Journal of Pioneering Medical Sciences
Received: July 01, 2025 | Accepted: August 24, 2025 | Published: September 05, 2025
Volume 14, Issue S02, Pages 102-107

DOI https://doi.org/10.47310/jpms202514S0216

Performance Evaluation of Large Language Models in Detecting Buccal Mucosal Lesions Using Smartphone-Based Imaging

Madan Kumar[1*], S. Rajeshwari[2], S. Savitha[3], C. Lavanya[4], K. Ranganathan[5] and Arthi Balasubramaniam[6]

[1,3,6]Department of Public Health Dentistry, Ragas Dental College and Hospital, Chennai, Tamil Nadu – 600 119, India
[2]ICMR Project, Ragas Dental College and Hospital, Chennai, Tamil Nadu – 600 119, India
[4,5]Department of Oral and Maxillofacial pathology, Ragas Dental College and Hospital, Chennai, Tamil Nadu – 600 119, India

Author Designation: [1,4,5]Professor, [2]Research Scientist, [3]Lecturer, [6]Reader

*Corresponding author: Madan Kumar (e-mail: madankumar21@yahoo.co.in).

Abstract **Background:** Early detection of oral potentially malignant disorders (OPMDs) is crucial for improving outcomes in oral cancer, particularly in resource-limited settings. Recent advances in large language models (LLMs) and smartphone imaging provide new opportunities for AI-driven diagnostic support; however, their use in detecting oral mucosal lesions remains underexplored. **Objective:** To evaluate and compare the diagnostic performance of Few-shot prompting, Retrieval-Augmented Generation (RAG), and RAG with Chain-of-Thought (RAG + COT) models inthe binary classification of smartphone-captured intraoral buccal mucosa images as either normal or abnormal. **Methods:** Using a standardized smartphone protocol, 250 intraoral images from 125 patients were categorized as normal, variations, or lesions and split equally into training and testing sets. Few-shot prompting used a subset only 10 test images — 5 normal and 5 lesion — which may produce unstable estimates, while RAG and RAG + COT models trained on the full training set. Expert annotations guided COT descriptors. Variation images were used only in model training for RAG and RAG+COT to improve contextual representation but excluded from binary performance evaluation. Performance was evaluated via sensitivity, specificity, accuracy, F1 score, precision, recall, and 95% confidence intervals. **Results:** Few-shot prompting achieved 80% sensitivity, 100% specificity, 90% accuracy, and an F1 score of 0.88 with wide CIs due to the very small test set. The RAG model, with 54% sensitivity and 91% specificity, showed limited true positive detection. Adding chain-of-thought (RAG + COT) improved sensitivity to 90% and accuracy to 82% (F1: 0.86), though specificity dropped to 64% leading to a higher false-positive rate with potential implications for screening follow-up, however, highlighting the value of structured logical reasoning in enhancing lesion detection. **Conclusion:** The RAG + COT model outperformed Few-shot and RAG models in mucosal lesion detection, demonstrating high sensitivity and improved diagnostic accuracy. However, its low specificity highlights the need for human review before acting on AI results. Findings are promising but preliminary, requiring validation in larger and more balanced datasets before clinical adoption. Its structured logical reasoning highlights the potential of LLMs with COT prompting to strengthen community-based oral cancer screening.

Key Words Large Language Models, Intraoral Imaging, Oral Cancer, Artificial Intelligence, Cancer Early Diagnosis, and Sensitivity and Specificity

INTRODUCTION

Oral cancer is one of the most prevalent cancers worldwide, and its burden is disproportionately higher in low- and middle-income countries, including India [1]. Oral potentially malignant disorders (OPMD) often serve as precursors to oral cancer, underscoring the critical importance of early detection for effective prevention [2]. However, timely diagnosis and treatment remain inaccessible in many settings particularly rural and underserved regions leading to significantly poorer outcomes. The overall five-year survival rate for oral cancer hovers around 50%, but varies widely by geographic and demographic factors; while it may reach up to 65% in developed countries, it can fall as low as 15% in some rural areas, depending on the tumour site [3]. Traditional screening approaches typically require trained personnel and specialized infrastructure, which are often lacking in resource-limited environments.

The widespread adoption of smartphones has opened promising pathways for point-of-care diagnostics. With the ability to capture high-resolution images suitable for clinical evaluation, smartphone cameras offer a practical and scalable solution for community-based screening programs [4]. Concurrently, advances in Artificial Intelligence (AI)particularly through advancements in Deep Learning (DL) and Natural Language Processing (NLP), has led to the emergence of large language models (LLMs), such as Generative Pre-trained Transformers (GPT) have demonstrated superior diagnostic performance over conventional feature-based methods in medical image analysis [5].These models possess the ability to efficiently process large volumes of data, integrating both current research and historical records, thereby establishing a novel paradigm for understanding and evaluating oncological conditions, including head and neck cancers [6]. Their capacity to organize and interpret complex information positions them as promising tools for clinical decision support, with the potential to augment diagnostic accuracy and streamline workflows in healthcare settings [7]. Oral cancer diagnosis using AI has historically relied on convolutional neural networks (CNNs) and transformer-based vision models. While effective, these approaches require extensive labelled datasets and often lack interpretability. LLM-based multimodal systems, such as GPT-4O, integrate visual and textual reasoning and may better handle variable-quality real-world data. Real-world smartphone-based imaging faces technical challenges including poor lighting, motion blur, variable framing, and heterogeneous device capabilities, which can degrade AI performance. One of the latest versions, ChatGPT 4.0, introduces voice and image recognition capabilities, broadening its applicability within healthcare [8]. Given that oral cavity and oropharyngeal lesions typically arise from the mucosal epithelium and are often easily imaged non-invasively, ChatGPT's image analysis capabilities may extend to detecting squamous cell carcinoma and Oral Potentially Malignant Disorders [9]. Such functionality could pave the way for AI-enabled screening tools in oral oncology and contribute to early self-assessment in oral health. Despite growing interest in LLMs within fields such as Periodontology, Endodontics, and Orthodontics, only few studies have specifically explored LLM-based approaches for oral mucosal lesion detection using community-acquired smartphone images [10]. This study explores the diagnostic potential of LLMs, specifically Retrieval-Augmented Generation (RAG) and RAG integrated with Chain-of-Thought (COT) reasoning in identifying OPMD and Oral Cancer from smartphone-captured intraoral images of the buccal mucosa. Buccal mucosa was chosen as the focus because it is easily accessible, well-illuminated, and a frequent site for OPMD in South Asian populations. We also employed few-shot prompting techniques to optimize performance on smaller datasets. This investigation aims to evaluate the strengths and limitations of large language models (LLMs) in the diagnostic assessment of OPMD and oral cancer, offering AI-driven solutions in oral healthcare.

METHODS

Capturing Intraoral Images

This Diagnostic accuracy study was conducted after obtaining approval from the Institutional Review Board (IRB Number: RIEC/20231021/PHD). Written informed consent was obtained from patients after explaining the purpose of the study and assuring them that their data would be protected at all times, ensuring their identity would not be revealed. Intraoral photographs were taken from 125 patients attending various spokes from rural area aged 18 years or older using a Samsung Galaxy M15 5G smartphone based on convenience sampling. The buccal mucosal images were captured following a standardized protocol in which the Region of Interest (ROI) was focused on the centre grid of the camera, covering more than 60% of the area [11]. Captured images were checked for quality, and if they did not meet the required standard, the images were recaptured. The final images were renamed anonymously using an alphanumeric ID and uploaded to the computer via either a direct line connection or a web-based server.

Intraoral Image Dataset

The dataset comprises 250 buccal mucosal images, which are categorized into three groups: normal images, variations from normal, and lesional images. The dataset was split into a training and a testing dataset, where the training dataset comprises 50 normal mucosal images, 50 variations from normal, and 25 lesional images, totalling 125 images. Similarly, the testing dataset contains the same number of images that fall into the same categories. The dataset information is split as shown in Figure 1. The healthy mucosa presents as homogeneous, pink, and shiny, with neither white nor red patches. Variations from normal present as a category that cannot be classified as either lesional or healthy, but exhibit some changes in the tissues. Variation images (n=100 across train/test) were included in RAG and RAG+COT training to enhance contextual knowledge but excluded from binary lesion/normal evaluation. The lesional images present as red or white changes, exhibiting changes similar to those of OPMD and Oral cancer.

The images were annotated using the VGG image annotation tool by three specialists from the Oral Pathology and Public Health Dentistry department, who had received prior training. Disagreements were resolved by consensus with a senior oral pathology expert. Inter-rater reliability was

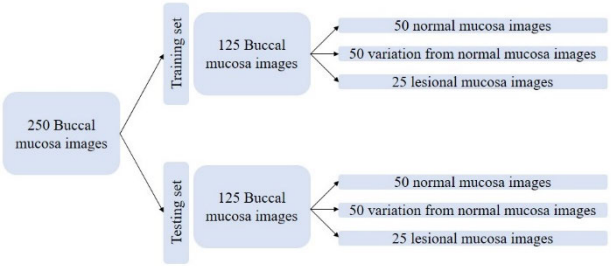

Figure 1: Dataset Distribution

high (Cohen's kappa = 0.87). Each image receives annotations concerning the appearance of symptoms and clinical records or histopathologic reports. Finally, another expert specialized in mucosal diseases with 10 years of clinical experience reviews each case label to confirm the initial assessment. Cases labelled as containing multiple oral disease conditions or as controversial by the experts are excluded from the dataset. The descriptors for the labels of each image were listed separately in an MS Excel sheet under a unique image ID, which is used for Chain of Thoughts in the RAG + COT model. The descriptors, such as location, colour, margin, surface texture, description of the lesion, and size were used as a Chain of Thoughts.

Network Framework and Training

This research employed Few-shot prompting for initial assessment, followed by RAG and RAG + COT. Few-shot prompting is the process of giving a language model a handful of demonstrations or examples within the prompt itself [12]. The goal was to perform a binary classification by categorizing images into two classes: either lesion or normal.

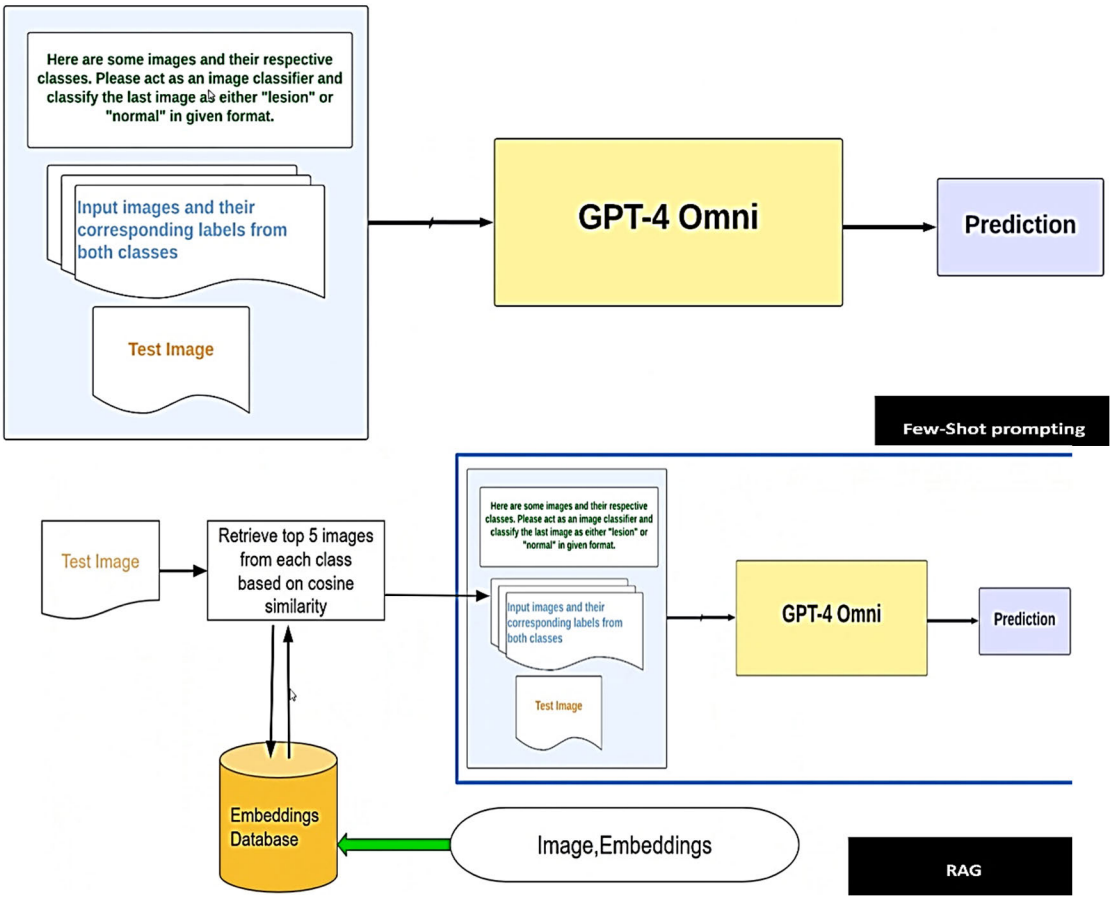

Figure 2: Workflow of various models in classifying intraoral images

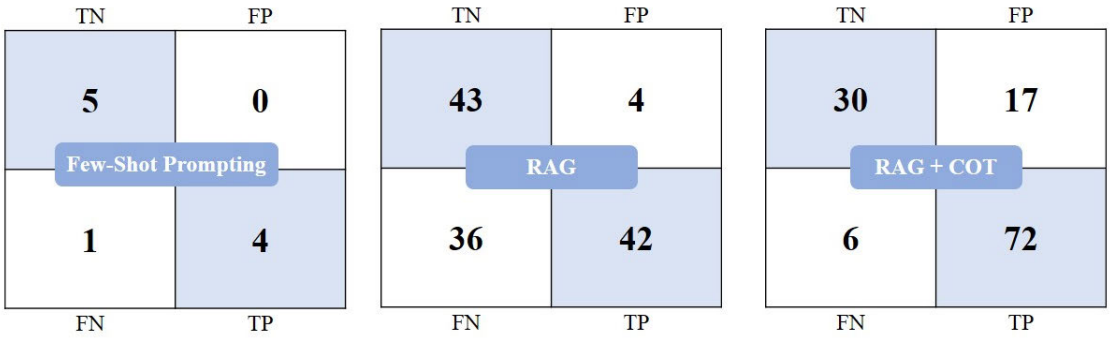

Figure 3: Confusion matrix of the network models

Table 1: Formula used for assessing the diagnostic performance of the models

Performance Metrics	Formula
Sensitivity	TP / (TP + FN)
Specificity	TN / (FP + TN)
Accuracy	(TP + TN) / (TP + TN + FP + FN)
F1 Score	2TP / (2TP + FP + FN)

Retrieval Augmented Generation (RAG) is an AI framework that improves the quality of responses from Large Language Models (LLMs) by augmenting LLMs with a specialized and mutable knowledge base. RAG works by combining external data with pre-trained LLMs to generate more accurate. The advantages include access to updated knowledge, retrieving real-time, current information from external sources, specific query handling, and retrieving relevant documents for accurate, specialized responses. The limitations are that it struggles with rare or niche queries and is limited to static, pre-trained data. RAG technique has been leveraged to retrieve relevant oral images from each class, followed by a few-shot prompting to enhance prediction accuracy. It retrieves the most relevant and up-to-date oral images from a large dataset, ensuring the model references current data for comparison and analysis, thereby providing access to updated and relevant information [13].

Additionally, it retrieves specific oral images related to enhancing the model's ability to provide accurate classifications, even for rare or unusual oral conditions, improving the handling of specific cases. The system comprises two primary components: the Retriever, which identifies and retrieves relevant images from a database based on the input query image for each class, and the Generator, which assigns labels to the retrieved images. This retrieval mechanism ensures the model has access to visually and contextually similar examples, enhancing interpretability and contextual grounding. Subsequently, the Generator processes both the retrieved and query images as part of a structured prompt to produce a textual classification output. In this study, GPT-4O serves as the Generator, leveraging its multimodal capabilities to perform the final diagnostic classification. Chain-of-Thought Prompting is a prompt engineering method that enhances the reasoning capabilities of large language models (LLMs) by encouraging them to break down their reasoning into a series of intermediate steps. In addition to providing an answer, Chain of Thought prompting requires the model to explain how it arrived at that final answer, offering more transparency and improving accuracy [14]. The COT template applied to normal mucosa includes, upon inspection of the {location}, an observation of a {colour} appearance with {margin} edges and a {texture} surface texture, which supports the conclusion that this is normal. For Variation in normal, upon inspecting the {location}, we observe a {colour} appearance with {margin} edges and a {texture} surface texture. The {description of the lesion} and {size} further support the conclusion that this is a {var}. {var} can be lesion or normal based on the requirement. In case of lesion mucosal images, upon inspecting the {location}, the model observes {colour} appearance with {margin, edges and a {texture} surface texture. The {description of the lesion} and {size} further supports the conclusion that it is a lesion. The workflow of the models is given in Figure 2.

Testing of the models was done using a testing dataset to obtain the performance metrics of the models. Of the 125 images in the training dataset, five exemplar images per class (normal and lesion) were selected to fit GPT-4O API token constraints and to maximize diversity of lesion presentations in training the Few-shot prompting model, followed by evaluation using the same number of images from the testing dataset. Following this, 125 images from the training dataset (50 normal, 50 variations from normal, and 25 lesion images) were used to train the RAG and RAG+COT models. The models were then evaluated using the same number of images from the testing dataset.

Diagnostic Performance Metrics

The performance of each model was evaluated using standard diagnostic metrics. True positives, true negatives, false positives, and false negatives were recorded in a confusion matrix (Figure 3). Sensitivity is defined as the ability of a model to correctly identify those with the disease (true positives), meaning it measures how well a model avoids false negatives. Specificity, on the other hand, refers to a model's ability to correctly identify those without the disease (true negatives), indicating how well it avoids false positives. Accuracy represents the number of correctly classified data instances over the total number of data instances. F1 score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives [15]. Sensitivity was the primary evaluation metric due to the screening context. Specificity, accuracy, F1 score, precision, and recall were secondary metrics. 95% confidence intervals were calculated for all metrics. McNemar's test was used for paired binary comparisons between models. All these metrics, calculated using the formulas given in Table 1. This study followed STARD 2015 guidelines for diagnostic accuracy reporting, and a completed checklist is provided in the supplementary material.

RESULTS

The study included 125 participants (82 males [65.6%], 43 females [34.4%]) with a mean age of 48.5 ± 12.3 years. This study evaluated the diagnostic performance of three LLM-based models. The few-shot prompting approach demonstrated a sensitivity of 80% and a specificity of 100%, indicating a strong ability to correctly identify true positive cases while completely eliminating false positives. The model achieved an overall accuracy of 90%, with an F1 score of 0.88, signifying a well-balanced and reliable diagnostic performance (Table 2). Results for the few-shot model should be interpreted with extreme caution given the very

small (n=10) test set; confidence intervals are wide. The standalone RAG model achieved a sensitivity of 54%, correctly identifying just over half of true positive cases of OPMD and oral cancer. This limited sensitivity highlights a high false negative rate, a critical concern in oncologic diagnostics, where delayed or missed detection can severely impact patient outcomes. Conversely, the model demonstrated high specificity at 91%, indicating strong performance in correctly classifying negative cases. This conservative behavior suggests a bias toward avoiding over-diagnosis, but at the expense of missing many true positives. The overall accuracy stood at 68%, meaning approximately two-thirds of the predictions were correct. The F1 score of 0.68 further reflects the model's inability to strike an effective balance between sensitivity and precision, underscoring its limitations in a diagnostic setting that prioritizes early and comprehensive detection. This indicates limited true positive detection despite high specificity.

Upon integrating Chain-of-Thought (COT) reasoning, the RAG + COT model demonstrated a marked improvement in diagnostic performance. Sensitivity increased substantially to 90%, indicating a significantly enhanced capacity to detect true positive cases. This improvement likely stems from the COT framework's ability to guide the model through step-by-step reasoning, enabling more accurate interpretation of lesion-related visual cues. However, this gain in sensitivity was accompanied by a decline in specificity to 64%, leading to a higher false-positive rate with potential implications for screening follow-up. It may lead to additional, potentially unnecessary clinical follow-ups. The overall accuracy improved to 82%, and the F1 score increased to 0.86, indicating a strong balance between precision and recall, with an appropriate leaning toward sensitivity. These metrics confirm the model's enhanced reliability and clinical relevance when structured logicalreasoning is incorporated. Precision and recall values were added to the performance table. RAG+COT's precision was 0.78 and recall 0.90; RAG's precision was 0.65 and recall 0.54; Few-shot's precision was 1.00 and recall 0.80.

Overall, the RAG+COT model appears well-suited for high-sensitivity screening tasks in community or tele-dentistry settings, where early detection of OPMD or oral cancer is paramount. While the higher false positive rate may raise the burden of follow-up diagnostics, it remains acceptable in preventive oncology, particularly when supported by expert clinical validation.

Table 2: Performance metrics of the network models

Models	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	F1 Score	Precision	Recall
Few-Shot prompting	0.80 (0.38-0.96)	1.00 (0.57-1.00)	0.90 (0.60-0.98)	0.88	1.00	0.80
RAG	0.54 (0.36-0.74)	0.91 (0.81-0.97)	0.68 (0.43-0.76)	0.68	0.78	0.56
RAG + COT	0.90 (0.75-0.98)	0.64 (0.49-0.77)	0.82 (0.62-0.92)	0.86	0.56	0.92

DISCUSSION

This study evaluated the diagnostic capabilities of Large Language Models (LLMs) in identifying Oral Potentially Malignant Disorders (OPMD) and Oral Cancer using intraoral images captured via smartphones. A comparative analysis between the standard Retrieval-Augmented Generation (RAG) model and its enhanced variant incorporating Chain-of-Thought (COT) reasoning revealed prominent differences in diagnostic efficacy. The RAG + COT model achieved a sensitivity of 90%, markedly exceeding the 54% sensitivity of the standard RAG model. This substantial improvement highlights the model's enhanced capability to accurately identify true positive cases of OPMD and oral cancer an essential advantage in screening settings where early detection is critical for timely intervention and reducing the risk of disease progression. The high false-positive rate (36% for RAG+COT) has important implications — in screening settings, this could lead to unnecessary referrals, patient anxiety, and additional costs. However, in preventive oncology, prioritizing sensitivity over specificity is often acceptable if expert review is available.

High sensitivity is especially vital in oncology, where undetected lesions may advance to more aggressive forms, reducing the likelihood of successful treatment. While the RAG model showed high specificity (91%), its sensitivity was insufficient to be considered a reliable diagnostic aid in a clinical or community screening context. A diagnostic model that fails to detect nearly half of actual disease cases presents a significant clinical risk, especially in environments with limited access to expert assessment [16]. In contrast, the RAG + COT model, despite a moderate specificity of 64%, provides a clinically acceptable balance by substantially reducing false negatives, thereby prioritizing early detection with an acceptable increase in false positives. Follow-up confirmatory evaluations may mitigate such overdiagnosis, but underdiagnosis in screening scenarios carries a far greater clinical risk.

The improved accuracy (82%) and F1 score (0.86) observed in the RAG + COT model further support its superior diagnostic balance. The incorporation of COT reasoning has improved the model's capacity, aligning with growing evidence that structured logical reasoning prompts enhance the decision-making performance of Large Language Models, especially in complex classification scenarios [17]. However, this research is subject to certain limitations. The improved performance of RAG+COT may not be solely due to chain-of-thought reasoning; dataset characteristics, retrieval quality, and bias toward lesion features could also contribute. Dataset imbalance (more normal than lesion images) may have affected performance; oversampling or augmentation could address this in future work. As this is a pilot study data augmentation was not carried out. Future enhancements should focus on optimizing classification thresholds and incorporating multimodal inputs such as habit history and demographic data to improve specificity while preserving sensitivity,

thereby enhancing clinical applicability. Moreover, while the dataset utilized in this study was representative, its limited size poses constraints on the model's broader applicability. Expanding the sample and incorporating a wider variety of intraoral image types would significantly enhance the model's generalizability. The study also indicates that although the standard RAG model adopts a conservative approach by reducing false positives, it is unsuitable as a standalone screening tool, despite its high specificity. The RAG + COT model, by contrast, offers a more clinically valuable performance profile, substantially improving sensitivity while maintaining high accuracy. Its ability to detect most true positive cases, even at the cost of increased false positives, makes it better suited for use as a frontline screening tool in tele-dentistry or community health settings. The integration of logical reasoning through Chain-of-Thought prompts enhances its interpretative capacity, reinforcing its potential role in AI-assisted oral health diagnostics. Incorporating annotated images from multiple clinical centres can further enhance data variability and minimize inherent biases, thereby strengthening the model's reliability and applicability in diverse patient populations.

CONCLUSIONS

In this preliminary evaluation, large language models demonstrated the potential to identify buccal mucosal lesions from smartphone-based images with varying accuracy, sensitivity, and specificity across prompting approaches. While the RAG + COT method achieved the highest sensitivity, the Few-Shot approach demonstrated perfect specificity in a limited sample. These findings suggest that LLMs could complement clinical decision-making in resource-constrained settings; however, the results should be interpreted cautiously given the relatively small and imbalanced test sets, potential sampling bias, and absence of external validation. Future research should involve larger, more diverse datasets, real-world clinical testing, and exploration of integration strategies with clinician workflows to assess the practical utility and safety of such systems.

Acknowledgement

This study is part of an ICMR project (Project ID IIRP-2023-1049) funded by Small Extramural Grants – 2023.

Conflicts of Interest

There are no conflicts of interest to disclose pertaining to the funding, conduct, or publication of this research.

Ethical Statement

Written informed consent was obtained from the study participants. All patient data were anonymized and securely stored. Unexpected findings were referred to specialists for further evaluation.

REFERENCES

[1] Shrestha, A.D., *et al.* "Prevalence and incidence of oral cancer in low- and middle-income countries: A scoping review." *European Journal of Cancer Care*, vol. 29, no. 2, March 2020.

[2] Abati, S., *et al.* "Oral cancer and precancer: A narrative review on the relevance of early diagnosis." *International Journal of Environmental Research and Public Health*, vol. 17, no. 24, December 2020, 9160.

[3] Warnakulasuriya, S. "Global epidemiology of oral and oropharyngeal cancer." *Oral Oncology*, vol. 45, no. 4–5, April 2009, pp. 309–316.

[4] Mantena, S., *et al.* "Improving community health-care screenings with smartphone-based AI technologies." *The Lancet Digital Health*, vol. 3, no. 5, May 2021, e280–e282.

[5] Garg, A., *et al.* "Prospect of large language models and natural language processing for lung cancer diagnosis: A systematic review." *Expert Systems*, vol. 41, no. 11, November 2024.

[6] De Souza, L.L., *et al.* "The role of large language models in advancing head and neck cancer research and care: A narrative review." *Journal of Medical Artificial Intelligence*, vol. 7, September 2024.

[7] Lin, C. and C.F. Kuo. "Roles and potential of large language models in healthcare: A comprehensive review." *Biomedical Journal*, 29 April 2025.

[8] Temsah, R., *et al.* "Healthcare's new horizon with ChatGPT's voice and vision capabilities: A leap beyond text." *Cureus*, vol. 15, no. 10, October 2023.

[9] Schmidl, B., *et al.* "Artificial intelligence for image recognition in diagnosing oral and oropharyngeal cancer and leukoplakia." *Scientific Reports*, vol. 15, no. 1, January 2025.

[10] Umer, F., I. Batool and N. Naved. "Innovation and application of large language models (LLMs) in dentistry – a scoping review." *BDJ Open*, vol. 10, no. 1, December 2024.

[11] Madan Kumar, P.D. *A Training Manual for Taking Intraoral Photographs for Community Health Workers* [manual]. Copyright Office, Government of India, 2025. Copyright Certificate No.: LD-20250166075.

[12] Perez, E., D. Kiela and K. Cho. "True few-shot learning with language models." *Advances in Neural Information Processing Systems*, vol. 34, December 2021, pp. 11054–11070.

[13] Gao, Y., *et al.* "Retrieval-augmented generation for large language models: A survey." *arXiv preprint* arXiv:2312.10997, vol. 2, no. 1, December 2023.

[14] Chen, Q., *et al.* "Towards reasoning era: A survey of long chain-of-thought for reasoning large language models." *arXiv preprint* arXiv:2503.09567, March 2025.

[15] Santini, A., A. Man and S. Voidăzan. "Accuracy of diagnostic tests." *The Journal of Critical Care Medicine*, vol. 7, no. 3, August 2021, pp. 241-248.

[16] Mabey, D., *et al.* "Diagnostics for the developing world." *Nature Reviews Microbiology*, vol. 2, no. 3, March 2004, pp. 231–240.

[17] Huang, Zhijian, *et al.* "Making large language models better planners with reasoning-decision alignment." *Lecture Notes in Computer Science*, September 09, October 04, 2024, Springer Nature Switzerland, Cham, pp. 73-90. http://dx.doi.org/10.1007/978-3-031-72764-1_5.