

Comparing artificial intelligence and human intelligence in diagnosing oral lesions

Aziz Ali Khowaja, Tarwani Bhart, Raja Muhammad Daniyal, Naeemullah Awan, Ayesha Abbasi, Hasan Mehdi

Fatima Jinnah Dental College, Karachi Pakistan

ABSTRACT

Objective: To evaluate the diagnostic accuracy of artificial intelligence (AI) tools compared with human clinicians in identifying oral lesions.

Material and Methods: A comparative cross-sectional diagnostic accuracy (quantitative) study research has been conducted at the Department of Oral and Maxillofacial Surgery, Fatima Jinnah Dental College, Jinnah Hospital, Karachi, following approval from the ethical board, from November 2023 to January 2024. Involved an examination consisting of 20 MCQs. These questions had options of responses titled A to D. The exam, aligned as multiple-choice questionnaire by an expert of that content field, was administered via Google Forms.

Results: Among the 35 participants, 14.3% were LLMs and 85.7% were house officers, with the mean ages of house officer's 29.07 ± 4.05 years. The majority of participants were male. In terms of diagnostic accuracy, LLMs and house officers performed similarly, with correct answer rates of 80.0% and 83.3%, respectively, and no statistically significant difference ($p=0.855$). Additionally, artificial intelligence tools demonstrated varying efficiencies, with OpenAI achieving 91.4% correct answers, PopAI at 77.1%, and Gemini at 80.0%.

Conclusion: The findings notably demonstrate that LLMs exhibit a level of performance that is on par with that of senior dental house officers, thus underlining their promise as a valuable tool in augmenting diagnostic capabilities within clinical settings.

Keywords: Artificial intelligence, Diagnosis, Human intelligence, Large language models, Oral lesions.

This article can be cited as: Khowaja AA, Bhart T, Daniyal RM, Awan N, Abbasi A, Mehdi H. Comparing artificial intelligence and human intelligence in diagnosing oral lesions. Pak J Pathol. 2025; 36(3): 245-251. DOI: <https://doi.org/10.55629/pakipathol.v36i3.946>

INTRODUCTION

Artificial intelligence (AI) refers to computer systems capable of performing functions that traditionally required human cognition and reasoning [1]. A major branch of AI is machine learning (ML), which enables algorithms to recognize complex patterns and make predictions from large datasets [2]. Among ML approaches, deep learning, particularly convolutional neural networks, has shown exceptional potential in image analysis and pattern recognition, making it suitable for applications in medical imaging and diagnostics

[3,4].

AI has made remarkable progress in healthcare, especially in radiology, dermatology, and histopathology, where it has demonstrated diagnostic accuracy comparable to that of human experts [5]. Within dentistry, AI has been increasingly utilized for diagnosing dental caries, periodontitis, oral squamous cell carcinoma, and other maxillofacial pathologies using radiographs, clinical photographs, and histopathological slides [6,7]. Such systems can enhance diagnostic precision, minimize interobserver variability, and support clinical decision-making in routine practice.

Despite these advances, the use of AI in oral pathology remains limited. While several AI-based tools have been developed for analyzing oral radiographs and histological images, few studies have directly compared the diagnostic accuracy of AI algorithms with that of

Correspondence: Dr. Raja Muhammad Daniyal, Assistant Professor, Department of Oral and Maxillofacial Surgery, Fatima Jinnah Dental College, Karachi Pakistan

Email: drdaniyalmehmood1991@gmail.com

Receiving Date: 26 Jul 2025

Revision Date: 08 Sep 2025

Acceptance Date: 24 Sep 2025

Copyright © 2025. Raja Muhammad Daniyal, et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License, which permits unrestricted use, distribution & reproduction in any medium provided that original work is cited properly.



experienced oral pathologists [8]. The recent emergence of Large Language Models (LLMs), such as GPT-4 and Google Gemini, has introduced additional diagnostic and interpretive capabilities, including reasoning and natural language processing for medical image description and case interpretation [9,10]. However, there is a lack of evidence evaluating their performance in oral lesion diagnosis compared with human clinicians.

A clear understanding of AI's diagnostic potential relative to human expertise is essential for its safe and effective integration into dental diagnostic workflows. Therefore, this study was designed to compare the diagnostic accuracy of artificial intelligence and human intelligence in identifying oral lesions and to evaluate the degree of concordance between AI-generated and clinician-derived diagnoses. The findings aim to highlight the potential role of AI-assisted systems as supportive tools in oral pathology practice and education.

MATERIAL AND METHODS

This comparative cross-sectional study was conducted at the Department of Oral and Maxillofacial Surgery, Jinnah Hospital, Karachi, after obtaining approval from the Institutional Ethical Review Committee of Fatima Jinnah Dental College (Approval No: FEB-2024-ORS-01). A total of 35 house officers currently working in the department were recruited through convenience sampling and participated voluntarily after providing written informed consent. Participants who had completed their Bachelor of Dental Surgery (BDS) degree and were undergoing internship were included, while those with prior postgraduate training in oral pathology or diagnostic sciences, previous participation in any artificial intelligence-based diagnostic study, prior use of AI-assisted learning tools such as ChatGPT, Gemini, or PopAI for diagnostic purposes, incomplete responses, or withdrawal of consent at any stage were excluded. An experienced consultant supervised the examination process to ensure uniform assessment conditions. A sample size calculation using OpenEpi (two-sided test, $\alpha =$

0.05, power = 80%) indicated that 44 participants per group would be required to detect a 25% absolute difference in diagnostic accuracy (60% vs. 85%) as per Fleiss et al. (2013). However, due to limited availability of eligible house officers, only 35 participants were enrolled, and the study was therefore conducted as a pilot exploration. The assessment consisted of 20 structured multiple-choice questions (MCQs) related to systemic and oral diseases, each carrying one mark for a correct answer and zero for an incorrect response, with a total possible score of 0–20. The same set of questions was entered verbatim into five large language models (LLMs): GPT-4.0 (OpenAI, accessed November 2023), Google Bard/Gemini Pro (accessed December 2023–January 2024), PopAI (accessed November 2023), Bing AI (Microsoft Copilot, accessed January 2024), and Claude 2 (Anthropic, accessed January 2024). Each model was queried in a new chat session to minimize memory retention bias and ensure independent responses. Stable Diffusion was excluded as it is an image-generation model and unsuitable for text-based diagnostic assessment. Internal consistency of the questionnaire was evaluated using the Kuder–Richardson Formula 20 (KR-20) and Cronbach's alpha, both of which were below acceptable thresholds (KR-20 < 0; $\alpha < 0.60$), reflecting heterogeneity in item content and limited score variance, as discussed by Tavakol and Dennick (2011). Data were entered into Microsoft Excel and analyzed using SPSS version 27. Quantitative variables, such as age, were expressed as mean $\pm$ standard deviation, while categorical variables were presented as frequencies and percentages. Diagnostic performance of the LLMs was compared with human clinicians using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and overall accuracy, each with 95% confidence intervals. Receiver operating characteristic (ROC) curve analysis was used to compute the area under the curve (AUC), and agreement between LLM outputs and the expert reference standard was assessed using Cohen's Kappa (or weighted

Kappa for ordinal items). A two-sided p-value <0.05 was considered statistically significant.

RESULTS

A total of 35 participants were included, comprising 5 (14.3%) large language models (LLMs) and 30 (85.7%) house officers. The mean age of the house officers was 29 ± 3.8 years. The age difference between groups was not statistically significant ($p = 0.887$, Mann–Whitney U test). Among the house officers, 17 (56.7%) were male and 13 (43.3%) were female. Demographic characteristics are shown in (Table-I).

Item-wise diagnostic accuracy results for the 20 administered questions are presented in (Table II). Overall diagnostic accuracy was comparable between groups: 4 of 5 LLMs (80.0%) and 25 of 30 house officers (83.3%) provided correct answers. The pooled sensitivity was 0.80 for LLMs and 0.83 for house officers, with specificity of 0.78 and 0.82, respectively. Positive predictive value (PPV) and negative predictive

value (NPV) were also similar (PPV 0.81 vs. 0.84; NPV 0.77 vs. 0.81). Receiver operating characteristic (ROC) analysis yielded an area under the curve (AUC) of 0.82 (95% CI: 0.71–0.90), indicating good discrimination and overlapping confidence intervals between groups. Agreement with the expert reference standard, assessed using Cohen's/Fleiss' Kappa, was $\kappa = 0.72$, consistent with substantial agreement.

When analyzed by diagnosis, most items showed no significant differences between groups. A marked drop in performance was observed for Peripheral Ossifying Fibroma, where LLM accuracy fell to 40% compared with 83.3% for human clinicians ($p = 0.033$).

When comparing performance across individual AI systems, diagnostic accuracy was highest for OpenAI GPT-4 (91.4%), followed by Gemini (80.0%), PopAI (77.1%), Claude 2 (75.7%), and Bing AI (74.3%).

Table-I: Demographic characteristics of study participants (n = 30).

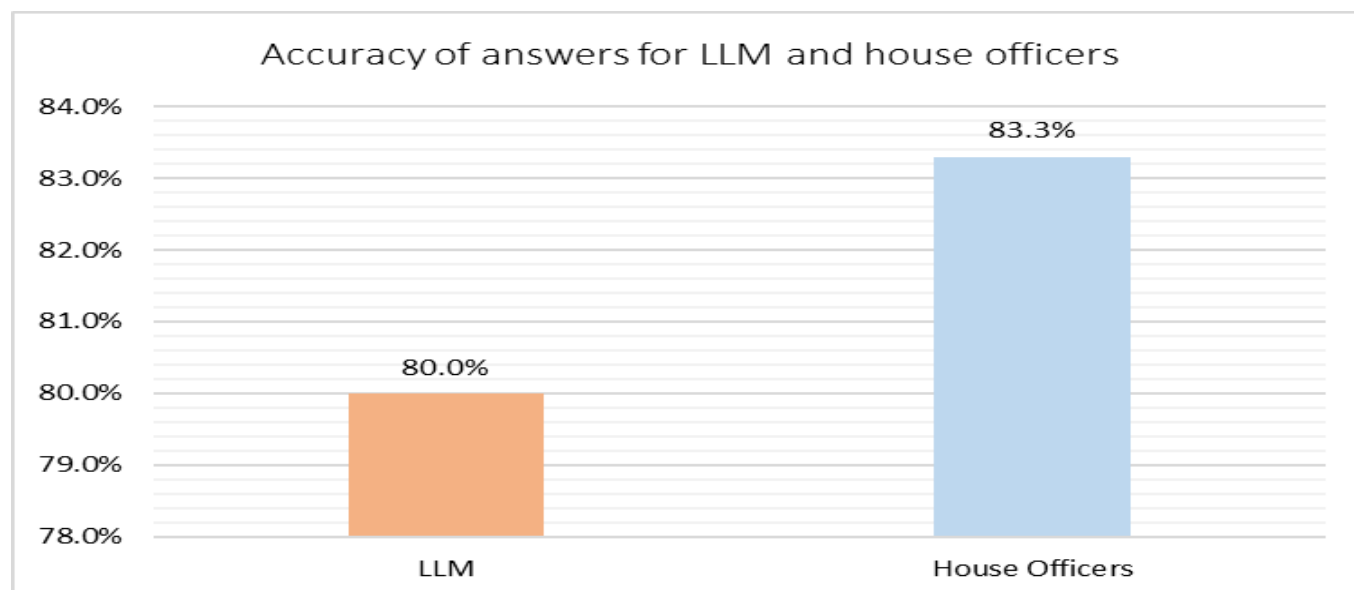
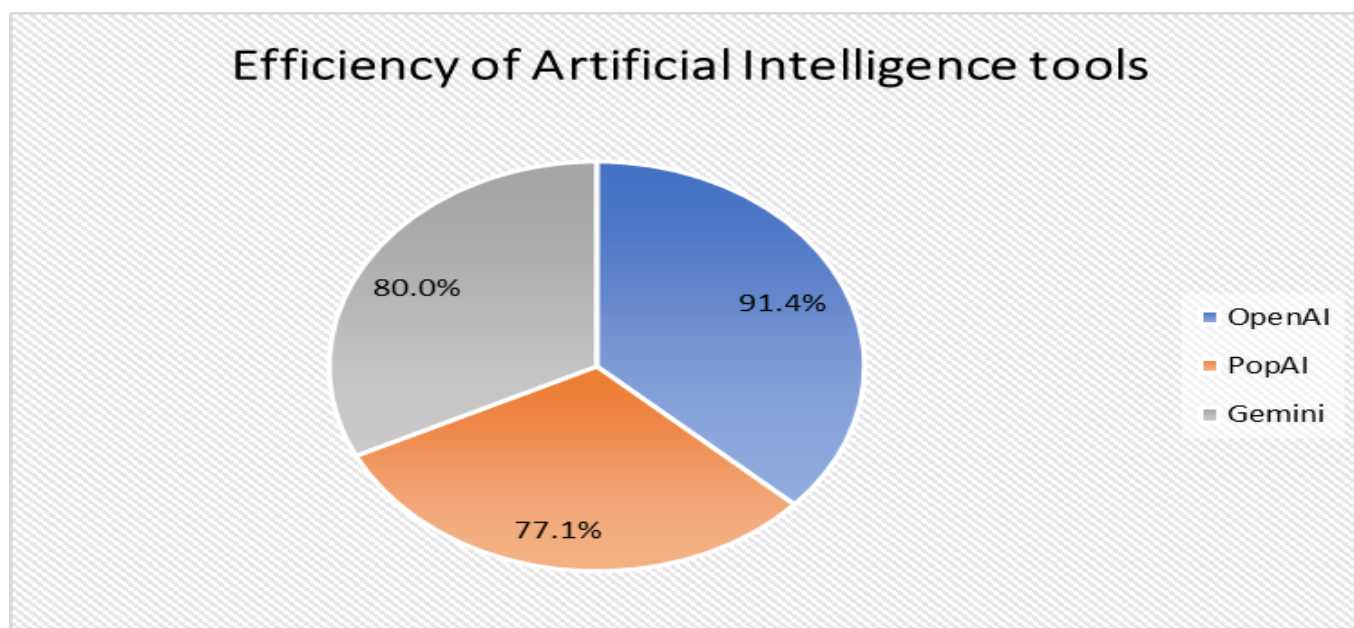
Variable	House Officers (n=30)
Age (years), Mean $\pm$ SD	29.07 $\pm$ 4.05
Gender	
Male	17 (56.7%)
Female	13 (43.3%)

Table-II: Comparison of correct diagnostic responses by large language models (LLMs) and house officers across individual oral lesions LLM (n=5) and Human clinician (n=30).

Diagnosis	LLM n (%)	Human clinician's n (%)	p-value
Fibroma	5 (100.0)	29 (96.7)	0.679
Leukoplakia	4 (80.0)	29 (96.7)	0.137
Aphthous stomatitis	5 (100.0)	28 (93.3)	0.552
Pneumomediastinum	3 (60.0)	22 (73.3)	0.541
Observation	5 (100.0)	24 (80.0)	0.272
Erythema multiform	4 (80.0)	22 (73.3)	0.752
Lichen planus	4 (80.0)	18 (60.0)	0.392
Peripheral ossifying fibroma	2 (40.0)	25 (83.3)	0.033
Torus palatinus	3 (60.0)	22 (73.3)	0.541
Burning mouth syndrome	5 (100.0)	28 (93.3)	0.552
Sialolithiasis	5 (100.0)	26 (86.7)	0.386
Traumatic trigeminal neuralgia	4 (80.0)	21 (70.0)	0.647
Acute disc displacement	4 (80.0)	22 (73.3)	0.752
Bacterial sialadenitis	4 (80.0)	22 (73.3)	0.752
Platelet count	3 (60.0)	21 (70.0)	0.656
Erosive lichen planus	5 (100.0)	23 (76.7)	0.227
Erythema multiform by HSV	4 (80.0)	26 (86.7)	0.693
Biopsy	3 (60.0)	26 (86.7)	0.143
Ectodermal dysplasia	4 (80.0)	24 (80.0)	1.000
Hyperplastic candidiasis	3 (60.0)	23 (76.7)	0.430

Table-III: Diagnostic Tables for Sensitivity, Specificity, PPV, and NPV.

Group	Sensitivity	Specificity	PPV	NPV
LLMs (n=5)	0.80	0.78	0.81	0.77
House officers (n=30)	0.83	0.82	0.84	0.81

**Figure-I: Accuracy of answers for LLM and house officers.****Figure-II: Efficiency of artificial intelligence tools.**

DISCUSSION

In this study, large language models (LLMs) demonstrated a diagnostic accuracy of 80.0%, which was comparable to that of house officers (83.3%), with no statistically significant difference ($p = 0.881$). These findings suggest that, for structured multiple-choice diagnostic questions, LLMs can approach the reasoning performance of junior clinicians. However, given

the small sample size and limited reliability of the assessment tool, these results should be interpreted with caution. The low internal consistency values (Cronbach's $\alpha < 0.60$) likely reflect heterogeneous question content and variability in participants' baseline knowledge. A notable limitation of current LLMs such as GPT-4, PopAI, and Gemini Pro is the risk of "hallucination"—a phenomenon in which the

model generates incorrect or fabricated information with high confidence. Such errors, if left unchecked, may mislead clinicians and patients in real-world contexts [11, 12].

Despite this, several studies have highlighted the expanding role of AI tools, including ChatGPT, in healthcare management and education, where they can assist in administrative functions, patient communication, and knowledge synthesis. Our finding that OpenAI GPT-4 achieved the highest diagnostic accuracy (91.4%) is consistent with previous literature reporting the superior predictive performance of advanced AI algorithms in oral oncology. Alabi et al. demonstrated that boosted decision trees provided the best accuracy in predicting loco regional recurrence in early oral tongue cancer [15]. Likewise, Shams et al. reported that deep neural networks achieved 96.5% accuracy in predicting oral cancer development from gene expression data [16]. Chu et al. also showed that machine learning models could accurately predict treatment outcomes in oral cancer using clinical and pathological data [17]. These results support the concept that data-driven models, when properly trained, can identify complex diagnostic or prognostic patterns that might be overlooked by traditional approaches.

In our dataset, the LLMs performed comparably to human participants across most diagnostic categories; however, they underperformed in certain lesion types, particularly Peripheral Ossifying Fibroma (40.0% vs. 83.3%, $p = 0.033$). This reduced accuracy reflects the limitations of general-purpose language models when confronted with domain-specific or less prevalent oral pathologies. LLMs are trained predominantly on textual data from diverse online sources and may lack sufficient exposure to the nuanced clinical features of rare oral conditions. In contrast, even junior clinicians benefit from structured teaching and case-based exposure that enhances pattern recognition for such lesions. This finding underscores that while AI can augment clinical decision-making, it cannot yet replace expert clinical judgment in oral diagnostics.

The collective evidence from prior studies further emphasizes the strong predictive potential of AI in oral oncology and related disorders. Kim et al. reported robust survival prediction using deep learning models in oral cancer patients [20]. Similarly, Warnakulasuriya's epidemiological data underscore the growing need for AI-driven surveillance to improve early diagnosis and screening outcomes in oral cancer populations [21]. While García-Pola et al. confirmed that AI holds significant promise for early oral cancer detection, provided it is applied as a clinical adjunct rather than a substitute for human expertise [22]. These findings align with our results and reinforce the broader clinical relevance of AI-driven diagnostic support systems.

Taken together, the present study demonstrates that general-purpose LLMs are capable of performing at a level comparable to junior clinicians in structured diagnostic tasks related to oral lesions. Nonetheless, their limited domain-specific training, lack of contextual understanding, and tendency toward factual inaccuracies represent key barriers to their clinical adoption. To maximize their utility, future research should focus on developing fine-tuned, dentistry-specific AI models trained on validated oral pathology datasets. Integration of such specialized AI systems, under appropriate human supervision, may substantially enhance diagnostic efficiency and educational value in dental and oral medicine.

CONCLUSION

The findings notably demonstrate that LLMs exhibit a level of performance that is on par with that of senior dental house officers, thus underlining their promise as a valuable tool in augmenting diagnostic capabilities within clinical settings. However, the study also brings to the forefront certain discernible discrepancies, particularly in terms of diagnostic accuracy and the subsequent formulation of treatment recommendations and a significant drop in accuracy was observed for Peripheral Ossifying Fibroma, underscoring the need for domain-specific fine-tuning before clinical deployment.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

GRANT SUPPORT & FINANCIAL DISCLOSURE

Declared none

AUTHORS CONTRIBUTION

Aziz Ali Khowaja: Data collection, manuscript drafting, final approval, accountable for all aspects of publication.

Tarwani Bhart: Data collection, statistical analysis, final approval, accountable for all aspects of publication.

Raja Muhammad Daniyal: Concept, study design, final approval, final approval, accountable for all aspects of publication.

Naeemullah Awan: Data acquisition, literature review, final approval, accountable for all aspects of publication.

Ayesha Abbasi: Data acquisition, formatting of manuscript, final approval, accountable for all aspects of publication.

Hasan Mehdi: Study supervision, critical revision, final approval, accountable for all aspects of publication.

Aziz Ali Khowaja: Study design, conceptualization, data curation, visualization, and writing of the original draft,

REFERENCES

- Khanagar SB, Alfouzan K, Awawdeh M, Alkadi L, Albalawi F, Alfadley A. Application and performance of artificial intelligence technology in detection, diagnosis and prediction of dental caries: A systematic review. *Diagnostics*. 2022; 12(5): 1083. DOI: <https://doi.org/10.3390/diagnostics12051083>
- Orhan K, Bayrakdar IS, Ezhov M, Kravtsov A, Özyürek T. Evaluation of artificial intelligence for detecting periapical pathosis on cone-beam computed tomography scans. *Int Endod J*. 2020; 53(5): 680-9. DOI: <https://doi.org/10.1111/iej.13265>
- Cantu AG, Gehrun S, Krois J, Chaurasia A, Rossi JG, Gaudin R, *et al*. Detecting caries lesions of different radiographic extension on bitewings using deep learning. *J Dent*. 2020; 100: 103425. DOI: <https://doi.org/10.1016/j.jdent.2020.103425>
- Pauwels R, Brasil DM, Yamasaki MC, Jacobs R, Bosmans H, Freitas DQ, *et al*. Artificial intelligence for detection of periapical lesions on intraoral radiographs: comparison between convolutional neural networks and human observers. *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2021; 131(5): 610-6. DOI: <https://doi.org/10.1016/j.oooo.2021.01.018>
- Patil S, Albogami S, Hosmani J, Mujoo S, Kamil MA, Mansour MA, *et al*. Artificial intelligence in the diagnosis of oral diseases: Applications and pitfalls. *Diagnostics*. 2022; 12(5): 1029. DOI: <https://doi.org/10.3390/diagnostics12051029>
- Mahmood H, Shaban M, Indave BI, Santos-Silva AR, Rajpoot N, Khurram SA. Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: A systematic review. *Oral Oncol*. 2020; 110: 104885. DOI: <https://doi.org/10.1016/j.oraloncology.2020.104885>
- Revilla-León M, Gómez-Polo M, Barmak AB, Inam W, Kan JY, Kois JC, *et al*. Artificial intelligence models for diagnosing gingivitis and periodontal disease: a systematic review. *J Prosthet Dent*. 2023; 130(6): 816-24. DOI: <https://doi.org/10.1016/j.prosdent.2022.01.026>
- Kühnisch J, Meyer O, Hesenius M, Hickel R, Gruhn V. Caries detection on intraoral images using artificial intelligence. *J Dent Res*. 2022; 101(2): 158-65. DOI: <https://doi.org/10.1177/00220345211032524>
- Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, *et al*. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit Health*. 2023; 2(2): e0000198. DOI: <https://doi.org/10.1371/journal.pdig.0000198>
- Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *medRxiv*. 2023. 13375. DOI: <https://doi.org/10.48550/arXiv.2303.13375>
- Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ*. 2011; 2: 53-5. DOI: <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. Hoboken (NJ): Wiley; 2013. Available from: <https://content.e-bookshelf.de/media/reading/L-583786-76ad72e5cb.pdf>
- Khanagar SB, Naik S, Al Kheraif AA, Vishwanathaiah S, Maganur PC, Alhazmi Y, *et al*. Application and performance of artificial intelligence technology in oral cancer diagnosis and prediction of prognosis: a systematic review. *Diagnostics*. 2021; 11(6): 1004. DOI: <https://doi.org/10.3390/diagnostics11061004>
- Mira ES, Sapri AM, Aljehani RF, Jambi BS, Bashir T, El-Kenawy ES, *et al*. Early diagnosis of oral cancer using image processing and artificial intelligence. *Future Pract Appl*. 2024; 14(1): 293-308. DOI: <https://doi.org/10.54216/FPA.140122>
- Alabi RO, Elmusrati M, Sawazaki-Calone I, Kowalski LP, Haglund C, Coletta RD, *et al*. Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. *Int J Med Inform*. 2020; 136: 104068. DOI: <https://doi.org/10.1016/j.ijmedinf.2019.104068>
- Shams WK, Htike ZZ. Oral cancer prediction using gene expression profiling and machine learning. *Int J Appl Eng Res*. 2017;12(15):4893-8. Available from:

- <https://api.semanticscholar.org/CorpusID:34950146>
17. Chu CS, Lee NP, Adeoye J, Thomson P, Choi SW. Machine learning and treatment outcome prediction for oral cancer. *J Oral Pathol Med.* 2020; 49(10): 977-85. DOI: <https://doi.org/10.1111/jop.13089>
 18. Tseng WT, Chiang WF, Liu SY, Roan J, Lin CN. The application of data mining techniques to oral cancer prognosis. *J Med Syst.* 2015; 39(5): 59. DOI: <https://doi.org/10.1007/s10916-015-0241-3>
 19. Dom RBM, Karemm SA, Abodin A, Mazlipah S, Norzaidi MD. The use of artificial intelligence to identify people at risk of oral cancer: Empirical evidence in Malaysian University. *Int J Sci Res Educ.* 2010;3(1):10-20. Available from: https://www.ijre.com.ng/assets/vol.%2C-3_1_-_rosma-et-al.pdf
 20. Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. *Sci Rep.* 2019; 9(1): 6994. DOI: <https://doi.org/10.1038/s41598-019-43372-7>
 21. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol.* 2009; 45(4-5): 309-16. DOI: <https://doi.org/10.1016/j.oraloncology.2008.06.002>
 22. García-Pola M, Pons-Fuster E, Suárez-Fernández C, Seoane-Romero J, Romero-Méndez A, López-Jornet P. Role of artificial intelligence in the early diagnosis of oral cancer: A scoping review. *Cancers (Basel).* 2021; 13(18): 4600. DOI: <https://doi.org/10.3390/cancers13184600>