously scrutinizes the tissue under a microscope, conducting various tests and analyses to determine whether cancer is present. The findings of the pathologist are consolidated in a comprehensive pathology report, providing crucial details about the diagnosis. These pathology reports play a pivotal role in the diagnostic process and contribute significantly to informing decisions about appropriate treatment options. Here are several common ways a biopsy sample can be obtained:

- Using a needle, medical practitioners perform procedures to extract tissue or fluid, commonly employed for tasks such as bone marrow aspirations, spinal taps, as well as biopsies for the breast, prostrate and liver.

- Endoscopy involves the use of slender , illuminated tube known as the endoscope to explore internal body regions. This method allows for the comprehensive examination of the targeted area and facilitates the removal of abnormal cell clusters. In some instances, a complete removal of the affected tissue occurs, occasionally involving the extraction of surrounding normal tissues to ensure thorough treatment.

- Upon identifying abnormal tissue during the examination, the doctor utilized the endoscope to remove the anomalous tissue along with a portion of adjacent normal tissue.

- Surgical procedures involve the removal of abnormal cells areas by a surgeon during an operation, with two primary approaches:excisional and incisional. (a) Excisional biopsy entails the complete removal of the abnormal cell area often accompanied by the extraction of surrounding normal tissue. (b) In contrast incisional biopsy involves the removal of only a portion of the abnormal area by the surgeon.

1.2 Medical Image Analysis, Artificial Intelligence and Deep Learning

Medical imaging, artificial intelligence(AI), and deep learning have converged to revolutionize healthcare diagnostics and treatment planning. Understanding the essence of this transformation requires a deep dive into the major imaging modalities, their unique contributions, and the pivotal role deep learning has played from its inception to its current impact on medical image analysis.

1.2.1 Medical Imaging Modalities

- **Radiography(X-rays) :** Radiography [8], commonly known as X-ray imaging, is the oldest and most widely used medical imaging technique. X-rays use electromagnetic energy beams to create images of the body's internal structures, particularly bones and dense tissues. The varying absorption of X-rays by different tissues produces contrast on the image: bones appear white, soft tissues appear gray, and air spaces appear black. Projectional radiographs are crucial for diagnosing fractures, infections, and tumors, while fluoroscopy provides real-time imaging for procedures like catheter placement.

- **Computed Tomography(CT) :** CT scans [9] expand on X-ray principles by capturing multiple images from different angles and using computer processing to create cross-sectional "slices" of the body. This modality provides detailed visualization of bones, blood vessels, and soft tissues, making it invaluable for diagnosing trauma, cancers, vascular diseases, and guiding biopsies or surgeries. CT imaging offers more detail than standard X-rays, enabling clinicians to detect subtle abnormalities and monitor disease progression.

- **Magnetic Resonance Imaging(MRI) :** MRI [10] uses strong magnetic fields and radio waves to generate highly detailed images of soft tissues, organs and the nervous system. Unlike X-rays and CT, MRI does not use ionizing radiation, making it safer for repeated use. MRI excels in imaging the brain, spinal cord, joints, and soft tissue structures, providing critical information for neurological, musculoskeletal, and cardiovascular diagnoses.

- **Ultrasound(Sonography) :** Ultrasound [11] imaging employs high-frequency sound waves to produce real-time images of internal organs, blood flow, and fetal development. It is non-invasive, radiation-free, and widely used in obstetrics, cardiology, and emergency medicine. Specialized probes allow for transvaginal, endorectal, and intravascular imaging, extending its diagnostic reach.

1.2.2 The Evolution of Deep Learning

- **Origins and Early Development :** Deep Learning [12], a subset of machine learning, traces its roots to the 1950s, inspired by efforts to simulate the human brain information capacity in neural networks. Early neural networks were limited by computational power and data availability. The first significant breakthrough came in the 1960s with the development of small, functional neural networks capable of basic pattern recognition. The 1980s saw the introduction of the "backpropagation algorithm" [13],which allowed for training multi-layer neural networks, significantly improving their ability to recognize shapes and patterns. Geoffrey Hinton, a key figure in deep learning, helped popularize neural networks and coined the term "deep learning" in 2006. Yann LeCunn's work on digit recognition further demonstrated the practical potential of neural networks.

- **Rise of Deep Learning Architectures :** The 2000s marked a turning point as computational power and access to large datasets grew. Deep learning architectures, particularly convolutional neural networks(CNNs) [14], became the backbone of image analysis. CNNs excel at automatically extracting hierarchical features from images, making them ideal for complex visual tasks.

- **Deep Learning in Medical Imaging :** The introduction of deep learning to medical imaging has been transformative. CNNs and related architectures have been applied to a range of tasks, including object localization, segmentation, and classification in medical images. These algorithms can process vast and intricate datasets, learning to identify patterns and anomalies they may be subtle and imperceptible to the human eye. Key advantages of deep learning in medical image analysis include:

  - **Automated Feature Extraction :** Deep learning models learn relevant features directly from raw image data, eliminating the need for manual feature engineering.
  - **High Accuracy and Consistency :** Studies have demonstrated that deep learning algorithms can match or exceed human experts in detecting diseases such as cancer, cardiovascular abnormalities, and neurological disorders.
  - **Scalability :** Once trained, deep learning models can analyze large volumes of images rapidly, supporting real-time diagnostics and reducing the workload on clinicians.
  - **Generalizability :** Advanced architectures can adapt to different imaging modalities(X-ray, CT, MRI, ultrasound) and clinical scenarios, making them versatile tools in healthcare.
  - **Enhancing Diagnostic Accuracy :** Deep learning models, particularly CNNs, have achieved remarkable success in identifying and classifying a wide range of medical conditions from images. Their ability to detect subtle features has led to earlier and more accurate diagnoses, improving patient outcomes.
  - **Supporting Precision Medicine :** By integrating imaging data with patient histories and genetic information, deep learning enables personalized treatment planning. Predictive models can forecast disease progression and response to therapy, guiding clinicians in selecting optimal interventions.
  - **Improving Efficiency and Access :** Automated image analysis reduces diagnostics turnaround time, enabling faster decision-making in critical care settings. This efficiency is especially valuable in resource-limited environments, where access to expert radiologists may be scarce.
  - **Expanding Research and Innovation :** Deep learning has accelerated biomedical research by enabling large-scale analysis of imaging datasets. Researchers can uncover new disease biomarkers, track epidemiological trends, and develop novel diagnostic tools.

  Despite its successes, deep learning in medical imaging faces challenges. Ensuring model generalizability across diverse populations and imaging devices is critical. Data privacy and security

must be maintained, especially when handling sensitive medical images. Interpretability remains a concern, as clinicians need to understand the rationale behind AI-driven decisions.

Ongoing research is focused on developing explainable AI models, improving data standardization, and integrating multi-modal data sources. The future promises even greater synergy between AI and medical imaging, with the potential to revolutionize diagnostics, treatment, and patient care. The advent of deep learning has amplified the value of these modalities by enabling automated, accurate, and scalable image analysis. From its early theoretical roots to its current practical applications, deep learning has become indispensable in modern medical imaging, driving advances in diagnosis, precision medicine, and healthcare delivery.

1.3 Challenges

The integration of artificial intelligence(AI) [15] and machine learning in healthcare is transforming diagnostics, patient care, and research. However, the path from concept to clinical deployment is fraught with multifaceted challenges. Each stage – dataset development, annotation, regulatory approval, frontline health worker training, ethical and privacy considerations, and handling image artifacts and variability – presents unique obstacles that must be addressed for safe, effective, and equitable AI adoption.

1. **Dataset Development**

    - **Fragmentation and Incompleteness :** Healthcare data is often fragmented across multiple sources – hospitals, clinics, insurers, and personal devices [16] – resulting in incomplete patient profiles. This fragmentation complicates the creation of comprehensive datasets necessary for building robust AI models. Data may be duplicated, inconsistent, or missing, leading to imprecise patient records and hindering large-scale studies.
    - **Data Diversity and Representation :** A critical challenge is ensuring datasets are representative of diverse populations [17]. Many datasets under-represent certain groups due to structural barriers in healthcare access, insufficient granularity in data collection, or legal and ethical restrictions on data sharing.
    - **Data Quality and Consistency :** Data quality if frequently compromised by missing entries, inconsistent coding and human error. For example, handwritten notes may not be digitized accurately, and different institutions may use varying terminologies for the same conditions. Poor data quality skews results, undermines clinical decision-making, and limits the generalization of AI solutions.
    - **Evolving Nature of Data :** Healthcare data is dynamic; patient information, diagnostic criteria, and treatment protocols change over time. Keeping datasets current is a continuous challenge, requiring ongoing data integration and validation.

2. **Data Annotation**

   - **Complexity and Expertise Requirements :** Medical data annotation [**?**] is uniquely challenging due to the complexity of healthcare data – ranging from imaging(e.g., X-rays, MRIs) to free-text notes and bio-signals. Annotation requires domain-specific expertise; only trained clinicians can accurately label subtle features in medical images or interpret nuanced clinical notes. The shortage of qualified annotators limits scalability and increase costs.
   - **Ambiguity and Subjectivity :** Medical data often presents ambiguities – different experts may interpret the same image or case differently. This subjectivity can lead to inconsistent annotations, which degrade model performance. Inter-annotator agreement metrics and multi-tiered review pipelines are necessary to ensure reliability.
   - **Bias in Labeling :** Annotation bias is a significant risk. Annotators may unconsciously favor certain outcomes or be more familiar with specific demographics, leading to skewed datasets. Such biases, if unaddressed, can be embedded in AI models, perpetuating health disparities.
   - **Cost and Time Overheads :** High-quality annotation is resource-intensive, especially for rare conditions that require multi-expert review. The process is slow and expensive, often limiting the size and diversity of annotated datasets.
   - **Tool Limitations :** Few annotation platforms fully support the specialized needs of medical data, such as DICOM standards for imaging or 3D visualization tools. This limits efficiency and accuracy in the annotation process.

3. **Ethical and Privacy Concerns**

   - **Patient Privacy and Data Security :** Healthcare data is highly sensitive, and its use in AI development raises significant privacy concerns. Strict regulations(e.g.,HIPAA, GDPR) govern data handling, requiring robust de-identification and secure storage. Breaches can have severe consequences for patients and institutions.
   - **Informed Consent :** Obtaining informed consent for the use of patient data in AI research is challenging, especially when data is repurposed for new applications. Patients may not fully understand how their data will be used, or may be reluctant to share information due to lack of trust.
   - **Bias and Fairness :** AI models trained on biased or unrepresentative data can perpetuate or exacerbate health disparities. Ensuring fairness requires deliberate efforts to include diverse populations in datasets and to monitor model performance across demographic groups.
   - **Transparency and Accountability :** Ethical AI development demands transparency about data sources, annotation processes, and model limitations. Clear accountability structures are needed to address errors or adverse outcomes linked to AI recommendations.

4. **Image Artifacts and Variability**

    - **Technical Artifacts :** Medical images often contain artifacts —- unwanted alterations caused by patient movement, equipment malfunction, or improper technique. These artifacts can obscure critical features, confuse annotators, and degrade AI model performance.
    - **Inter-Device and Inter-Operator Variability :** Images acquired on different devices or by different operators may vary in quality, contrast, and resolution. Such variability complicates annotation and model training, as AI systems may learn device - or operator - specific patterns rather than true pathology.
    - **Standardization Challenges :** Lack of standardized imaging protocols across institutions leads to heterogeneous datasets. Without harmonization, AI models may not generalize well to new settings or populations.
    - **Impact on Model Robustness :** AI models trained on artifact - laden or variable - quality images may perform poorly in real - world clinical environments, where image quality cannot be guaranteed. This undermines trust and limits adoption.

1.4 Scope and Contributions

The following are the technical contributions from the author as part of the thesis:
Scope :

1. **AI-Driven Screening for OPMDs :** The thesis explores the development and validation of AI-based models for the detection of OPMDs using photographic images captured with smartphones. It focuses on the use of deep learning architectures(e.g., DenseNet, Swin Transformers) for image classification, feasibility of deploying such models in real-world, resource-constrained environments and the role of front-line health workers(FHWs) [18] in capturing images and facilitating early detection.

2. **Dataset Creation and Diversity :** A significant scope of the thesis is the creation and curation of large, annotated datasets comprising thousands of oral cavity images. The datasets includes:(a) Images taken by both trained and untrained FHWs, reflecting real-world variability in image quality. (b) Diverse cases, including both suspicious and non-suspicious lesions, to ensure robust model performance and generalizability.

3. **Evaluation and Validation :** The thesis systematically evaluates the performance of AI models on internal and independent test sets, addressing the impact of image quality and FHW [18] training on model accuracy along with generalizability of models to new, unseen data from different sources highlighting limitation of AI in point-of-care screening.

4. **Implementation Challenges :** The study also includes a critical discussion of challenges in scaling AI solutions in low- and middle- income countries(LMICs), ethical, regulatory and privacy considerations in digital health deployments.

Contributions :

1. **Novel AI Screening Pipeline :** The primary contribution is the design and validation of a novel AI-assisted screening pipeline for OPMDs. The studies compare state-of-the-art deep learning models(DenseNet201, Swin Transformer) and demonstrate their effectiveness in classifying oral lesions from smartphone images. The best-performing models achieve high F1-scores(up to 0.84 on internal test sets), establishing new benchmarks for non-invasive, image-based OPMD screening.

2. **Real-World Dataset and Evaluation :** The creation of a large, annotated dataset of over 2,000 oral images, including a unique independent test set collected by trained and untrained FHWs, is a significant resource for the research community.

3. **Empowerment of Front-Line Health Workers :** The studies demonstrate that, with minimal training, FHWs can effectively capture images suitable for AI analysis, enabling scalable screening in remote or underserved areas. The research provides a blueprint for integrating AI tools into existing health systems, supporting early detection and referral.

4. **Addressing Implementation Barriers :** By systematically evaluating the impact of image quality and operator training, the work highlights key barriers to implementation and suggests pathways for improving robustness.The work also discuss the ethical implications of AI-driven screening, including data privacy, informed consent, and the need for regulatory frameworks tailored to digital health innovations.

5. **Societal and Health Impact :** The proposed solutions have the potential to significantly improve early detection rates for OPMDs, particularly in LMICs where specialist access is limited. The use of ubiquitous smartphone technology and AI enables scalable, cost-effective screening programs that can be deployed at the community level.

1.5 Thesis Outline

Chapter 2 presents a comprehensive evaluation of artificial intelligence (AI)-assisted screening for oral potentially malignant disorders (OPMDs) [19] using photographic images captured via smartphones. The research addresses a critical gap in the early detection of oral cancer, particularly in low- and middle-income countries where access to specialists and advanced diagnostic tools is limited.

The chapter introduces a deep learning pipeline utilizing two state-of-the-art architectures—DenseNet201 and Swin Transformer—to classify oral cavity images as either ‘suspicious’ (potentially malignant) or

'non-suspicious.' The dataset comprised 2,178 images collected during oral cancer screening camps in India. It then describes the dataset acquisition process, annotation process, transfer learning protocols, followed by the experimental setup, results, a discussion on explainability, and a summary of findings.

Chapter 3 summarizes the key findings of the research, highlighting the contributions made to the field of mobile diagnostics at the community screening level and concluding the thesis. It also discusses the limitations of the current study and suggests directions for future research to further advance the field.

Chapter 2

AI-Assisted Screening of Oral Potentially Malignant Disorders Using Smartphone-Based Photographic Images

Oral Cancer and oral potentially malignant disorders(OPMDs) are significant public health concerns worldwide, particularly in regions where risk factors such as tobacco, alcohol, and betel nut use are prevalent. These conditions are closely linked, as OPMDs often serve as precursors to oral cancer. Early recognition, prevention, and intervention are critical to reducing morbidity and mortality associated with these diseases.

2.1 Introduction

2.1.1 Oral Cancer

Oral cancer refers to malignant neoplasms arising from the lining of the oral cavity, mostly commonly squamous cell carcinoma, which accounts for over 90% of cases. The oral cavity includes the lips, anterior two-thirds of the tongue, floor of the mouth, buccal mucosa (cheeks), hard palate, retromolar trigone, and alveolar ridges. Globally, oral cancer ranks among the top ten most common cancers, with an estimated 377,713 new cases and 177,757 deaths reported in 2020 alone. The burden is particularly high in South and Southeast Asia, parts of Europe, and among certain populations in North America, reflecting the distribution of risk factors [20].

Epidemiological Highlights [21] :

- **Geographical Variation :** Incidence rates are highest in India, Pakistan, Sri Lanka, and Bangladesh, where oral cancer constitutes up to 40% of all cancers.
- **Gender and Age :** Oral cancer is more common in men than women, likely due to higher exposure to risk factors. The majority of cases occur in individuals over 40, but younger patients are increasingly affected, especially with rising Human Papillomavirus(HPV) prevalence.
- **Socioeconomic factors :** Lower socioeconomic status is associated with higher incidence, likely due to increased risk factor exposure and reduced access to healthcare.

Etiology and Risk Factors [21] : The development of oral cancer is multi-factorial, involving both environmental and genetic influences. The main risk factors are :

- **Tobacco Use :** Smoking cigarettes, cigars, pipes, and using smokeless tobacco (chewing tobacco, snuff) are the most significant risk factors. Tobacco contains carcinogens such as nitrosamine and polycyclic aromatic hydrocarbons, which cause genetic mutations in oral epithelial cells.
- **Alcohol Consumption :** Alcohol acts synergistically with tobacco, increasing the permeability of oral mucosa to carcinogens and generating acetaldehyde, a known carcinogen. Heavy drinkers who also smoke have a 15-fold increased risk compared to non-users.
- **Betel Nut(Areca Nut) Chewing :** Common in South Asia and among migrant populations, betel nut contains arecoline, which promotes fibrosis and carcinogenesis. Betel quid often includes tobacco, further increasing risk.

Clinical Presentation : Oral cancer often presents with subtle, non-specific symptoms in its early stages, leading to delayed diagnosis. Common signs and symptoms include :

- **Non-healing Ulcer :** A persistent ulcer or sore in the mouth that does not heal within two weeks is a classic sign.
- **Lump or Mass :** A swelling, thickening, or lump in the oral tissues, often painless initially.
- **White or Red Patches :** Leukoplakia(white patch) [22] and erythroplakia(red patch) are common presentations. Erythroplakia carries a higher risk of malignancy.
- **Difficulty Chewing or Swallowing :** Tumors involving tongue, floor of mouth, or oropharynx can impair mastication and deglutition.

Early lesions may be asymptomatic or mistaken for benign conditions, emphasizing the importance of thorough oral examinations, especially in high-risk individuals.
Diagnosis : Accurate and timely diagnosis of oral cancer is essential for optimal outcomes. The diagnostic process typically involves :

- **Clinical Examination :** A thorough inspection and palpation of the oral cavity, oropharynx, and neck to identify suspicious lesions and lymphadenopathy.
- **Imaging Studies :**
  - **Radiographs :** Useful for assessing bone involvement.
  - **CT and MRI :** Provide detailed information about tumor size, depth and spread to adjacent structures.
  - **PET Scans :** Help detect distant metastases and assess metabolic activity.

- **Biopsy :** The definitive diagnosis relies on histopathology examination. Techniques include :
  - **Incisional Biopsy :** Removal of a representative portion of the lesion.
  - **Excisional Biopsy :** Removal of the entire lesion, typically for small, accessible lesions.

Management : The management of oral cancer is multidisciplinary, involving surgery, radiation, chemotherapy, and supportive care. The choice of treatment depends on the tumor's site, size, stage, and patient factors.

- **Surgery :** The primary treatment for most oral cancers. Surgical options include wide local excision, mandibulectomy (removal of part of the jaw), and neck dissection for lymph node involvement. Reconstruction with grafts or flaps may be necessary for large defects.
- **Radiation Therapy :** Used as primary treatment for early-stage tumors, as adjuvant therapy after surgery for advanced disease, or for palliation in inoperable cases. Techniques include external-beam radiation and interstitial brachytherapy.
- **Chemotherapy :** Often combined with radiation (chemoradiation) for advanced, unresectable, or metastatic disease. Common agents include cisplatin, 5-fluorouracil, and taxanes.
- **Rehabilitation :** Post-treatment rehabilitation is vital for restoring speech, swallowing, and appearance. Prosthodontic appliances, speech therapy, and nutritional support are often required.
- **Palliative Care :** For advanced disease, the focus shifts to symptom control and quality of life.

Prognosis : Prognosis depends largely on the stage at diagnosis. Early-stage oral cancers (stage I and II) have a 5-year survival rate of 70–90%, while advanced-stage disease (stage III and IV) drops to 30–50%. Factors influencing prognosis include tumor size, depth of invasion, lymph node involvement, histological grade, and patient comorbidities. Delayed diagnosis, common in resource-limited settings, contributes to poor outcomes.

2.1.2 Oral Potentially Malignant Disorders(OPMDs)

OPMDs are a diverse group of lesions and conditions of the oral mucosa that carry an increased risk of progression to cancer. They serve as clinical markers for heightened surveillance and preventive intervention.

Types of OPMDs :

1. **Leukoplakia [22] :** Defined as a white patch or plaque that cannot be characterized clinically or pathologically as any other disease. It is the most common OPMD, with a malignant transformation rate ranging from 1% to 20% over time. Risk is higher for non-homogeneous (speckled or nodular) types and lesions on the tongue or floor of mouth.

2. **Erythroplakia :** Appears as a fiery red, velvety patch that cannot be attributed to any other condition. Although less common than leukoplakia, erythroplakia has a much higher risk of malignant transformation (up to 50% in some studies).

3. **Oral Submucous Fibrosis(OSMF) :** A chronic, progressive condition characterized by fibrosis of the oral mucosa, leading to stiffness, trismus (reduced mouth opening), and burning sensation. It is strongly associated with betel nut chewing and carries a malignant transformation rate of 7–13%.

4. **Lichen Planus :** A chronic inflammatory condition, presenting as white, lacy patches (reticular type) or erosive/ulcerative lesions. The erosive type has a small but definite risk of malignancy (about 1%).

Pathogenesis and Risk of Transformation : The process of malignant transformation in OPMDs involves a series of genetic and epigenetic changes leading to dysplasia and eventual carcinoma. Key factors influencing risk include :

- **Clinical Appearance :** Non-homogeneous lesions (speckled, nodular, or verrucous) are more likely to become malignant than homogeneous, flat lesions.
- **Site :** Lesions on the lateral border of the tongue and floor of mouth have higher malignant potential.
- **Histopathology :** The degree of epithelial dysplasia(mild, moderate, severe) is the most important predictor. Severe dysplasia or carcinoma insitu warrants prompt intervention.
- **Duration and Size :** Larger and long-standing lesions have higher risk.
- **Patient Factors :** Age, gender, immune status, and genetic predisposition may influence risk.

Diagnosis of OPMDs : Early detection and accurate diagnosis are essential for preventing malignant progression. The diagnostic approach includes :

- **Clinical Examination :** Careful inspection and palpation of the oral mucosa during routing dental or medical visits, especially in high-risk individuals.
- **Biopsy :** The gold standard for diagnosis and grading of dysplasia. Incisional biopsy is preferred for large or suspicious lesions.
- **Imaging :** Not routinely required unless there is suspicion of deep tissue involvement.

Management of OPMDs : Management aims to eliminate risk factors, treat dysplasia, and prevent progression to cancer.

- **Elimination of Risk Factors :**

  - **Tobacco Cessation :** Counseling, nicotine replacement therapy, and behavioral interventions.
  - **Alcohol Reduction :** Education and support for reducing or abstaining from alcohol.
  - **Betel Nut Cessation :** Community-based programs to discourage use.

- **Medical Management :**
  - **Topical Steroids :** Used for symptomatic relief in lichen planus and other inflammatory OPMDs.
  - **Antioxidants and Retinoids :** Some studies suggest benefit in reducing dysplasia, but evidence is mixed.

- **Surgical Management :**
  - **Excision :** Indicated for lesions with moderate or severe dysplasia, non-homogeneous leukoplakia, erythroplakia, any lesion with suspicious features.
  - **Laser Ablation :** Alternate methods for removing or destroying dysplastic tissue.

Prevention and Public Health Strategies : Given the preventable nature of many risk factors, public health initiatives play a crucial role in reducing the burden of oral cancer and OPMDs.

- **Tobacco and Alcohol Control :** Legislation, taxation, advertising restrictions, and public education campaigns have proven effective in reducing use.
- **Betel Nut Regulation :** Policies to restrict sale and use, particularly among youth, are needed in high-prevalence regions.
- **Awareness and Education :** Training healthcare professionals to recognize early signs and educating the public about risk factors and symptoms.
- **Screening Programs :** Targeted screening in high-risk population demography can facilitate early detection and intervention.

Oral cancer and oral potentially malignant disorders [23] are interrelated conditions with significant health, social, and economic implications. Most cases are attributable to modifiable risk factors, making prevention and early intervention highly effective strategies. Regular oral examinations, prompt biopsy of suspicious lesions, and multidisciplinary management are essential to improving outcomes. Ongoing research into molecular markers and targeted therapies holds promise for further advances in diagnosis and treatment. Ultimately, reducing the global burden of oral cancer and OPMDs requires a comprehensive approach involving individuals, healthcare providers, and public health systems.

2.2 Related Works

2.2.1 Advancements in Digital Diagnostics for Oral Potentially Malignant Disorders

Oral potentially malignant disorders (OPMDs) encompass a group of conditions, including oral leukoplakia [24], erythroplakia, and oral lichen planus, that carry a significant risk of progressing to oral squamous cell carcinoma (OSCC), a prevalent and aggressive form of oral cancer. With over 300,000 new oral cancer cases diagnosed globally each year, early detection of OPMDs is critical to improving patient outcomes and reducing mortality rates, as the five-year survival rate for OSCC remains below 50% when diagnosed at advanced stages (State of the Art in OPMD Diagnosis). Traditional diagnostic methods, such as visual inspection and biopsy, have limitations, including subjectivity and invasiveness, prompting the development of digital diagnostics. These technologies, encompassing artificial intelligence (AI), machine learning, advanced imaging, and mobile applications, offer innovative solutions for accurate, non-invasive, and timely detection of OPMDs. This study explores the advancements in digital diagnostics for OPMDs as of 2025, highlighting their applications, benefits, challenges, and future potential.

2.2.1.1 Traditional Methods of Diagnosing OPMDs

Historically, the diagnosis of OPMDs has relied on clinical examination and histopathological analysis. During a clinical examination, healthcare providers visually inspect the oral mucosa for abnormalities such as white or red patches, ulcers, or irregular tissue changes. However, this method is subjective, heavily dependent on the clinician's expertise, and may fail to detect subtle changes indicative of early OPMDs. Distinguishing between benign and potentially malignant lesions based solely on visual inspection can be challenging, leading to potential misdiagnoses.

Biopsy, the gold standard for confirming OPMD diagnoses, involves removing a tissue sample for microscopic examination to assess for dysplasia [25] or malignancy. While effective, biopsies are invasive, can cause patient discomfort, and carry risks such as infection or bleeding. Additionally, interpreting biopsy results can be complex, particularly when histological features are ambiguous. These limitations underscore the need for alternative diagnostic approaches that are more objective, less invasive, and capable of detecting OPMDs at an earlier stage.

2.2.1.2 Advancements in Digital Diagnostics

Digital diagnostics have revolutionized the detection of OPMDs by leveraging cutting-edge technologies to enhance accuracy and accessibility. These advancements include AI and machine learning, mobile applications, and advanced imaging techniques, each contributing to improved diagnostic capabilities.

1. **Artificial Intelligence and Machine Learning :** AI, particularly deep learning, has emerged as a powerful tool for diagnosing OPMDs. Deep learning models analyze large datasets of images

to identify patterns and features indicative of disease, often achieving accuracy comparable to or surpassing that of human experts. A notable advancement is the use of vision transformers for multi-class detection of OPMDs. A 2024, study [26] developed a Mask R-CNN model with a Swin Transformer backbone to classify clinical photographs of oral leukoplakia, oral lichen planus, OSCC, and healthy mucosa. The model demonstrated high performance, with an average specificity of 0.946, sensitivity of 0.831, precision of 0.856, accuracy of 0.928, and F1 score of 0.842 across these conditions (Advancements in Diagnosing OPMDs). The model's ability to differentiate between disorders with high accuracy highlights its potential as a diagnostic aid. Another significant development is the integration of AI with confocal laser endomicroscopy (CLE) for real-time digital oral microscopy. CLE enables in vivo microscopic examination of oral tissues, producing high-resolution images that AI algorithms can analyze to detect cellular abnormalities. This technology facilitates rapid, precise, and reproducible diagnoses, potentially reducing the reliance on invasive biopsies (AI-Driven Oral Microscopy). By automating image interpretation, AI-driven CLE addresses the challenge of operator expertise, making it a promising tool for widespread clinical adoption.

2. **Mobile Applications :** Mobile applications have become increasingly important in the screening and diagnosis of OPMDs. These apps use algorithms to analyze images of oral lesions submitted by patients or healthcare providers, offering risk assessments and facilitating remote consultations. A 2023, systematic review highlighted the potential of mobile applications in risk assessment, screening, diagnosis, and treatment monitoring for oral cancer and OPMDs. Benefits include timely referrals, self-monitoring, and improved adherence to treatment plans (Mobile Apps in Oral Cancer). For example, apps can enable patients in remote or underserved areas to capture images of oral lesions and receive preliminary assessments, connecting them with specialists for further evaluation. However, the effectiveness of mobile applications depends on image quality and algorithm accuracy, necessitating ongoing validation and refinement.

3. **Benefits and Challenges :** The advancements in digital diagnostics offer numerous benefits for OPMD detection:

    - Enhanced Accuracy: AI and machine learning models provide objective analysis, reducing the subjectivity of visual inspections.
    - Early Detection: Digital tools can identify subtle changes, enabling intervention before malignant transformation.
    - Improved Accessibility: Mobile apps and tele-medicine extend diagnostic capabilities to underserved regions, enhancing healthcare equity.

    Despite these advantages, challenges remain:

    - Data Limitations : AI models require large, diverse datasets, which can be scarce for rare conditions like OPMDs.

- Clinical Integration : Incorporating digital diagnostics into routine practice requires training and infrastructure upgrades.
- Cost and Accessibility : Advanced imaging systems and AI tools can be expensive, limiting their availability in low-resource settings.
- Validation Needs : Rigorous clinical trials and regulatory approvals are necessary to ensure the reliability and safety of these technologies.

2.2.2 Existing Works

Oral Potentially Malignant Disorders(OPMDs) [27] exhibit an increased risk of malignant transformation. These lesions present an array of clinical variations, including white, red, or mixed red-white lesions with verrucous [28], papillary [28], corrugated, atrophic, and ulcerated presentations. In addition, lesions like frictional keratosis, chemical injury, luekoedema, candidasis, denture-associated stomatitis, and desquamative or autoimmune disorders exhibit overlapping clinical features, making the diagnosis of OPMDs challenging. Though oral cancers can develop de novo, OPMDs share numerous risk factors and molecular/genetic alterations with oral cancers. Studies indicate that most habit-associated oral cancers evolve from pre-existing OPMDs. Preliminary epidemiological research and systematic reviews report that upto 40% of leukoplakia develops into oral cancer. Hence, the early diagnosis and differentiation of OPMDs from clinically similar-appearing lesions are vital for limiting the possible malignant change and improving treatment outcomes. The oral cavity can be easily visualized without special instruments compared to other internal organs. A Conventional Oral Examination(COE) which involves a visual inspection by a specialist, is the clinical standard for detecting oral lesions. The clinical assessment of OPMDs is subjective, and biopsies remain the gold standard for their definitive diagnoses. However, many high-risk individuals in low- and middle-income countries lack access to specialists or adequate health services, leading to delays in diagnoses and referrals for patients with OPMDs [27] and oral cancer. On the other hand, diagnoses based on biopsies are not ideal for screening due to their invasive nature and limited availability of experts of point-of-care or remote locations. Therefore, there is a definite need to develop an easy-to-use, non-invasive oral screening tool that enhances the existing system for managing OPMDs. Comprehensive clinical assessments, swift patient referrals for biopsies, and the cessation of habits/risk factors are keys to better patient care. Different studies have evaluated auto-fluorescence imaging devices as clinical adjuncts to COE for detecting OPMDs and oral cancer [29]. These studies showed that combining auto-fluorescence visualization with COE provides better accuracy than either method alone. Multi-spectral screening devices incorporating different lights(white, violet, and green-amber) have shown promise in maximizing the advantages of white light and fluorescence-based examinations for detecting OPMD. An accurate interpretation of results requires training and an understanding of oral pathology. Since an oral examination by a expert is not always feasible in primary care or community settings, implementing an automatic classification system based on oral cavity imaging would be beneficial. Increasing evidence shows that deep learning techniques can match or surpass human experts in diverse prediction tasks, including classifying dif-

ferent types of cancers like skin, breast, and cervical cancer. Artificial Intelligence(AI) in healthcare is poised to improve the experience of both clinicians and patients.
In this study, we examined the potential of AI in detecting OPMDs from the photographic images of oral cavities in the Indian population. A large dataset of oral cavity images captured using a regular smartphone camera from the community screening camps in India was used for this purpose. This dataset comprises photographic images of normal oral cavities, OPMDs, and a smaller set of oral cancer images. The major objective is to evaluate the performance of different state-of-the-art deep learning models to identify suspicious lesions comprising of OPMDs and oral cancer using white light imaging. Convolutional Neural Networks(CNNNs) are well-known deep learning architectures widely used in image classification tasks, including in the medical domain, for identifying various diseases. The success of transformers in natural language processing has led to their adaptation to computer vision problems. Dosovitskiy et al showed that vision transformers [30], self-attention based architectures, can attain excellent performance compared to CNNs in various imaging benchmarks, requiring fewer computational resources to train.

2.3 Biocon Foundation Dataset

2.3.1 Dataset Collection

Intraoral smartphone-based white light images were systematically collected as part of a large-scale community outreach initiative aimed at early detection and prevention of oral cancer. This program was a collaborative effort between the Biocon Foundation and the Department of Oral Medicine and Radiology at KLE Society's Institute of Dental Sciences, Bengaluru. The entire program was thoroughly reviewed and approved by the institutional review board, ensuring adherence to ethical standards and patient safety.

The collection of intraoral images was carried out by front-line healthcare workers(FHWs) who underwent comprehensive skill training delivered oral medicine specialists. The training module was meticulously designed to cover a broad spectrum of knowledge areas including the epidemiology and burden of oral cancer, the importance of awareness, principles of early detection, and preventive strategies. The training tools employed were diverse and interactive, comprising PowerPoint presentations for theoretical knowledge, focus group discussions to facilitate peer learning and clarify doubts, and in situ simulations to provide hands-on experience. The in situ simulation component was particularly impactful, involving a realistic setup with a chair, a patient, and a step-be-step guide that walked the FHWs through the process of examining the oral cavity. This simulation emphasized the identification of normal oral mucosa as well as the recognition of various tobacco-induced lesions, which are critical in the context of oral cancer screening.

To assess the effectiveness of the training, a pre-test in the form of a questionnaire was administered to all participating FHWs before the commencement of the training. This pre-test served to gauge the baseline knowledge and identify areas requiring further emphasis. Upon completion of the training, a

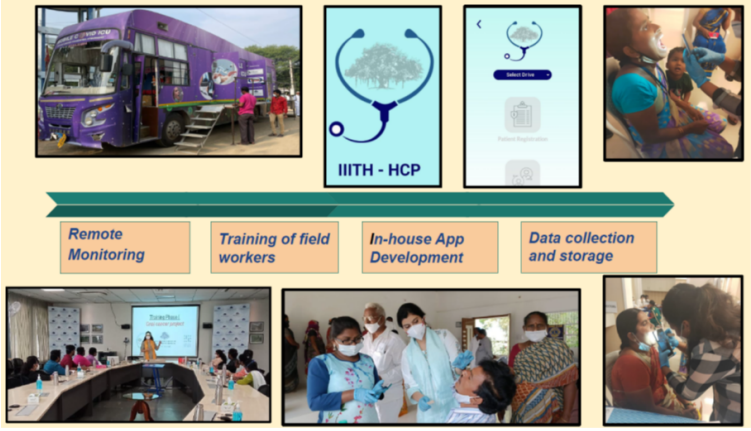


Figure 2.1 Frontline health workers data collection

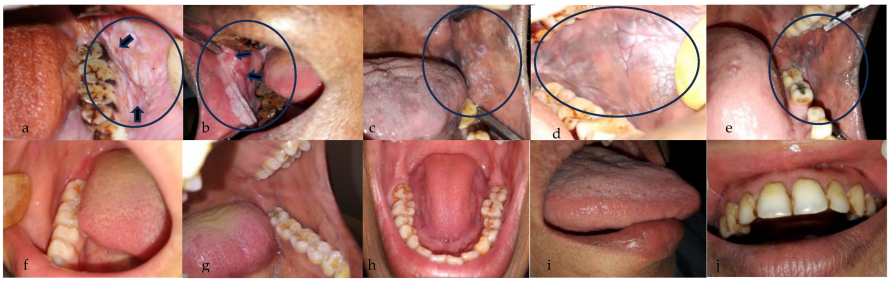


Figure 2.2 Intraoral images of suspicious and non-suspicious lesions. (a,b): Buccal mucosa showing a white lesion with few red areas (indicated by arrows) suggestive of non-homogeneous leukoplakia. (c): Left buccal mucosa showing a reticular, lacy white lesion suggestive of lichen planus. (d,e): Left buccal mucosa showing a white patch suggestive of homogeneous leukoplakia. (f,g): Normal appearance of left and right buccal mucosa. (h,i): Dorsal and lateral surface of a tongue showing no abnormalities. (j) Upper labial mucosa and vestibule showing no abnormalities. The black circle indicates the region of interest.

post-test was conducted to evaluate the knowledge gained and the overall effectiveness of the training program. In cases where FHWs did not achieve satisfactory scores on the post-test, they were provided with additional training to ensure competency. At the conclusion of the training, each FHW was given a clinical manual for future reference, enabling them to reinforce their learning and maintain high standards in the field.

In the field setting, FHWs used smartphone cameras with a minimum resolution of 5 megapixels to capture high-quality intraoral images, focusing predominantly on the buccal mucosa (as illustrated in Figure 2.2). Alongside image capture, detailed patient demographic information was collected, including age, gender, and relevant medical history. Habitual risk factors such as alcohol consumption, tobacco use (in both smoked and smokeless forms), pan chewing, and other related behaviors were meticulously documented. The study population was restricted to adults aged 18 years and above, and

written informed consent was obtained from all participants, ensuring ethical compliance and respect for patient autonomy.

2.3.2 Data Annotation

Once collected, the intraoral images underwent a detailed annotation process conducted by three experienced oral medicine specialists. Each image was independently reviewed and classified into one of two categories: suspicious or non-suspicious. The suspicious category primarily included images depicting oral potentially malignant disorders (OPMDs), as defined by standard clinical criteria. These OPMDs included homogenous leukoplakia, non-homogenous leukoplakia, erythroplakia, verrucous leukoplakia, oral lichen planus, oral submucous fibrosis, and tobacco pouch keratosis . Additionally, the dataset contained a number of images representing oral cancers, characterized by ulceroproliferative or exophytic growths. In contrast, the non-suspicious category comprised images of normal oral mucosa, normal anatomical variations, and benign lesions that did not raise suspicion for malignancy.

The specialist diagnosis served as the reference standard for image classification in this study. This approach was grounded in previous research by Birur et al.(2019, 2022), which demonstrated that remote specialist evaluations were as accurate as those conducted onsite for the diagnosis of OPMDs and oral cancer. Specifically, onsite specialist diagnoses exhibited high sensitivity (94%) when compared to histological findings, while remote specialist diagnoses achieved high accuracy relative to onsite assessments, with a sensitivity of 95% and specificity of 84% . These findings validated the reliability of using specialist annotation as the reference standard for the present dataset.

To further ensure the integrity of the dataset, all images were manually reviewed to exclude those with significant blurring, flash distortions, being out-of-focus, or where the lesions of interest were not adequately visible. This meticulous curation process resulted in a final dataset comprising 2,178 high-quality images, of which 1,120 were classified as suspicious and 1,058 as non-suspicious.

For the purpose of developing and evaluating machine learning models for automated image classification, the curated dataset was randomly divided into three subsets: a training set (1344 images), a validation set (412 images), and a testing set (422 images).

Table 2.1 Dataset Description

Image Category	Train	Validation	Test	Total
Suspicious	670	216	234	1120
Non-Suspicious	674	206	178	1058
Total	1344	412	422	2178

To further assess the generalizability and robustness of the best-performing models, an independent test set (set I) was assembled. This set consisted of 440 photographic images, equally divided between suspicious (220 images) and non-suspicious (220 images) categories. Notably, these images were captured by untrained FHWs during oral cancer screening camps organized by the Grace Cancer Foundation in Telangana. The inclusion of this independent test set provided a stringent evaluation of

model performance in real-world conditions, where image quality and consistency may vary due to the lack of specialized training among image collectors.

2.3.3 Quality Control Guidelines

To maintain the quality and diagnostic utility of the images, oral medicine specialists periodically reviewed the images submitted by FHWs. If any images were found to be of suboptimal quality—due to factors such as blurring, flash-induced distortions, or being out-of-focus the respective patients were re-screened to obtain suitable images. This rigorous quality control process ensured that only diagnostically useful images were included in the subsequent analysis.

In summary, this comprehensive approach—spanning rigorous training of FHWs, systematic image collection and annotation, stringent quality control, and robust dataset curation—ensured the development of a high-quality resource for advancing the early detection and prevention of oral cancer through both clinical and technological innovations. The dataset not only supports the evaluation of machine learning algorithms but also serves as a valuable reference for future research and community-based screening programs.

2.4 Methods

2.4.1 Supervised Learning

Supervised learning is a foundational technique in the field of machine learning and artificial intelligence. It enables computers to learn from labeled data - data where the correct output is already known - so that they can make accurate predictions or classifications when presented with new, unseen data. The model learns the underlying relationships between input features and their corresponding outputs, enabling it to predict correct outputs for new, unlabeled data. Labeled data consists of example data points paired with the correct answers, explicitly teaching the model to identify the relationships between features and data labels. As the algorithm processes these examples, it adjusts its internal parameters to minimize the difference between its predictions and the actual outputs. The primary aim of supervised learning is to make sense of data within the context of a specific question, such as classifying emails as spam or not spam, or predicting the price of a house based on its features.

Supervised Learning in deep neural networks is a structured process that involves training models on labeled data, validating their performance, and optimizing them using specific loss and objective functions. This approach enables neural networks to learn complex patterns and make accurate predictions. Below is a detailed exploration of these components from a neural network perspective.

2.4.1.1 Training, Validation, and Test Data

- **Training Data :** Neural networks learn by adjusting their internal parameters(weights) based on labeled examples from the training dataset. The model iteratively processes this data, identifying

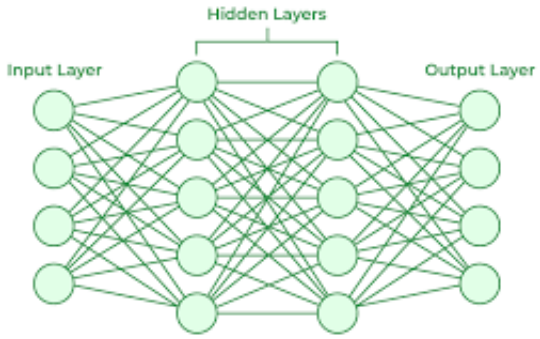


Figure 2.3 Neural Network

patterns and relationships that allow it to predict outcomes for new, unseen inputs. For instance, in image classification, the network learns to associate visual features with specific labels by minimizing the error between its predictions and the true labels during training.

- **Validation Data :** The validation dataset is used to evaluate the model's performance during training but is not used to update the model's weights. Instead, it helps in tuning hyperparameter(such as the number of layers, learning rate, or number of hidden units) and in making decisions about model architecture. Monitoring performance on the validation set helps prevent overfitting, as the model may otherwise become too specialized to the training data and lose generalizability. Techniques like early stopping rely on validation loss: training halts when validation error starts increasing, signaling potential overfitting.

- **Test Data :** The test dataset is reserved for the final, unbiased evaluation of the model after all training and validation steps are complete. It provides an estimate of how the model will perform on truly unseen data, ensuring that the reported accuracy or error is not inflated by repeated exposure during training or validation.

2.4.1.2 Loss Functions in Neural Networks

A loss function quantifies the difference between the neural network's predictions and the actual target values (ground truth). It is the mathematical foundation that guides the learning process. The loss function measures how well (or poorly) the model is performing on a specific task. The objective during training is to minimize this loss, thereby improving the model's predictive accuracy.

Types of Loss Functions :

- **Mean Squared Error :** Used for regression tasks, MSE calculates the average of the squared differences between predicted and actual values. It penalizes larger errors more heavily, making

it suitable for continuous output predictions such as stock prices or house values.

$$\mathcal{L}_{\text{MSE}}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{k=1}^{K} \big(\hat{y}_k - y_k\big)^2. \tag{2.1}$$

- **Cross-Entropy Loss(Log Loss) :** Commonly used for classification tasks, cross-entropy loss measures the difference between the predicted probability distribution and the true distribution (labels). It is particularly effective when the network outputs probabilities for each class, as in image or text classification

$$\mathcal{L}_{\text{BCE}}(\hat{y}, y) = -\big[y \log(\hat{y}) + (1-y) \log(1-\hat{y})\big] \tag{2.2}$$

The loss function is a function of the network's weights. During training, algorithms like gradient descent adjust these weights to minimize the loss, thereby improving the model's predictions.

2.4.1.3 Objective Functions and Optimization

- **Objective Function :** In neural networks, the objective function is typically synonymous with the loss function. The network's sole objective during training is to minimize this function. Sometimes, the objective function may include additional terms, such as regularization, to penalize complexity and prevent overfitting.

- **Empirical Risk Minimization (ERM) :** The process of minimizing the loss over the training dataset is known as empirical risk minimization. The goal is to find the set of parameters that minimizes the average loss (risk) on the training data, with the hope that this will generalize well to new data.

- **Optimization Process :**

  - **Forward Pass :** The network processes input data and generates predictions.
  - **Loss Calculation :** The loss function computes the error between predictions and actual labels.
  - **Backward Pass(Backpropagation) :** The network calculates gradients of the loss with respect to its weights and updates them to reduce the loss.
  - **Iteration :** This process repeats over many epochs (full passes through the training data), gradually improving the model's performance.

- **Regularization :** To enhance generalization, regularization terms(such as L1 or L2 penalties) can be added to the objective function. These terms discourage overly complex models that might overfit the training data.

2.4.1.4 Strengths of Supervised Learning

Supervised learning offers several advantages :

- **Predictive Power :** Models can make accurate predictions or classifications on new data if trained well.
- **Human Interpretability :** Since humans label the data, the decisions made by supervised models are often easier to understand and validate.
- **Versatility :** Applicable to a wide range of problems, from medical diagnosis to financial forecasting.
- **Performance Optimization :** With labeled data and clear objectives, models can be fine-tuned for optimal performance using various evaluation metrics.

2.4.1.5 Challenges and Limitations

Despite its strengths, supervised learning has notable limitations :

- **Data Requirements :** Requires large amounts of accurately labeled data, which can be expensive and time-consuming to obtain.
- **Overfitting :** Models may perform well on training data but fail to generalize to new, unseen data if not properly validated.
- **Limited to Known Classes :** Supervised models can only predict classes they have been trained on; they struggle with novel or out-of-distribution data.
- **Human Intervention :** Continual human involvement is needed for labeling, validation, and retraining as new data emerges..
- **Bias and Quality Issues :** Poorly labeled or unbalanced data can introduce bias and reduce model reliability.

From a neural network perspective, supervised learning is a cycle of training on labeled data, validating model choices, and minimizing a loss (objective) function. The careful division of data into training, validation, and test sets ensures robust evaluation and prevents overfitting. Loss functions and objective functions are central to this process, guiding the optimization of network parameters to achieve high predictive accuracy. Regularization and empirical risk minimization further refine the model, ensuring it generalizes well to new, unseen data.

2.4.2 Transfer Learning

Transfer learning is a powerful technique in machine learning where knowledge gained from solving one problem is leveraged to address a different, but related, problem. Instead of starting the learning process from scratch, transfer learning allows a model to use patterns and features it has already learned, significantly reducing the amount of data and computational resources required for the new task. At its core, transfer learning involves taking a model that has been pre-trained on a large dataset—often for a generic task—and adapting it to a new, often more specific, task. For example, a neural network trained on millions of images to recognize everyday objects (like animals, vehicles, or household items) can be repurposed to identify specific medical anomalies in X-ray images. The early layers of such deep neural networks typically learn general features, such as edges or textures, which are useful across a wide range of visual recognition tasks. By reusing these learned features, transfer learning enables rapid development of high-performing models even when the new task has limited labeled data. The process of transfer learning generally follows these steps :

1. **Select a Pre-Trained Model :** Choosing a model that has been trained on a large, relevant dataset. In computer vision, models like ResNet, VGG, or Inception are commonly used; in natural language processing, models like BERT or GPT are popular.

2. **Adapt the Model :** Replace or fine-tune the final layers of the pre-trained model to suit the new task. For instance, the output layer might be changed to reflect the number of classes in the new classification problem.

3. **Fine-Tune with New Data :** Train the adapted model on the new, smaller dataset. Often, only the later layers are updated, while the earlier layers remain unchanged or are updated at a slower rate to preserve previously learned features.

Transfer learning is especially valuable in fields like computer vision and natural language processing, where deep learning models typically require vast amounts of labeled data and computational power to train from scratch. By leveraging pre-trained models, organizations can save significant time and resources, making advanced AI applications more accessible and commercially viable. The advantages of transfer learning includes :

- **Reduced Data Requirements :** It enables effective model training even when only a small amount of labeled data is available for the new task.

- **Faster Training :** Since the model starts from a knowledgeable state, it converges more quickly than models trained from scratch.

- **Improved Performance :** Leveraging knowledge from related tasks often results in better generalization and higher accuracy, especially when the new task is similar to the original one.

In summary, transfer learning enables the rapid deployment of robust models in domains where data is scarce or expensive to obtain. Its impact is evident in the success of applications ranging from image recognition and natural language understanding to generative AI systems like ChatGPT and Google Gemini.

2.4.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a cornerstone of modern deep learning, especially in domains involving visual data such as image and video analysis, object recognition, and medical imaging. Their architecture, inspired by the human visual cortex, leverages spatial hierarchies in data to efficiently extract features and perform complex tasks with high accuracy. Over the past decade, numerous CNN-based architectures have been developed, each introducing innovations to improve performance, efficiency, and adaptability to diverse tasks.

Core Concepts and Building Blocks

- **Convolutional Layer :** The convolutional layer is the fundamental building block of a CNN. It consists of a set of learnable filters (kernels) that slide over the input data, computing dot products to produce feature maps. Each filter is designed to activate in response to specific patterns, such as edges or textures, in localized regions of the input known as receptive fields. The use of shared weights across the spatial domain drastically reduces the number of parameters compared to fully connected networks, enabling deeper architectures and mitigating overfitting.

- **Pooling Layer :** Pooling layers, also known as downsampling layers, reduce the sampling dimensions of feature maps by aggregating information within local neighbourhoods. The most common pooling operations are max pooling and average pooling. Pooling serves to decrease computational complexity, introduce a degree of translational invariance, and control overfitting by reducing the number of parameters.

- **Fully Connected Layer :** After several convolutional and pooling layers, the high-level features are flattened and passed to one or more fully connected(dense) layers. These layers perform the final classification or regression tasks. In image classification, the output is typically a softmax layer that produces class probabilities.

- **Activation Functions :** Non-linear activation functions, such as ReLU(Rectified Linear Unit), are applied after each convolutional or dense layers to introduce non-linearlity, enabling the network to learn complex patterns.

- **Hyperparameters :** Key hyperparameters in CNNs include the number and size of filters, stride, padding, and the number of layers. These settings influence the model's capacity, computational requirements, and performance.

2.4.3.1 Major CNN-Based Architectures

Over the years, several landmark CNN architectures have been proposed, each addressing specific challenges and pushing the boundaries of what is possible in deep learning.Below is an overview of some of the most influential architectures used in this research.

1. **VGG-19 [31]**

   - **Architecture :** VGG-19 consists of 19 weight layers (16 convolutional + 3 fully connected), using only 3×3 convolutional filters and 2×2 max-pooling throughout. This uniformity simplifies the design while increasing depth compared to predecessors like AlexNet. The small receptive fields capture intricate patterns, and ReLU activations enhance non-linearity. The model ends with three fully connected layers for classification.
   - **Performance and Applications :**
     - Achieved 92.7% top-5 accuracy on ImageNet
     - Used in medical imaging (e.g., diabetic retinopathy detection with 96% accuracy)
     - Key strengths: Simplicity and transfer learning efficacy.
     - Limitations: High computational cost (138M parameters) and memory footprint (533MB).
   - **Innovations :**
     - Demonstrated that depth (via stacked 3×3 convolutions) improves feature extraction.
     - Omitted Local Response Normalization (LRN) to reduce training time without sacrificing accuracy.

2. **Inception-ResNet-v2 [32]**

   - **Architecture :** This architecture merges Inception modules with residual connections. Key components:
     - **Stem block :** Initial feature extraction.
     - **Inception-ResNet blocks :** Parallel convolutions (1×1, 3×3, 5×5) with residual shortcuts.
     - **Reduction blocks :** Dimensionality reduction between stages.

     Trained with 164 layers and 56M parameters, it uses batch normalization and dropout for stability
   - **Performance and Applications :**
     - Excels in medical imaging (e.g., 88.7% accuracy in pneumonia detection from X-rays).
     - Outperforms pure Inception networks in object detection backbones (e.g., Faster R-CNN [33]).
     - Strengths: Accelerated convergence and mitigated vanishing gradients via residual links.

- **Innovations :**
  - Replaced filter concatenation with residual additions, boosting gradient flow.

3. **MobileNet-v2 [34]**

   - **Architecture :** Optimized for mobile devices, MobileNet-v2 introduces :
     - **Inverted Residual Blocks :** Expand channels (1×1 conv), apply depthwise conv, then project back.
     - **Linear bottlenecks :** ReLU activations prevent information loss in low-dimensional spaces
     - **Depthwise separable convolutions :** Factorize standard conv into depthwise (spatial) and pointwise (channel-wise) operations, slashing computation by 8–9×.
   - **Performance and Applications :**
     - Achieves near-state-of-the-art accuracy on ImageNet [35] with only 9.4M parameters.
     - Ideal for real-time applications: object detection (MobileNet-SSD), AR/VR, and edge devices.
   - **Innovations :**
     - Linear bottlenecks preserve features in low dimensions.
     - Expansion layers increase capacity without proportional compute cost.

4. **DenseNet [36]**

   - **Architecture :** DenseNets connect each layer to every subsequent layer ("dense blocks"):
     - **Dense Connectivity :** Concatenates feature maps from all prior layers, enhancing gradient flow.
     - **Transition Layers :** 1×1 convs and pooling reduce dimensionality between blocks.
     - **Growth Rate :** Controls new features per layer (e.g., DenseNet-121 has k=32).
   - **Performance and Applications :**
     - DenseNet-201 achieves 77.9% top-1 ImageNet accuracy with 20M parameters.
     - Excels in tasks requiring feature hierarchy (e.g., semantic segmentation).
     - Strengths: Mitigates vanishing gradients and encourages feature reuse.
   - **Innovations :**
     - **Composite Functions :** BatchNorm-ReLU-Conv sequences in each layer.
     - **Bottleneck layers :** 1×1 convolutions before 3×3 convolutions to reduce computation.

Each architecture represents a strategic evolution: VGG-19 prioritized depth, Inception-ResNet-v2 balanced complexity and accuracy, MobileNet-v2 optimized efficiency, and DenseNet [37] maximized

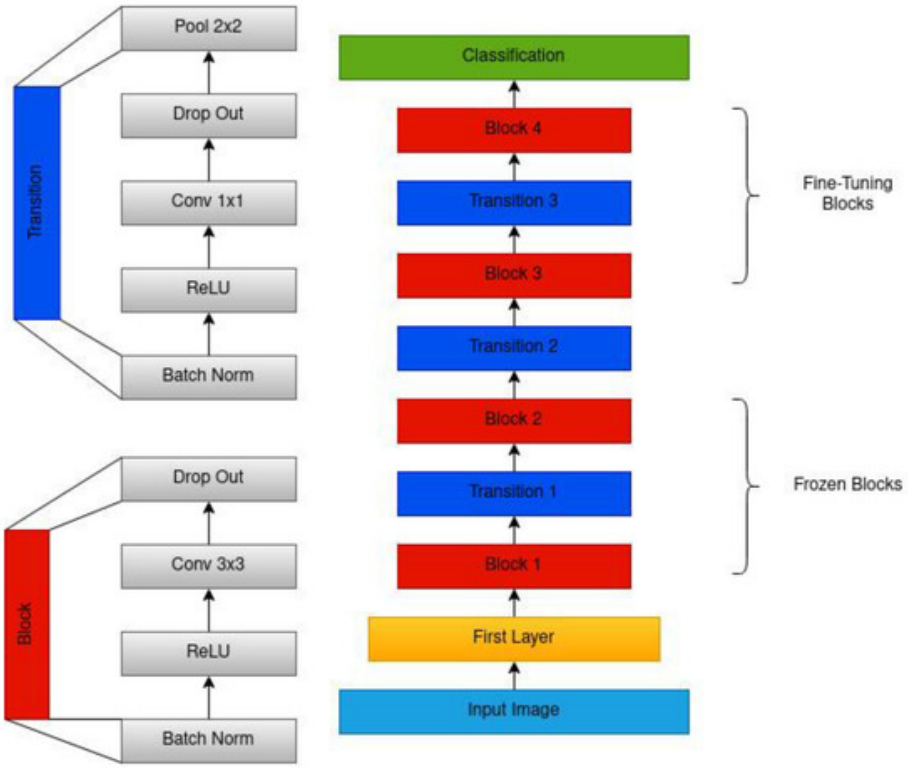


Figure 2.4 DenseNet201

feature propagation. These innovations collectively advanced CNN capabilities across domains from healthcare to embedded systems.

2.4.4 Vision Transformers

Transformers, originally developed for natural language processing, have revolutionized computer vision through architectures like the Vision Transformer (ViT) [30], Data-efficient Image Transformer (DeiT), and Swin Transformer. Each model represents a significant step in adapting transformer-based learning to visual tasks, addressing core challenges such as data efficiency, scalability, and hierarchical representation.

1. **Vanilla Vision Transformer**

   - **Background and Motivation :** Traditional convolutional neural networks (CNNs) have long dominated computer vision, leveraging their inductive biases (e.g., locality, translation invariance) to efficiently learn from relatively modest datasets. However, transformers, with their global self-attention [38] mechanism, have shown remarkable results in NLP. The Vision Transformer (ViT) was the first major attempt to apply a pure transformer architecture

to image recognition, minimizing architectural changes to test how much image structure the model could learn from data alone.

- **Architecture :**
  - **Patch Embedding :** Instead of pixels, ViT treats an image as a sequence of fixed-size non-overlapping patches (e.g., 16x16 pixels). Each patch is flattened and linearly projected into a vector, analogous to word embeddings in NLP transformers.
  - **Position Embedding :** Since transformers are permutation-invariant, learnable position embeddings are added to each patch vector to encode spatial information.
  - **Transformer Encoder :** The sequence of patch embeddings is processed by a standard transformer encoder stack, comprising multi-head self-attention [38] and feed-forward layers.
  - **Classification Token :** A special [CLS] token is prepended to the sequence, and its output embedding is used for classification.
- **Training and Performance :**
  - **Data Requirements :** ViT's lack of strong image-specific inductive biases means it requires very large datasets to generalize well. When trained on ImageNet (1M images), ViT underperforms compared to state-of-the-art CNNs. However, with larger datasets like ImageNet-21k (14M images) or JFT-300M (300M images), ViT matches or outperforms CNNs, achieving up to 88.55% top-1 accuracy on ImageNet and 99.50% on CIFAR-10
  - **Compute Efficiency :** ViT attains these results with fewer computational resources than comparable CNNs when sufficient data is available.
  - **Robustness and Applications :** ViT has demonstrated robustness to input distortions and has been applied to tasks beyond classification, including object detection, segmentation, and even generative modeling.
- **Limitations :**
  - **Data-Hungry :** ViT's main limitation is its need for massive labeled datasets to reach its full potential, making it less accessible for domains with limited data.
  - **Lack of Locality Bias :** Unlike CNNs, ViT does not inherently encode local spatial relationships, which can hinder performance on smaller datasets or tasks requiring fine-grained spatial reasoning.

2. **Data-Efficient Image Transformer**

- **Background and Motivation :** Recognizing ViT's data inefficiency, Meta (formerly Facebook AI) introduced the Data-efficient Image Transformer (DeiT) [39] to democratize transformer-based vision models. DeiT's goal is to enable training high-performance vision transform-

ers on standard datasets like ImageNet, without requiring hundreds of millions of images or massive compute resources.

- **Key Innovations :**
  - **Efficient Training :** DeiT can be trained on a single 8-GPU server in just a few days, achieving 84.2% top-1 accuracy on ImageNet—competitive with leading CNNs.
  - **Augmentation and Regularization :** The training regime borrows heavily from best practices in CNNs, including aggressive data augmentation, optimization tweaks, and regularization strategies to simulate larger datasets.
  - **Knowledge Distillation with Distillation Token :**
    - **Teacher-Student Framework :** DeiT introduces a novel distillation approach where a CNN (teacher) guides the transformer (student) during training.
    - **Distillation Token :** A unique learnable token is added to the transformer's input sequence. This token is trained to predict the teacher's output, while the standard [CLS] token predicts the ground-truth class. This separation allows the model to balance learning from both the teacher and the data.
    - This transformer-specific distillation improves performance without requiring changes to the transformer architecture, and it is more effective than traditional distillation approaches.
  - **Performance and Impact :**
    - **Data Efficiency :** DeiT achieves state-of-the-art performance on ImageNet using only 1.2 million images, a significant reduction compared to the data requirements of vanilla ViT.
    - **Accessibility :** By lowering the computational and data barriers, DeiT makes transformer-based vision models accessible to a broader research community.
  - **Limitations :**
    - **Still Relies on Augmentation :** While more data-efficient than ViT, DeiT's performance is still heavily dependent on sophisticated augmentation and regularization.
    - **No Explicit Locality :** Like ViT, DeiT does not encode spatial locality, although its training regime compensates for this to some extent.

3. **Swin Transformer**

- **Background and Motivation :** While ViT and DeiT demonstrated the viability of transformers in vision, they struggled with two key issues:
  - Inefficiency at high image resolutions due to quadratic complexity in self-attention.
  - Lack of hierarchical feature maps, which are crucial for dense prediction tasks like object detection and segmentation.

The Swin Transformer [40] (Shifted Window Transformer) was introduced by Microsoft Research to address these limitations and serve as a general-purpose backbone for a wide range of vision tasks.

- **Key Innovations :**
  - **Hierarchial Representation :** Swin Transformer builds hierarchical feature maps by merging image patches in deeper layers, similar to how CNNs downsample feature maps. This enables multi-scale representation, essential for dense prediction tasks
  - **Shifted Window Attention :**
    - **Local Self-Attention :** Instead of global self-attention, Swin computes self-attention within local windows, dramatically reducing computational complexity from quadratic to linear with respect to image size.
    - **Window Shifting :** To allow cross-window information flow, windows are shifted between layers, enabling interactions between neighboring regions and preserving global context over depth.
  - **Scalability and Flexibility :** Swin's design supports variable input sizes and can be scaled up for larger models or higher-resolution images without prohibitive computational costs.
- **Performance and Applications :**
  - **General-Purpose Backbone :** Swin Transformer achieves state-of-the-art results across a variety of vision tasks, including image classification, object detection, and semantic segmentation.
  - **Efficiency :** By combining hierarchical design with local attention, Swin is both computationally efficient and highly performant, making it suitable for real-world applications and deployment.
- **Limitations :**
  - **Complexity :** The shifted window mechanism and hierarchical merging introduce additional architectural complexity compared to vanilla ViT.
  - **Implementation Overhead :** Efficiently implementing the window shifting and hierarchical merging requires careful engineering, especially for large-scale models.

2.5 Experiments and Implementation Details

The experimental framework was meticulously designed to evaluate the effectiveness of both convolutional neural networks (CNNs) and transformer-based architectures in medical image analysis. The original input images, which ranged from 3 to 5 megapixels in resolution, were first resized to a uniform dimension of 224×224 pixels. This resizing step ensured compatibility with the standard input requirements of the pre-trained models and facilitated efficient batch processing during training. For

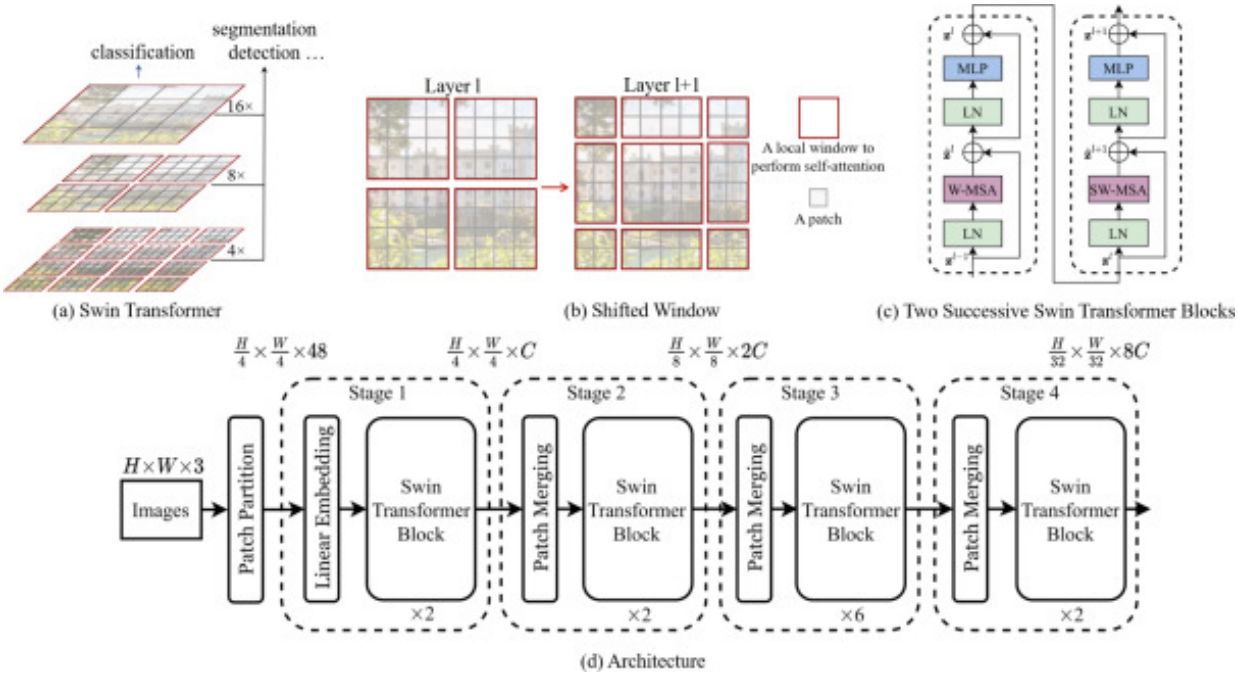


Figure 2.5 An overview of the Swin Transformer. (a) Hierarchical feature maps for reducing computational complexity. (b) Shifted window approach which was used when calculating self-attention. (c) Two successive Swin Transformer Blocks which presented at each stage. (d) Core architecture of the Swin.

CNN-based experiments, several state-of-the-art architectures pre-trained on the ImageNet [35] dataset were utilized, including VGG19, Inception ResNet-V2, MobileNet-V2, and the DenseNet [37] family. Transfer learning was employed to leverage the rich feature representations learned from large-scale natural image datasets. Fine-tuning strategies were carefully tailored to each architecture: in the case of VGG19, Inception ResNet-V2, and MobileNet-V2, approximately one-third of the network layers were frozen. This approach preserved the ability of the lower layers to extract fundamental low-level features such as edges, lesion size, and color variation, which are crucial for medical image interpretation. For the DenseNet [37] family, a more conservative strategy was adopted by freezing half of the network layers, further emphasizing the retention of foundational feature extraction while allowing the remaining layers to adapt to the specific characteristics of the target dataset. The training process for all models was standardized to ensure fair comparison. The Adam optimizer was selected for its robust performance and adaptive learning rate capabilities, with an initial learning rate set to 0.0001. Each model was trained for 50 epochs, utilizing a batch size of 16 to balance computational efficiency and gradient stability. Categorical cross-entropy loss was employed as the objective function, reflecting the multi-class nature of the classification task. To prevent overfitting, early stopping was implemented by monitoring the validation loss and halting training if no improvement was observed over a set number of epochs. For transformer-based architectures, a cosine annealing learning rate scheduler was incorporated. This scheduler featured a warm-up phase, during which the learning rate increased linearly,

followed by a gradual decrease according to a cosine decay schedule. This approach has been shown to improve convergence and generalization in deep learning models. All experiments were conducted using PyTorch version 2.0.1, leveraging the computational power of an Nvidia A100 GPU. This consistent hardware and software environment ensured reproducibility and efficient execution of all training and evaluation procedures.

2.5.1 Experiment Setup

The primary metrics to measure the performance of models were Precision, Recall, Specificity and F1-score.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2.3}$$

$$\text{Recall (Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2.4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{2.5}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}, \tag{2.6}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2.7}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{2.8}$$

Precision measures the proportion of the model's positive predictions that were correct. Recall (sensitivity) measures the proportion of true positive samples correctly identified by the model. Specificity measures the proportion of true negative samples correctly identified by the model. The F1-score is the harmonic mean of precision and recall. All the metrics were calculated using true positives(TP), true negatives(TN), false positive(FP),and false negative(FN) samples. Additionally, AUC i.e area under the receiver operator characteristic curve, was measured as secondary metric which shows the plot of the true positive rate(TPR) against the False Positive rate(FPR).

2.6 Results

Six pre-trained convolutional neural networks (CNNs) were re-trained using a dataset comprising 1,344 oral cavity images to evaluate their effectiveness in detecting suspicious lesions. Each of these neural networks exhibited varying levels of performance on the test set, as summarized in Table 2.2. Among the tested models, the DenseNet family, specifically DenseNet201, emerged as the top performer. DenseNet201 achieved impressive metrics, with a precision of 86%, recall (sensitivity) of 85%,

specificity of 83%, and an F1-score of 86%. These results indicate that DenseNet201 is highly reliable in accurately identifying suspicious lesions while minimizing both false positives and false negatives.

The pre-trained networks evaluated in this study differed significantly in their architectures, particularly in terms of the number of layers and the required input image size. These structural differences play a crucial role in determining the models' overall performance, computational speed, and memory requirements. When selecting a model for clinical or real-world deployment, it is essential to consider the trade-off between these factors. A model with high accuracy but excessive computational demands may not be practical for rapid or resource-constrained applications.

The DenseNet family stood out not only for its superior detection performance but also for its balance between speed and network size. DenseNets are known for their efficient use of parameters and ability to facilitate feature reuse through dense connections, which likely contributed to their success in this application. In contrast, VGG19, another well-known CNN architecture, demonstrated the lowest performance among the models tested, highlighting the importance of network design in medical image analysis tasks.

Table 2.2 Comparison of convolution-style architectures on the internal test dataset (n = 422) in Table 2.1. Macro-averaged precision, recall (sensitivity), specificity, and F1-score are reported.

Method	Parameters	Precision	Recall	F1-Score	Specificity
VGG19	138M	0.69	0.68	0.68	0.68
InceptionResNet-V2	56M	0.72	0.72	0.72	0.72
MobileNet-V2	9.4M	0.75	0.75	0.75	0.73
DenseNet121	8M	0.85	0.85	0.85	0.83
DenseNet169	14M	0.84	0.83	0.84	0.78
DenseNet201	**20M**	**0.86**	**0.85**	**0.86**	**0.83**

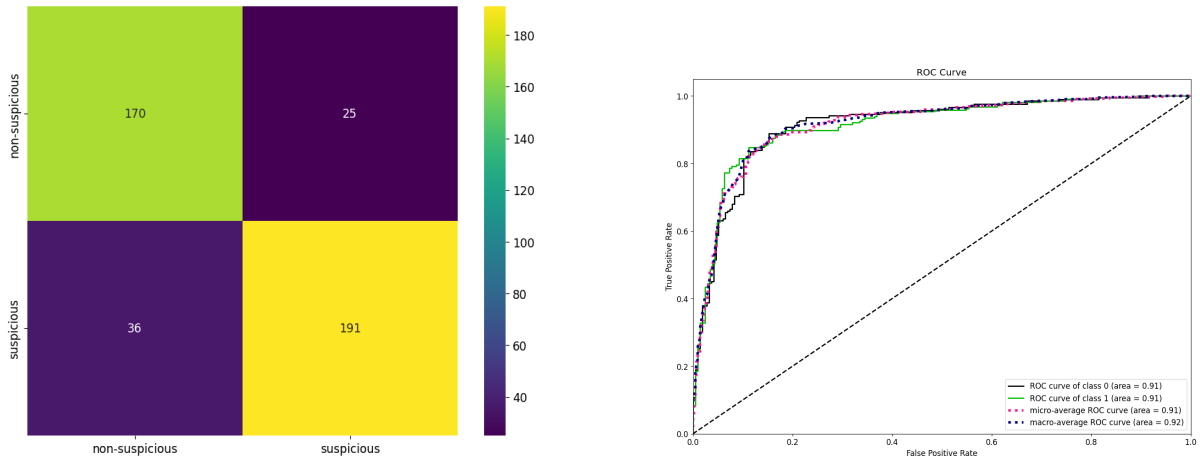
As an alternative approach to conventional convolutional neural networks (CNNs), vision transformer (ViT) architectures were also explored for the task of detecting suspicious lesions in oral cavity images. Vision transformers [41] represent a significant shift in deep learning for image analysis, as they are based on the concept of self-attention rather than convolutional operations. This allows them to capture long-range dependencies and contextual information more effectively across the entire image.

In this study, three variants of Swin Transformers—tiny, small, and base—were evaluated alongside the original Vision Transformer (ViT) and Data-efficient Image Transformer (DeiT) models, with their comparative results presented in Table 2.3. Swin Transformers consistently outperformed both ViT and DeiT, achieving approximately a 10% increase in key performance metrics, including precision, recall, specificity, and F1-score. Among the Swin Transformer variants, the base model demonstrated the highest performance, achieving precision and recall (sensitivity) of 86%, specificity of 83%, and an F1-score of 86%. These results match the best performance observed with DenseNet201, indicating that transformer-based architectures are highly effective for this medical imaging task.

However, a notable distinction between these top-performing models lies in their network size and computational complexity. The Swin Transformer (base) model contains 88 million parameters, which

Table 2.3 Comparison of transformer-style architectures on the internal test dataset (n = 422) in Table 2.1. Macro-averaged precision, recall (sensitivity), specificity, and F1-score are reported.

Method	Parameters	Precision	Recall	F1-Score	Specificity
ViT	86M	0.77	0.77	0.77	0.77
DeiT	86M	0.77	0.75	0.75	0.76
Swin(Tiny)	29M	0.84	0.84	0.84	0.73
Swin(Small)	50M	0.85	0.85	0.85	0.75
Swin(Base)	**88M**	**0.86**	**0.86**	**0.86**	**0.83**

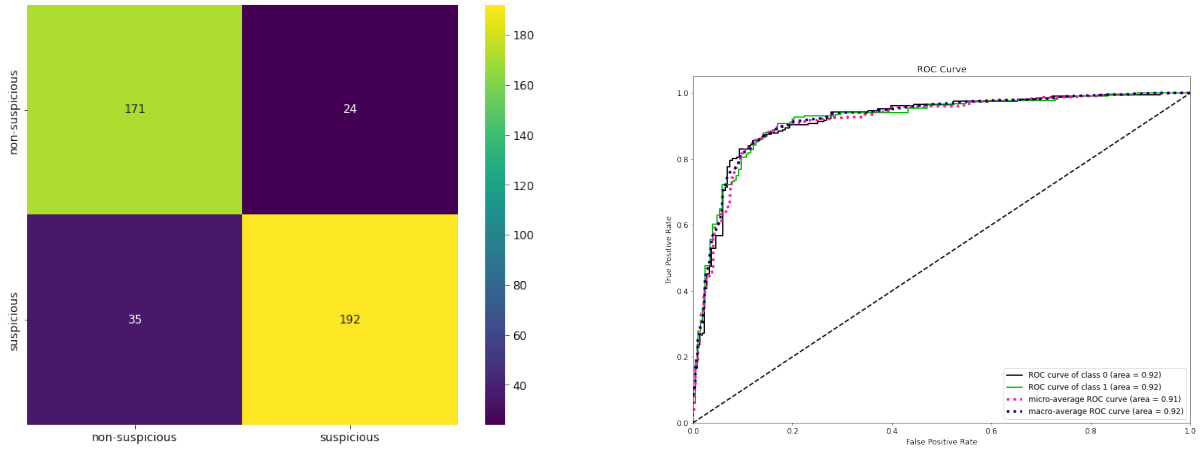


(a) (b)

Figure 2.6 Confusion Matrix and ROC Curve (a) DenseNet201 (b) ROC Curve

is substantially larger than the 20 million parameters of DenseNet201, as shown in Tables 2.2 and 2.3. This difference has important implications for practical deployment, as larger models generally require more memory and computational resources, potentially limiting their use in resource-constrained environments such as point-of-care or mobile applications. The confusion matrix of the best-performing models, DenseNet201 and Swin Transformer (base), shows only subtle differences with AUC values greater than 90% for both cases. The ROC curve shows the relationship between false positive and true positive rates and is used for calculating the AUC (area). Class 0 is non-suspicious, and Class 1 is suspicious.

Clinically, most false positives in the internal test set were associated with cases exhibiting tobacco stains, physiologic melanosis, aphthous ulcers, and various periodontal diseases such as gingivitis, abscesses, and recession (see Figure 2.8(A)). Some discrepancies may have resulted from challenges in lesion localization, anatomical topography, or normal physiological variations. Regarding false negatives, the majority involved lesions that were ultimately diagnosed as non-suspicious, including those that appeared as early speckled (white-red) areas, gingival desquamation, or traumatic keratosis (refer to Figure 2.8(B)). When combined with conventional oral examination (COE) by general dentists, the rate



(a) (b)

Figure 2.7 Confusion Matrix and ROC curve of Swin Transformer(base) (a) Swin(base) Transformer (b) ROC curve

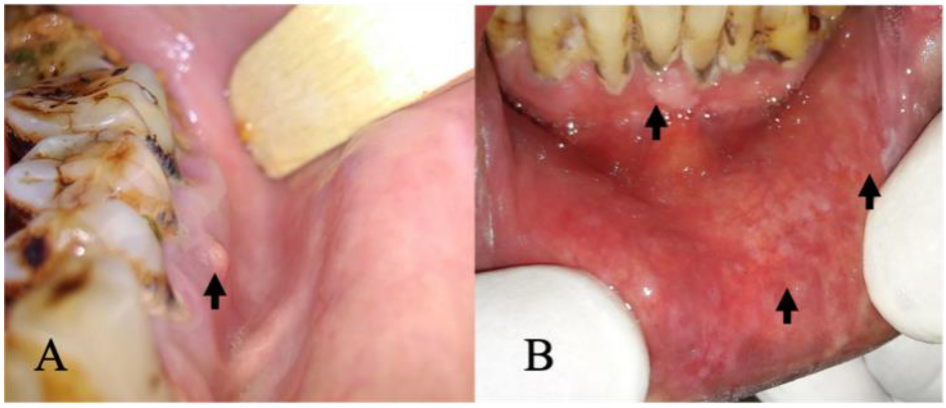


Figure 2.8 (A) False positive: intraoral image showing elevated lesion suggestive of periapical abscess (indicated by arrow). (B) False negative: intraoral image showing lower labial mucosa with white areas (indicated by arrow) and periodontitis in the lower anterior region (indicated by arrow).

of false positives can be effectively minimized. However, the primary challenge remains in reducing false negatives and improving the sensitivity of the AI system.

A five-fold cross-validation (CV) was conducted for both DenseNet201 and Swin Transformer (base) to assess the generalizability of these models across different training and validation splits. For this process, the training and validation datasets described in Table 2.1 were combined and randomly divided into five folds, ensuring class balance was maintained throughout. The average model performance test sets (Figure 2.9) was calculated and reported with 95% confidence intervals (CI) (see Table 2.4).

DenseNet201 achieved the highest average F1-score of 0.84 (CI: 0.79–0.89), while the Swin Transformer (base) followed closely with an average F1-score of 0.83 (CI: 0.78–0.88). The Youden Index, which evaluates a diagnostic test's ability to balance sensitivity and specificity, was 0.71 for

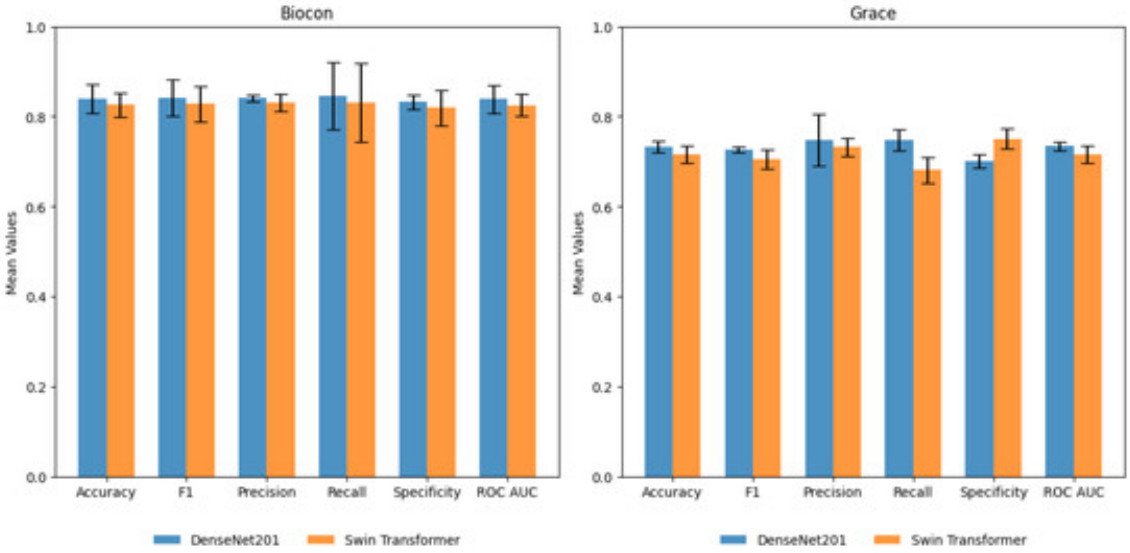


Figure 2.9 Performance of DenseNet201 and Swin Transformer (base) on the Biocon and Grace test sets. The average value of performance metrics with a 95% confidence interval are shown.

DenseNet201 and 0.67 for Swin Transformer (base). Notably, the confidence interval for recall (sensitivity) was wider than that for specificity, indicating greater variability and uncertainty in the models' ability to detect suspicious cases. multirow booktabs array

It is important to highlight that a noticeable drop in performance occurred in only one out of the five cross-validation runs, suggesting that the models are generally stable but may be sensitive to certain data splits. This observation underscores the need for further data sampling to obtain more precise confidence interval estimates and to better understand the models' variability across different subsets.

Additionally, these models were evaluated on an independent test set I, comprising 440 images sourced from the Grace Foundation. On this dataset, DenseNet201 achieved an average recall (sensitivity) of 0.75 (CI: 0.68–0.82), specificity of 0.70 (CI: 0.68–0.72), and an F1-score of 0.73 (CI: 0.67–0.78). DenseNet201 outperformed the Swin Transformer on test set I, as illustrated in Figure 2.9. This independent set included images captured by untrained field health workers (FHWs), resulting in considerable variability in image quality and focus. The substantial decline in data quality contributed to a noticeable drop in the AI models' performance. Moreover, the presence of a small subset of advanced lesions or oral cancers, along with differences in lesion localization and surface characteristics, further influenced the rates of false negatives and false positives.

In contrast, the internal test set from Biocon was composed of images taken by trained FHWs, ensuring consistent image quality and close resemblance to the training and validation datasets. This similarity in data quality contributed to the stable performance of the AI models on the internal test set. The findings from the independent test set I underscore the necessity for more robust model training strategies. Incorporating greater variability into the training dataset can help improve the models' performance and enhance their clinical utility for routine screening of oral potentially malignant disorders

Table 2.4 Performance of DenseNet201 and Swin Transformer (Base) on Biocon and Grace test datasets. Values are reported as point estimate (CI lower–upper).

Biocon (test)					
DenseNet201	0.84 (CI 0.80–0.88)	0.84 (CI 0.83–0.85)	0.85 (CI 0.75–0.94)	0.84 (CI 0.79–0.89)	0.83 (CI 0.81–0.85)
Swin Transformer (base)	0.83 (CI 0.79–0.86)	0.83 (CI 0.80–0.85)	0.83 (CI 0.72–0.94)	0.83 (CI 0.78–0.88)	0.82 (CI 0.77–0.87)
Grace (test)					
DenseNet201	0.73 (CI 0.72–0.75)	0.75 (CI 0.68–0.82)	0.75 (CI 0.68–0.82)	0.73 (CI 0.67–0.78)	0.70 (CI 0.68–0.72)
Swin Transformer (base)	0.72 (CI 0.69–0.74)	0.73 (CI 0.71–0.76)	0.68 (CI 0.65–0.72)	0.70 (CI 0.68–0.73)	0.75 (CI 0.72–0.78)

(OPMDs). Expanding the diversity of images—by including a wider range of lesion types, imaging conditions, and operator experience—will likely make AI models more resilient and effective in real-world screening scenarios.

2.7 Explainability

Gradient-weighted Class Activation Mapping (Grad-CAM) [42] is an explainable AI technique used to visualize which regions of an input image most influence a deep neural network's prediction. Grad-CAM works by computing the gradients of the target class score with respect to the feature maps of the last convolutional layer. These gradients are globally averaged to obtain importance weights, which are then used to create a heatmap highlighting the most relevant areas for the prediction. This heatmap, when overlaid on the original image, helps interpret and verify the model's decision-making process by showing where the model is "looking".

To provide visual explanations for model decisions, class-activation maps [43] were generated using Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM results demonstrated that the top-performing models, DenseNet201 and Swin Transformer (base), concentrated on the relevant regions of the images when making predictions (see Figure 2.10). However, the resulting heatmaps tended to be broad, which is likely due to the models being trained with image-level labels rather than precise region-of-interest annotations. This suggests that incorporating region-specific annotations in future training could further refine the model's focus and improve interpretability.

Figure 2.10 GradCAM visual explanation for the model decision. The colour heatmap highlights the areas in the input image contributing to the decision made by the model, with red regions representing a high score for the class.

2.8 Summary

This study demonstrates the promising application of artificial intelligence (AI) models in diagnosing oral potentially malignant disorders (OPMDs) and oral cancers using images of the oral cavity. By leveraging lightweight deep learning frameworks, the proposed solution enables efficient and accessible screening, requiring only smartphone-captured images. The AI models were evaluated on multiple independent test sets, revealing strong diagnostic potential but also highlighting the need to train on diverse, heterogeneous, and sometimes noisy data to ensure robust performance in real-world conditions.

The simplicity and portability of the solution make it especially suitable for deployment in resource-limited and remote settings, where access to specialized care is often lacking. AI-based screening can empower frontline healthcare workers (FHWs) to identify high-risk patients and facilitate timely referrals, potentially leading to earlier interventions and improved patient outcomes. The ability to automate image analysis and deliver rapid, accurate results directly at the point of care represents a significant advance over traditional methods, which often require expert interpretation and specialized equipment.

Chapter 3

Conclusion and Future Works

The early detection of oral potentially malignant disorders (OPMDs) and oral cancer is a crucial public health challenge, especially in countries like India where the prevalence of these conditions is high and access to specialized care is limited. Delayed diagnosis often leads to advanced-stage disease, increased mortality, and a significant reduction in the quality of life for patients. Traditional screening methods, while effective, are constrained by the limited availability of trained specialists and the logistical difficulties of reaching remote or underserved communities. In this context, the integration of artificial intelligence (AI) into community-based screening programs emerges as a transformative solution that can bridge these gaps and significantly improve early detection rates.

This thesis focused on developing and validating AI-assisted screening methods for the detection of OPMDs using photographic images of the oral cavity, specifically targeting the Indian population. The study leveraged recent advances in deep learning, evaluating the performance of lightweight models such as DenseNet201 and the Swin (base) Transformer. These models are particularly suited for deployment in resource-constrained and mobile settings, making them ideal for large-scale community screening initiatives. The models were trained and tested on a robust dataset of oral cavity images, simulating the variability and challenges encountered in real-world community screening programs.

The results of the study are promising. Both DenseNet201 and Swin Transformer models demonstrated high performance, with F1-scores of 0.84 and 0.83, respectively, on the internal test set. These findings underscore the potential of AI-based solutions to accurately identify suspicious oral lesions using only white light photographic images, eliminating the need for additional imaging modalities or specialized equipment. DenseNet201, in particular, stands out for its parameter efficiency, ease of training, and ability to mitigate the vanishing gradient problem, making it especially suitable for mobile health applications. The Swin Transformer complements this by dynamically focusing on the most informative regions of the image, capturing both local and global features through its hierarchical attention mechanism.

A key strength of the proposed approach lies in the simplicity of the imaging protocol. By standardizing the process to ensure consistency in lesion positioning and focal distance, the need for specialized expertise in image acquisition is minimized. This enables trained community healthcare workers to conduct large-scale screening with minimal technical barriers, thereby expanding the reach and impact

of early detection programs. The ability of the AI models to operate on standard smartphone-captured images further enhances their accessibility and scalability, making them particularly well-suited for low-resource settings.

Despite these strengths, the study also identifies several challenges and limitations. The performance of the models declines when tested on images with greater variability in quality, such as those captured by untrained frontline healthcare workers, highlighting the need for further model refinement to ensure robust generalizability across diverse imaging conditions and user skill levels. Additionally, the current validation was conducted on a limited number of images and within a relatively homogeneous demographic, emphasizing the need for broader validation across different regions, populations, and device types to establish the models' reliability and applicability on a larger scale.

Nevertheless, the evidence presented in this thesis strongly supports the viability of AI-assisted screening using white light imaging as a practical and effective strategy for the early detection of OPMDs in India. By simplifying the screening process and reducing reliance on specialist expertise, this approach has the potential to significantly improve patient outcomes, reduce the burden of advanced oral cancers, and ultimately save lives. The deployment of such solutions aligns with the broader movement toward digital health and AI-powered medical imaging, which are increasingly recognized as pivotal for population health screening in India and other similar settings.

In summary, this research lays a solid foundation for the integration of AI-based image analysis into community screening programs for oral cancer and its precursors. The demonstrated effectiveness of lightweight deep learning models, combined with the practical advantages of smartphone-based imaging, positions this approach as a scalable and impactful solution for addressing the pressing public health challenge of oral cancer in India.

Future Work

Looking ahead, several pathways emerge for advancing and scaling the impact of AI-assisted oral cancer screening in India and similar settings. The expansion and diversification of the image dataset is a primary area of focus. By collecting a larger and more varied set of oral cavity images that reflect a broader range of demographics, geographic regions, and device types, the model's robustness and generalizability can be significantly improved. Including images that capture a wide spectrum of disease profiles, lighting conditions, and oral mucosal presentations—such as variations in pigmentation and the presence of teeth or retractors—will help the model better distinguish OPMDs from other oral conditions.

Enhancing model performance and interpretability is another critical direction. Further optimization of deep learning architectures, including the exploration of additional lightweight models and hybrid approaches, may yield improvements in both accuracy and computational efficiency. Incorporating object detection and localization techniques, such as the use of bounding box annotations to specify regions of interest, can help the model focus on relevant lesion areas and reduce the influence of background noise. Additionally, developing explainable AI methods that provide visual or textual rationales for model predictions will enhance clinical trust and facilitate adoption by healthcare workers and specialists.

Real-world validation and clinical integration are essential next steps. Large-scale, prospective validation studies in diverse community settings are needed to assess the real-world performance and impact of the AI-assisted screening solution. Collaborations with public health authorities, non-governmental organizations, and academic institutions can facilitate the integration of the AI tool into existing oral health programs and infrastructure. Establishing standardized protocols for image acquisition, data handling, and referral pathways will help ensure consistency and quality across different screening sites.

Operational and ethical challenges must also be addressed to ensure the responsible deployment of AI-powered screening tools. Data privacy, security, and compliance with regulatory frameworks are paramount. User-friendly interfaces and training modules for community healthcare workers will support effective adoption and minimize errors in image capture and interpretation. It is also important to address potential biases in the dataset and model predictions, particularly with respect to underrepresented groups, to ensure equitable access and outcomes.

The integration of AI screening results with electronic health records and clinical decision support systems presents another promising avenue. Linking AI outputs with patient records can facilitate seamless referral, follow-up, and monitoring of high-risk patients. Incorporating patient habit history, such as tobacco and betel nut use, and other risk factors into the model input may further enhance predictive accuracy and enable personalized risk stratification.

Exploring multimodal and longitudinal approaches could further improve diagnostic sensitivity and specificity. While this study focused on white light imaging, integrating additional imaging modalities, such as autofluorescence or narrow-band imaging, could provide complementary information. Longitudinal studies that track lesion progression over time, combined with AI analysis, may offer valuable insights into malignant transformation risk and optimal intervention timing.

Policy advocacy and public health impact assessment are also vital components of future work. Engaging with policymakers to advocate for the inclusion of AI-assisted screening in national oral health strategies can accelerate the scale-up and sustainability of these innovations. Evaluating the cost-effectiveness, health outcomes, and social impact of AI-powered screening programs will provide the evidence needed to support widespread adoption and investment.

Continuous model updating and the establishment of feedback loops are necessary to maintain the relevance and effectiveness of the AI solution. Ongoing data collection, model retraining, and performance monitoring will ensure that the AI tool remains up-to-date and responsive to evolving epidemiological trends and technological advancements. Incorporating feedback from frontline users and specialists into the development cycle will drive user-centered improvements and foster ongoing engagement.

In conclusion, the integration of AI-assisted screening methods into oral health programs in India holds significant promise for transforming the early detection and management of OPMDs and oral cancer. The work presented in this thesis demonstrates the feasibility and effectiveness of lightweight deep learning models applied to smartphone-based white light images, paving the way for scalable, accessible, and impactful solutions in resource-constrained settings. By addressing the outlined future

directions, researchers, clinicians, and policymakers can collectively advance the field toward realizing the full potential of AI in improving oral health outcomes and reducing the burden of cancer in India and similar contexts worldwide.

List of Related Publications

[P1] **Vivek Talwar** , Pragya Singh , Nirza Mukhia , Anupama Shetty , Praveen Birur , Karishma M. Desai , Chinnababu Sunkavalli , Konala S. Varma , Ramanathan Sethuraman , C. V. Jawahar and P. K. Vinod **"AI-Assisted Screening of Oral Potentially Malignant Disorders Using Smartphone-Based Photographic Images"**, in proceedings of *Cancers*, 2023.

Related co-author publications:

[P1] Karishma Madhusudan Desai , Pragya Singh , Mahima Smriti , **Vivek Talwar** , Manav Chaudhary , George Paul , Subhas Chandra Kolli , Parisa Sai Raghava , Golla Vamshi Krishna , C. V. Jawahar , P. K. Vinod , Varma Konala and Ramanathan Sethuraman **"Screening of oral potentially malignant disorders and oral cancer using deep learning models"**, in proceedings of *Nature Scientific Reports*, 2025.

Bibliography

[1] S. Warnakulasuriya, O. Kujan, J. M. Aguirre-Urizar, J. V. Bagan, M. Á. González-Moles, A. R. Kerr, G. Lodi, F. W. Mello, L. Monteiro, G. R. Ogden *et al.*, "Oral potentially malignant disorders: A consensus report from an international seminar on nomenclature and classification, convened by the who collaborating centre for oral cancer," *Oral diseases*, vol. 27, no. 8, pp. 1862–1880, 2021.

[2] K. R. Coelho, "Challenges of the oral cancer burden in india," *Journal of cancer epidemiology*, vol. 2012, no. 1, p. 701932, 2012.

[3] P. Kumari, P. Debta, and A. Dixit, "Oral potentially malignant disorders: etiology, pathogenesis, and transformation into oral cancer," *Frontiers in pharmacology*, vol. 13, p. 825266, 2022.

[4] S. Sriharikrishnaa, P. S. Suresh, and S. Prasada K, "An introduction to fundamentals of cancer biology," in *Optical Polarimetric Modalities for Biomedical Research*. Springer, 2023, pp. 307–330.

[5] F. Kamangar, G. M. Dores, and W. F. Anderson, "Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world," *Journal of clinical oncology*, vol. 24, no. 14, pp. 2137–2150, 2006.

[6] S. A. Fedewa, A. G. Sauer, R. L. Siegel, and A. Jemal, "Prevalence of major risk factors and use of screening tests for cancer in the united states," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 24, no. 4, pp. 637–652, 2015.

[7] A. Elangovan and T. Jeyaseelan, "Medical imaging modalities: a survey," in *2016 International Conference on emerging trends in engineering, technology and science (ICETETS)*. ieee, 2016, pp. 1–4.

[8] X. Ou, X. Chen, X. Xu, L. Xie, X. Chen, Z. Hong, H. Bai, X. Liu, Q. Chen, L. Li *et al.*, "Recent development in x-ray imaging technology: Future and challenges," *Research*, 2021.

[9] P. J. Withers, C. Bouman, S. Carmignato, V. Cnudde, D. Grimaldi, C. K. Hagen, E. Maire, M. Manley, A. Du Plessis, and S. R. Stock, "X-ray computed tomography," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 18, 2021.

[10] D. Formica and S. Silvestri, "Biological effects of exposure to magnetic resonance imaging: an overview," *Biomedical engineering online*, vol. 3, pp. 1–12, 2004.

[11] P. N. Wells, "Ultrasound imaging," *Physics in medicine & biology*, vol. 51, no. 13, p. R83, 2006.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," *Nature Reviews Neuroscience*, vol. 21, no. 6, pp. 335–346, 2020.

[14] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[15] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature medicine*, vol. 28, no. 1, pp. 31–38, 2022.

[16] W.-Q. Wei, C. L. Leibson, J. E. Ransom, A. N. Kho, P. J. Caraballo, H. S. Chai, B. P. Yawn, J. A. Pacheco, and C. G. Chute, "Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 219–224, 2012.

[17] K. E. Paik, R. Hicklen, F. Kaggwa, C. V. Puyat, L. F. Nakayama, B. A. Ong, J. N. Shropshire, and C. Villanueva, "Digital determinants of health: Health data poverty amplifies existing health disparities—a scoping review," *PLOS Digital Health*, vol. 2, no. 10, p. e0000313, 2023.

[18] N. P. Birur, K. Gurushanth, S. Patrick, S. P. Sunny, S. A. Raghavan, S. Gurudath, U. Hegde, V. Tiwari, V. Jain, M. Imran *et al.*, "Role of community health worker in a mobile health program for early detection of oral cancer," *Indian Journal of Cancer*, vol. 56, no. 2, pp. 107–113, 2019.

[19] V. Talwar, P. Singh, N. Mukhia, A. Shetty, P. Birur, K. M. Desai, C. Sunkavalli, K. S. Varma, R. Sethuraman, C. Jawahar *et al.*, "Ai-assisted screening of oral potentially malignant disorders using smartphone-based photographic images," *Cancers*, vol. 15, no. 16, p. 4120, 2023.

[20] P. Mohan, "Assessment of the feasibility of opportunistic screening for oral potentially malignant disorders and oral cancer at dental colleges in india: A public health initiative from bengaluru," Ph.D. dissertation, University of Canterbury, 2022.

[21] P. Gupta, R. Bhonsle, P. Murti, D. Daftary, F. S. Mehta, and J. Pindborg, "An epidemiologic assessment of cancer risk in oral precancerous lesions in india with special reference to nodular leukoplakia," *Cancer*, vol. 63, no. 11, pp. 2247–2252, 1989.

[22] S. Warnakulasuriya and A. Ariyawardana, "Malignant transformation of oral leukoplakia: a systematic review of observational studies," *Journal of Oral Pathology & Medicine*, vol. 45, no. 3, pp. 155–166, 2016.

[23] M. Essat, K. Cooper, A. Bessey, M. Clowes, J. B. Chilcott, and K. D. Hunter, "Diagnostic accuracy of conventional oral examination for detecting oral cavity cancer and potentially malignant disorders in patients with clinically evident oral lesions: Systematic review and meta-analysis," *Head & neck*, vol. 44, no. 4, pp. 998–1013, 2022.

[24] J. M. Aguirre-Urizar, I. Lafuente-Ibáñez de Mendoza, and S. Warnakulasuriya, "Malignant transformation of oral leukoplakia: systematic review and meta-analysis of the last 5 years," *Oral Diseases*, vol. 27, no. 8, pp. 1881–1895, 2021.

[25] B. Song, S. Sunny, R. D. Uthoff, S. Patrick, A. Suresh, T. Kolur, G. Keerthi, A. Anbarani, P. Wilder-Smith, M. A. Kuriakose *et al.*, "Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning," *Biomedical optics express*, vol. 9, no. 11, pp. 5318–5329, 2018.

[26] S. Vinayahalingam, N. van Nistelrooij, R. Rothweiler, A. Tel, T. Verhoeven, D. Tröltzsch, M. Kesting, S. Bergé, T. Xi, M. Heiland *et al.*, "Advancements in diagnosing oral potentially malignant disorders: leveraging vision transformers for multi-class detection," *Clinical oral investigations*, vol. 28, no. 7, p. 364, 2024.

[27] G. Tanriver, M. Soluk Tekkesin, and O. Ergen, "Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders," *Cancers*, vol. 13, no. 11, p. 2766, 2021.

[28] C. L. Dunlap and B. Barker, "A guide to common oral lesions," *Department of Oral and Maxillofacial Pathology UMKC School of Dentistry*, 2016.

[29] N. Birur, B. Song, S. Sunny, G. Keerthi, P. Mendonca, N. Mukhia *et al.*, "Field validation of deep learning based point-of-care device for early detection of oral malignant and potentially malignant disorders. sci rep 12 (1): 14283," 2022.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[31] S. Karen and A. Zisserman, "Very deep convolutional networks for large-scale image recognition. arxiv 2014," *arXiv preprint arXiv:1409.1556*.

[32] S. Christian, I. Sergey, V. Vincent, and A. Alex, "Inception-v4 inception-resnet and the impact of residual connections on learning. 2016," *arXiv preprint arXiv:1602.07261*.

[33] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[34] M. Z. A. Z. Liang-Chieh, C. M. Sandler, and A. Howard, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[36] H. Gao, L. Zhuang, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, no. 2, 2017, pp. 3–12.

[37] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, P. Jantana, and S. Vicharueang, "Ai-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer," *Plos one*, vol. 17, no. 8, p. e0273508, 2022.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[39] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[40] L. Ze, L. Yutong, C. Yue, H. Han, W. Yixuan, Z. Zheng, L. Stephen, and G. Baining, "Swin transformer: Hierarchical vision transformer using shifted windows, 2021," *CoRR, abs/2103.14030*.

[41] T. Flügge, R. Gaudin, A. Sabatakakis, D. Tröltzsch, M. Heiland, N. van Nistelrooij, and S. Vinayahalingam, "Detection of oral squamous cell carcinoma in clinical photographs using a vision transformer," *Scientific Reports*, vol. 13, no. 1, p. 2296, 2023.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[43] S. Camalan, H. Mahmood, H. Binol, A. L. D. Araujo, A. R. Santos-Silva, P. A. Vargas, M. A. Lopes, S. A. Khurram, and M. N. Gurcan, "Convolutional neural network-based clinical predictors of oral dysplasia: class activation map analysis of deep learning results," *Cancers*, vol. 13, no. 6, p. 1291, 2021.