npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01818-5

# Characteristics, licensing, and ethical considerations of openly accessible oral-maxillofacial imaging datasets: a systematic review



Jing Hao[1], Andrew Nalley[1], Andy Wai Kan Yeung[1], Ray Tanaka[1], Qi Yong H. Ai[2], Walter Yu Hang Lam[3], Zhiyi Shan[4], Yiu Yan Leung[5], Abeer AlHadidi[6], Michael M. Bornstein[7], James Kit Hon Tsoi[1], Colman McGrath[1] & Kuo Feng Hung[1] ✉

Several open-source oral-maxillofacial imaging datasets have been created but their characteristics, ethical clearance, and licensing for reuse remain unclear. This study aimed to systematically identify these datasets and investigate their characteristics, ethical approvals, and licensing requirements for reuse. Open-source oral-maxillofacial imaging datasets were identified through electronic databases and dataset platforms. 105 datasets with 437538 images and 100 intraoral videos from patients across twenty-one countries were included. The datasets comprise imaging modalities, including photographs, periapical, panoramic, and cephalometric radiographs, CBCT, MRI, surface scans, videos, and histopathological images. Nearly 80% of them provide annotations, but only 25.7% specified the annotators' qualification. The majority (83.8%) did not disclose whether ethical approval was obtained, while 61.9% specified terms or licenses for dataset reuse. There is an urgent need to develop standardized guidelines for reusing image datasets and to establish AI-specific consents to fully inform patients about potential uses of their data in AI projects.

Dentistry has progressed swiftly towards digitalization in the last two decades, largely attributed to its significant dependency on advanced imaging techniques with computer-aided design and manufacturing. These technologies play a crucial role in various stages of dental practice, such as diagnosis, treatment planning, guided surgery, post-surgical evaluation, prosthodontic workflows including CAD/CAM applications, and follow-up assessment, and have even expanded to facilitate remote consultations through tele-dentistry[1]. The image data, generated during daily practices and easily accessible from dental clinic database systems, forms the backbone of most artificial intelligence (AI) models proposed in the field of dentistry[2–4]. Currently, many innovative dental AI models have been developed using images to automatically perform complex tasks, such as multimodal image registration[3], segmentation of anatomical structures and pathologies in the oral and maxillofacial region[5,6], detection of various dental diseases[7], generation of 3D dental models[8,9], and interpretation of dental radiographs[10]. Such AI tools have the potential to push the progression of digitalization in oral healthcare.

While certain dental AI models have demonstrated performance on par with or exceeding dental professionals with internal images, most of these models have not been externally validated due to a lack of external image data. This deficiency has been reported in previous studies highlighting a significant performance drop in dental AI models when tested with external images, likely a result of the absence of diverse image data used during model training[11]. The lack of large datasets, comprising images with varying conditions, greatly limits the development and validation of robust and widely applicable dental AI models. One potential solution to enhance the robustness and generalizability of AI models is to integrate images from multiple sources into the training and validation stages[12].

[1]Division of Applied Oral Sciences and Community Dental Care, Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China. [2]Department of Diagnostic Radiology, The University of Hong Kong, Hong Kong SAR, China. [3]Division of Restorative Dental Sciences, Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China. [4]Division of Paediatric Dentistry and Orthodontics, Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China. [5]Division of Oral and Maxillofacial Surgery, Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China. [6]Department of Oral and Maxillofacial Pathology, Radiology and Medicine, New York University, New York, NY, USA. [7]Department of Oral Health & Medicine, University Center for Dental Medicine Basel UZB, University of Basel, Basel, Switzerland. ✉e-mail: hungkfg@hku.hk

In recent years, a growing number of publicly accessible datasets, such as TED3[6], Ctooth[13], IO150K[14], have been introduced. A previous study has identified 16 publicly available dental imaging datasets and summarized their characteristics to facilitate the use of dental imaging data in AI research[15]. More recently, an increasing number of AI studies have been published along with open-access oral-maxillofacial imaging datasets. However, the sources and characteristics of these recent public datasets for oral-maxillofacial imaging including annotation details, have not yet been systematically investigated. Without a thorough understanding of these datasets prior to their use in AI model training and testing, there is an increased risk of unintended biases, such as data leakage. This can occur when training and test sets contain duplicate images from various repackaged datasets, potentially leading to overly optimistic performance estimates as the model could learn from identical data in both phases. Moreover, the significance of understanding ethical considerations, specific terms, and licensing requirements for reusing these datasets is becoming more widely recognized[16–18]. Using these datasets without clear understanding of ethical and licensing information may incur substantial ethical and legal risks. The issue of whether AI models trained on datasets that prohibit commercial use can be licensed for commercial purposes remains controversial. Currently, the ethical clearance and specific terms regulating their reuse in AI projects are unclear. Therefore, the primary objective of this systematic review, reported in accordance with the PRISMA guideline[19], was to provide a comprehensive summary of openly accessible datasets containing images from the oral-maxillofacial region, including details such as the year and purpose of dataset creation, creators, country and institution of origin, imaging modality, image type and format, patient and image count, imaging device manufacturer, image annotation details, annotators' qualifications, and dataset access. The secondary objective was to investigate the ethical approvals, specific terms, and licenses for the reuse of these datasets.

## Results

### Image datasets included in this systematic review

The initial search conducted through PubMed and Google scholar yielded a total of 181 articles. After removing duplicates, 176 datasets remained. Following the screening of titles and abstracts, thirty-six studies were deemed eligible for full-text reading. Among these thirty-six studies, twelve were excluded due to the use of a blocking technique obscuring the oral cavity region ($n = 5$), issues with accessibility ($n = 5$), and unclear descriptions ($n = 2$). Consequently, twenty-four studies[14,20–42] providing information on the eligible datasets were included.

A total of 786 datasets were identified through Google Dataset Search, Kaggle, and Hugging Face. After removing duplicates, 614 datasets remained. Upon initial screening, 86 datasets were deemed eligible. However, seven of these datasets were subsequently excluded due to inaccessibility ($n = 4$), incorrect descriptions ($n = 2$), and degraded image quality ($n = 1$), resulting in 79 datasets included. Additionally, three datasets, which were recommended by experts in the field and met the inclusion criteria, were further included in this systematic review[43–47].

A single duplicate was identified upon cross-checking between the literature and platform searches, resulting in a total of 105 datasets included in this systematic review. Figure 1 illustrates the flowchart of the study and dataset selection process. The two reviewers exhibited high inter reviewer agreement for the selection process with Cohen's kappa values ranging from 0.83 to 0.92.

### General information on the datasets

The 105 datasets were created between 2018 and 2024, comprising a total of 437,538 images and 100 intraoral videos (Table 1; Fig. 2). The number of images per dataset ranged from 17 to 150,000 with 52 (49.5%) datasets

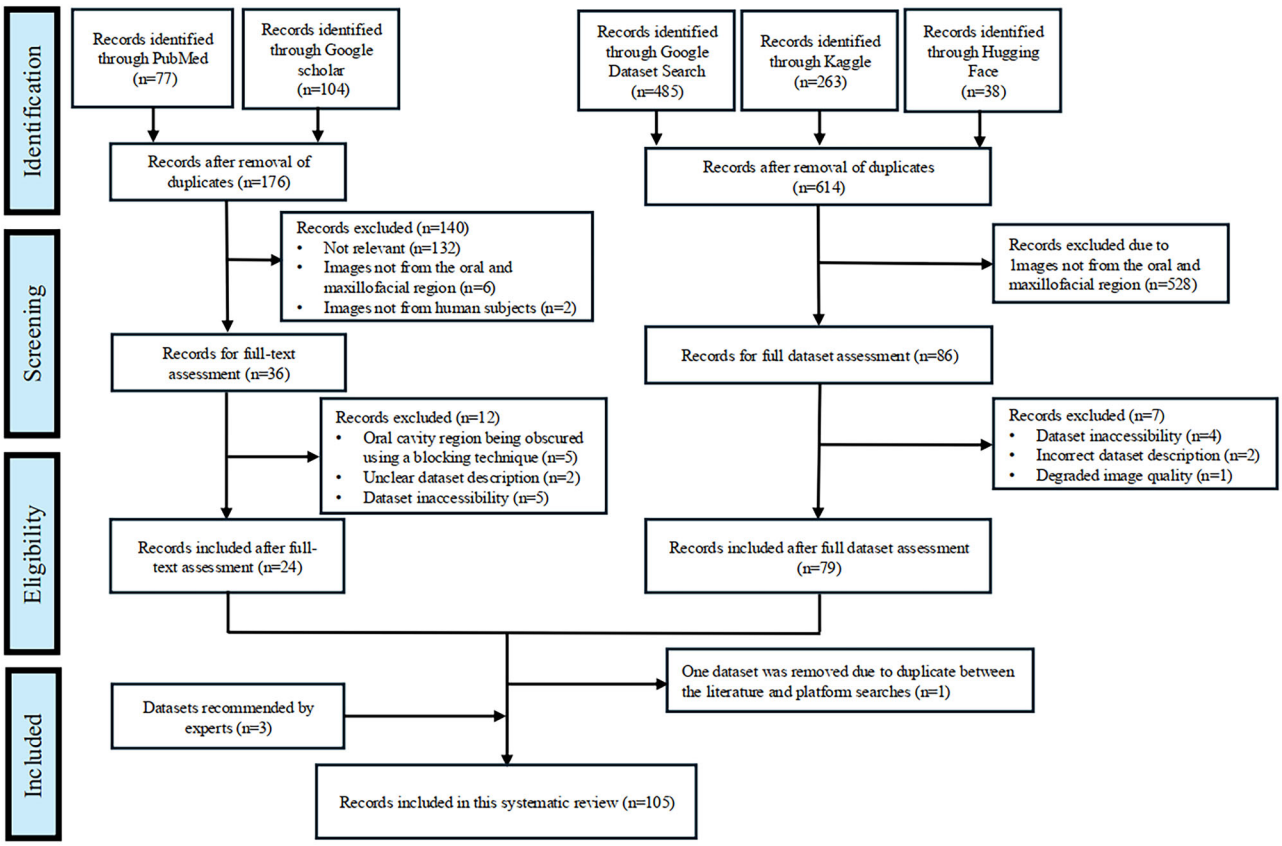


**Fig. 1 | The flowchart of the study and dataset selection process.** The flowchart illustrates the systematic process for selecting datasets relevant to the oral and maxillofacial region. Initially, records were identified from PubMed, Google scholar, Google Dataset Search, Kaggle, and Hugging Face, with duplicates removed. During screening, records were excluded for irrelevance, non-human subjects, or images outside the target region. The eligibility assessment further excluded datasets with obscured regions, unclear descriptions, or inaccessibility. Final inclusion involved expert recommendations and removal of duplicates, resulting in 105 records.

https://doi.org/10.1038/s41746-025-01818-5



**Table 1 | Sources and characteristics of the included openly accessible datasets**

| Dataset no. | Year | Purpose of dataset creation | Dataset creator | Country of origin | Institution of origin | Imaging modality | Image type | Image format | Number of patients | Number of images/scans | Imaging device |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | 2021 | Not reported | Hasnita Dita | Not reported | Not reported | Panoramic radiographs | 2D | PNG | Not reported | 100 | Not reported |
| 2. | 2020 | Mandible segmentation | Amir Hossein Abdi[20,68] | Iran | Noor Medical Imaging centre, Qom | Panoramic radiographs | 2D | PNG | 116 | 116 | Soredex Cranexd Digital Panoramic X-Ray Unit |
| 3. | 2022 | Tooth instance segmentation & maxillomandibular semantic segmentation & report generation | Karen Panetta[21] | USA | Tufts University Institutional Research Board | Panoramic radiographs | 2D | TIFF/JPG | 1000 | 1000 | OP100 Orthopantomograph and Plammeca Promax 2D Radiographic Units |
| 4. | 2022 | Dental implant detection | Unknown | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 2376 | Not reported |
| 5. | 2023 | Tooth instance segmentation | Walid Brahmi[69] | Tunisia | Two Dental Clinics in Sidi Bouzid | Panoramic radiographs | 2D | JPG | Not reported | 107 | Not reported |
| 6. | 2023 | Implant detection | Unknown | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 122 | Not reported |
| 7. | 2023 | Caries segmentation | Unknown | Not reported | Not reported | Panoramic radiographs | 2D | PNG | Not reported | 200 | Not reported |
| 8. | 2023 | Detection of periapical lesions | Association for the Advancement of Artificial Intelligence (AAAI) | Not reported | Not reported | Panoramic radiographs | 2D | PNG | Not reported | 292 | Not reported |
| 9. | 2023 | Not reported | Satishwar Iyer | India | A dental clinic in India | Panoramic radiographs | 2D | JPG | Not reported | 379 | Not reported |
| 10. | 2023 | Not reported | Satishwar Iyer | India | A dental clinic in India | Panoramic radiographs | 2D | PNG | Not reported | 1488 | Not reported |
| 11. | 2023 | Tooth instance segmentation | Nanyang Technological University | Singapore | Nanyang Technological University | Panoramic radiographs | 2D | PNG | Not reported | 565 | Not reported |
| 12. | 2023 | Tooth instance segmentation | Humans in the Loop | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 598 | Not reported |
| 13. | 2023 | Detection of multiple dental conditions | DENTEX | Not reported | Not reported | Panoramic radiographs | 2D | PNG | Not reported | 705 | Not reported |
| 14. | 2023 | Detection of multiple dental conditions | Unknown | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 739 | Not reported |
| 15. | 2023 | Diagnosis of need for apicoectomy surgery | Ilia Farzi[70] | Not reported | Not reported | Panoramic radiographs | 2D | PNG/JPG | Not reported | 857 | Not reported |
| 16. | 2023 | Not reported | Mohamadreza Momeni | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 1269 | Not reported |
| 17. | 2023 | Detection of multiple dental conditions | Unknown | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 1303 | Not reported |
| 18. | 2023 | Tooth semantic segmentation | Unknown | Not reported | Not reported | Panoramic radiographs | 2D | PNG | Not reported | 2000 | Not reported |
| 19. | 2023 | Segmentation of abnormal cells indicating malignancy | Unknown | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 2624 | Not reported |
| 20. | 2023 | Segmentation of teeth and detection of multiple dental conditions | Yifan Zhang[22] | China | Acquired in part at the Diagnostic Imaging Center of the Southwest State University of Bahia | Panoramic radiographs | 2D | PNG | Not reported | 2885 | Not reported |

**Table 1 (continued) | Sources and characteristics of the included openly accessible datasets**

| Dataset no. | Year | Purpose of dataset creation | Dataset creator | Country of origin | Institution of origin | Imaging modality | Image type | Image format | Number of patients | Number of images/scans | Imaging device |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | (UESB) and in part by the Hangzhou Lishui Dental Hospital | | | | | | |
| 21. | 2023 | Detection of multiple dental conditions | Ibrahim Ethem Hamamci et al[23]. | Not reported | Three distinct institutions | Panoramic radiographs | 2D | PNG | Not reported | 3654 | Not reported |
| 22. | 2023 | Detection of multiple dental conditions | Reem Salah Shehab | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 5543 | Not reported |
| 23. | 2023 | Detection of alveolar bone loss | Reem Salah Shehab | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 9521 | Not reported |
| 24. | 2023 | Not reported | Ali Noranian | Not reported | Not reported | Panoramic radiographs | 2D | PNG | Not reported | 112 Original (336 Augmented) | Not reported |
| 25. | 2023 | Detection of multiple dental conditions | MiaMIA Group | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 3834 (Panoramic radiographs) / 15,240 (segmented image patches) | Not reported |
| 26. | 2024 | Caries segmentation | K Vipul Arya | Not reported | Not reported | Panoramic radiographs | 2D | PNG | Not reported | 64 | Not reported |
| 27. | 2024 | Classification of tooth types | Unknown | Not reported | Not reported | Panoramic radiographs | 2D | PNG | Not reported | 221 | Not reported |
| 28. | 2024 | Detection of multiple dental conditions | Rubaba Binte Rahman[65] | Bangladesh | Three centres: Prescription Point Ltd, Lab Aid Specialized Hospital, and Ibn Sina Diagnostic and Imaging Centre | Panoramic radiographs | 2D | JPG | 232 | 232 | Not reported |
| 29. | 2024 | Not reported | R. Zannah | Not reported | Not reported | Panoramic radiographs | 2D | PNG | 389 | 389 | Xiaomi Redmi Note 9 Pro with a 64 MP Camera for Capturing the Raw Images of the Patients |
| 30. | 2024 | Segmentation and classification of tooth types | Devichand Budagam[71] | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 425 | Not reported |
| 31. | 2024 | Detection of multiple dental conditions | Wenbo Zhou[24] | Not reported | China-Japan Union Hospital of Jilin University, Wuxi Stomatology Hospital, People's Hospital of Zhengzhou, and three dental clinics | Panoramic radiographs | 2D | PNG | Not reported | 2000 | Orthophos XG, Planmeca ProMax, and Bondream 1020 |
| 32. | 2024 | Detection of multiple dental conditions | Maria Waqas[72] | Saudi Arabia | King Faisal University, Shaheed Zulfiqar Ali Bhutto Medical University, NED University of Engineering and Technology | Panoramic radiographs | 2D | PNG | Not reported | 622 | Not reported |
| 33. | 2024 | Detection of multiple dental conditions | Unknown | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 772 | Not reported |
| 34. | 2024 | Tooth instance segmentation | Anas Jamal | Not reported | Not reported | Panoramic radiographs | 2D | TIFF | Not reported | 772 | Not reported |

https://doi.org/10.1038/s41746-025-01818-5



**Table 1 (continued) | Sources and characteristics of the included openly accessible datasets**

| Dataset no. | Year | Purpose of dataset creation | Dataset creator | Country of origin | Institution of origin | Imaging modality | Image type | Image format | Number of patients | Number of images/scans | Imaging device |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 35. | 2024 | Gender classification | Wenchi K[25] | China | Sichuan University, Chengdu | Panoramic radiographs | 2D | JPG | Not reported | 979 | Not reported |
| 36. | 2024 | Detection of multiple dental conditions | Abdulrahman Burham | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 3666 | Not reported |
| 37. | 2024 | Detection of multiple dental conditions | Reem Salah Shehab | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 5509 | Not reported |
| 38. | 2024 | Detection of multiple dental conditions | Henrique Rezer Mosquér | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 8188 | Not reported |
| 39. | 2024 | Detection of multiple dental conditions | Reem Salah Shehab | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 8423 | Not reported |
| 40. | 2024 | Detection of multiple dental conditions | Aya Ali Alnozahyy | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 9206 | Not reported |
| 41. | 2024 | Tooth Instance Segmentation | Jing Hao[5] | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 16,317 | Not reported |
| 42. | 2024 | Tooth Semantic Segmentation | Jing Hao[6] | Not reported | Not reported | Panoramic radiographs | 2D | JPG | Not reported | 20,953 | Not reported |
| 43. | 2024 | Tooth Instance Segmentation And Numbering | Niha Adnan[43] | Pakistan | Dental clinics of Aga Khan University Hospital | Panoramic radiographs | 2D | PNG | 250 | 250 | Orthophos XG 3-D (Dentsply Sirona GmbH, Bensheim, Deutschland) |
| 44. | 2021 | Dental radiography image enhancement | Julio César Mello Román[44] | Paraguay | Department of Radiology of the Faculty of Dentistry, Universidad Nacional de Asunción. | Panoramic radiographs | 2D | JPG | 598 | 598 | I-Max touch, Owandy Radiology, France |
| 45. | 2019 | Classification of mouth/ tongue states | Stefan Murga[26] | Canada | Carleton University | Photographs (oral cavity) | 2D | PNG | 17 | 30,231 | Not reported |
| 46. | 2020 | Oral cancer classification | Namrata Sengupta[73] | India | ENT hospitals of Ahmedabad | Photographs (the mouth and oral cavity) | 2D | JPG | Not reported | 131 | Not reported |
| 47. | 2021 | Classification of oral mucosal lesions | Chandrashekar HS[74] | India | Different Hospitals and Colleges in Karnataka | Photographs (oral cavity) | 2D | JPG | Not reported | 323 | Not reported |
| 48. | 2021 | Detection of multiple oral hard and soft tissue conditions | Manodeep Ray | Not reported | Not reported | Photographs (the mouth and oral cavity) | 2D | JPG | Not reported | 3357 | Not reported |
| 49. | 2021 | Detection of mouth and teeth | Vũ Tuấn Hải | Not reported | Not reported | Photographs (the mouth) | 2D | PNG | Not reported | 750 (Dataset 1) and 1787 (Dataset 2) | Not reported |
| 50. | 2022 | Caries segmentation | Atreya Majumdar and Anvit Jaykar | Not reported | Not reported | Photographs (oral cavity) | 2D | TIFF | Not reported | 43 | Not reported |
| 51. | 2022 | Oral cancer classification | N. Sengupta | Not reported | Not reported | Photographs (oral cavity) | 2D | JPG /PNG | Not reported | 131 | Not reported |
| 52. | 2023 | Detection of tooth discoloration | Unknown | Not reported | Not reported | Photographs (oral cavity) | 2D | JPG | Not reported | 199 | Not reported |
| 53. | 2023 | Classification of oral mucosal lesions | Javed Rashid[27] | Pakistan | Dental clinics in Okara, Punjab, Pakistan, and other locations | Photographs (oral cavity) | 2D | JPG | Not reported | 516 | Not reported |
| 54. | 2023 | Classification of tooth types | Unknown | Not reported | Not reported | Photographs (oral cavity) | 2D | JPG | Not reported | 724 | Not reported |

https://doi.org/10.1038/s41746-025-01818-5



**Table 1 (continued) | Sources and characteristics of the included openly accessible datasets**

| Dataset no. | Year | Purpose of dataset creation | Dataset creator | Country of origin | Institution of origin | Imaging modality | Image type | Image format | Number of patients | Number of images/scans | Imaging device |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 55. | 2023 | Tooth segmentation | Pawan Valluri | Not reported | Not reported | Photographs (oral cavity) | 2D | JPG | Not reported | 2495 | Not reported |
| 56. | 2024 | Oral cancer grade classification | Manodeep Ray | Not reported | Not reported | Photographs (oral cavity) | 2D | JPG | Not reported | 204 | Not reported |
| 57. | 2024 | Segmentation of the tongue | Worapan Kusakunniran[28] | Not reported | Not reported | Photographs (the tongue) | 2D | JPG | Not reported | 300 | Not reported |
| 58. | 2024 | Classification of oral cancer cells | Unknown | Not reported | Not reported | Photographs (oral cavity) | 2D | JPG/ PNG | Not reported | 407 | Not reported |
| 59. | 2024 | Oral cancer classification | Mohd Zaid Rashid | Not reported | Not reported | Photographs (Oral Cavity) | 2D | JPG | Not reported | 950 | Not reported |
| 60. | 2024 | Diagnosis of gingivitis | Duy Hoang Bao[75] | Vietnam | Hanoi Medical University | Photographs (oral cavity) | 2D | JPG | Not reported | 1096 | Not reported |
| 61. | 2024 | Oral cancer classification | Manodeep Ray | Not reported | Not reported | Photographs (the mouth and oral cavity) | 2D | JPG /PNG | Not reported | 2006 | Not reported |
| 62. | 2024 | Detection of multiple oral hard and soft tissue conditions | Reem Salah Shehab | Not reported | Not reported | Photographs (oral cavity) | 2D | JPG | Not reported | 3011 | Not reported |
| 63. | 2024 | Oral cancer classification | Manodeep Ray | Not reported | Not reported | Photographs (the mouth and oral cavity) | 2D | JPG/ PNG | Not reported | 3327 | Not reported |
| 64. | 2024 | Classification of calculus, mouth ulcers, tooth discoloration, caries, and missing teeth | Raja Priyanshu | Not reported | Not reported | Photographs (oral cavity) | 2D | JPG | Not reported | 5048 | Not reported |
| 65. | 2024 | Detection and classification of dental caries | Reem Salah Shehab | Not reported | Not reported | Photographs (oral cavity) | 2D | JPG | Not reported | 6845 | Not reported |
| 66. | 2024 | Detection of multiple dental conditions | Salman Sajid | Not reported | Not reported | Photographs (oral cavity) | 2D | JPG | Not reported | 13,862 | Not reported |
| 67. | 2024 | Tooth instance segmentation | Bo Zou[14] | Not reported | Not reported | Photographs (oral cavity) | 2D | PNG | Not reported | 150,000 | Not reported |
| 68. | 2019 | Assessment of lingual papillae patterns | Hanhui | Not reported | Not reported | Photographs (the tongue) | 2D | PNG | Not reported | 1250 | Not reported |
| 69. | 2020 | Not reported | Parth Chokhra | Not reported | Not reported | Periapical radiographs | 2D | JPG | Not reported | 120 | Not reported |
| 70. | 2023 | Caries segmentation | Evident | Not reported | Not reported | Periapical radiographs | 2D | PNG | Not reported | 504 | Not reported |
| 71. | 2023 | Not reported | Nisrean Thalji[76] | Jordan | Jadara University | Periapical radiographs | 2D | PNG | Not reported | 929 | Not reported |
| 72. | 2023 | Not reported | Muhammad Sajad | Not reported | Not reported | Periapical radiographs | 2D | JPG | Not reported | 8620 | Not reported |
| 73. | 2024 | Classification of endodontic and periodontal diseases | Muhammad Sajad | Pakistan | Armed Forces Institute of Dentistry Rawalpindi | Periapical radiographs | 2D | JPG | Not reported | 534 | Not reported |
| 74. | 2024 | Not reported | Abdulrahman Burham | Not reported | Not reported | Periapical radiographs | 2D | JPG | Not reported | 926 | Not reported |
| 75. | 2024 | Caries classification | Walid | Not reported | Not reported | Periapical radiographs | 2D | JPG | Not reported | 957 | Not reported |
| 76. | 2024 | Periapical instance segmentation | Nguyen Thai Tung | Not reported | Not reported | Periapical radiographs | 2D | JPG | Not reported | 1321 | Not reported |

https://doi.org/10.1038/s41746-025-01818-5



**Table 1 (continued) | Sources and characteristics of the included openly accessible datasets**

| Dataset no. | Year | Purpose of dataset creation | Dataset creator | Country of origin | Institution of origin | Imaging modality | Image type | Image format | Number of patients | Number of images/scans | Imaging device |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 77. | 2024 | Not reported | Nada Aglan | Not reported | Not reported | Periapical radiographs | 2D | JPG | Not reported | 1885 | Not reported |
| 78. | 2024 | Unclear detection task | Nada Aglan | Not reported | Not reported | Periapical radiographs | 2D | JPG | Not reported | 1899 | Not reported |
| 79. | 2024 | Apical lesions detection | Viet Do[77] | Vietnam | Hanoi Medical University | Periapical radiographs | 2D | JPG | Not reported | 3926 | Not reported |
| 80. | 2024 | Detection of multiple dental conditions | Engineering Ubu | Not reported | Not reported | Periapical radiographs | 2D | JPG/PNG | Not reported | 6578 | Not reported |
| 81. | 2018 | For teaching dental radiology to dental students | Che-Wei Liao[29] | Not reported | Not reported | micro-CT | 2D slices | TIFF | Not reported | 1031 | The Self-Assembled micro-CT System and an Advanced Commercially Available micro-CT System (Skyscan 2211) |
| 82. | 2022 | Generation of volumetric meshes of jawbone and teeth | Torkan Gholamalizadeh[30] | France | Dentofacial Orthopedics Department, University of Bordeaux, Bordeaux | Jaw and tooth model images generated from CBCT | 3D | OBJ/STL | 17 | 17 | Not reported |
| 83. | 2022 | Mandibular canal segmentation | Marco Cipriano[31] | Italy | The Affidea Center Located in Modena | CBCT | 3D | NUMPY | Not reported | 347 | Newtom/Ntvgimk4 |
| 84. | 2024 | Tooth segmentation | Yaqi Wang[32] | China | Hangzhou Dental Hospital and Hangzhou Qiantang Dental Hospital | Panoramic radiographs and CBCT | 2D&3D | PNG / NUMPY | Not reported | 4000 (Panoramic radiographs) & 362 (CBCT) | OP300 Manufactured with the Instrumentarium Orthopantomograph |
| 85. | 2021 | Landmark localization | Robin Andlauer[33] | Not reported | Not reported | Face images | 2D | JPG | Not reported | 7690 | Not reported |
| 86. | 2022 | Not reported | Törpe | Not reported | Not reported | Intraoral scans | 3D | STL | Not reported | 108 | Not reported |
| 87. | 2022 | Tooth detection | Bitcamp | Not reported | Not reported | Intraoral scan images | 2D | JPG | Not reported | 10,952 | Not reported |
| 88. | 2022 | Localization, segmentation, and labelling of teeth from intraoral 3D scans | Achraf Ben-Hamadou[34] (3DTeethSeg22_challenge ToothFairy or Teeth3DS) | France And Belgium | Acquired By Orthodontists/Dental Surgeons with More than 5 years of professional experience from partner dental clinics located mainly in France and Belgium | Intraoral scans | 3D | DICOM | 900 | 1800 | Primescan, Trios3, iTero Element 2 Plus |
| 89. | 2023 | Automated design of complete denture metal base, tooth point cloud segmentation | Li Cheng[78] | China | Dental laboratories in Chinese hospitals | Scans of dental cast models from dentulous patients | 3D | Point cloud | 950 | 950 | Not reported |
| 90. | 2024 | Classification of the severity of craniosynostosis | Tareq Abdel-Alim[35] | Germany | Department of Oral and Maxillofacial Surgery of The Heidelberg University Hospital | Head and face scans | 3D | PLY | Not reported | 300 | A 3D Image Recording System (Canfield Vectra-360-Nine-Pod, Canfield Science, Fairfield, NJ, USA) |
| 91. | 2021 | Oral cancer cell classification | Chandran Venkatesan | Not reported | Not reported | Histopathological images | 2D | JPG | Not reported | 528 | Not reported |
| 92. | 2022 | Oral exfoliative cytology, location: tongue | Sakaecho-Nishi | Japan | Department of Pathology, Nihon University School of Dentistry | Histopathological images | 2D | PNG | Not reported | 9593 patches | Olympus Dp74 |

https://doi.org/10.1038/s41746-025-01818-5



**Table 1 (continued) | Sources and characteristics of the included openly accessible datasets**

| Dataset no. | Year | Purpose of dataset creation | Dataset creator | Country of origin | Institution of origin | Imaging modality | Image type | Image format | Number of patients | Number of images/scans | Imaging device |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 93. | 2022 | Classification of various nucleus | André Victória Matias | Brazil | Universidade Federal De Santa Catarina | Histopathological images | 2D | JPG | Not reported | 9797 patches | Not reported |
| 94. | 2023 | Classification of oral squamous cell carcinoma and leukoplakia | Maria Clara Falcão Ribeiro De Assis[36] | Brazil | The Oral Diagnosis Project (NDB) of the Federal University of Espírito Santo (UFES) | Histopathological images | 2D | JPG | 77 | 237 | Not reported |
| 95. | 2023 | Classification of normal epithelium of the oral cavity and oral squamous cell carcinoma (OSCC) | Tabassum Yesmin Rahman[37] | India | Dr. B. Borooah Cancer Institute and Ayursundra Healthcare Pvt. Ltd, Guwahati | Histopathological images | 2D | JPG | 230 | 1224 | Leica ICC50 HD Microscope |
| 96. | 2024 | Segmentation and classification of oral epithelial dysplasia | Adriano Barbosa Silva[38] | Brazil | The Federal University of Uberlândia | Histopathological images | 2D | TIFF | 30 | 456 | Histological Slides were Digitized with Leica DM500 Optical Microscope |
| 97. | 2024 | Oral cancer cell classification | Jane Hsieh[39] | Not reported | Not reported | Histopathological images | 2D | TIFF | Not reported | 800 | Not reported |
| 98. | 2024 | Segmentation and classification of papanicolaou-stained cells | Maikel M. Rönnau[40] | Brazil | Faculty of Dentistry, Federal University of Rio Grande do Sul, R. Ramiro Barcelos | Papanicolaou-stained images of the oral mucosa cells | 2D | JPG | 52 | 1563 | Nikon Eclipse Si Microscope with A Nikon Prime Cam 6 Camera |
| 99. | 2024 | Classification of normal epithelium of the oral cavity and oral squamous cell carcinoma (OSCC) | Vidit Gandhi | Not reported | Not reported | Histopathological images | 2D | JPG | Not reported | 5192 | Not reported |
| 100 | 2023 | Segmentation of speech MR images | Matthieu Ruthven[41] | London, UK | School of Biomedical Engineering & Imaging Sciences, King's College London, King's Health Partners, St Thomas' Hospital | MRI | 3D | DICOM | 5 | 392 | 3.0 T TX Achieva MRI Scanner and A 16-Channel Neurovascular Coil |
| 101 | 2024 | Cephalometric landmark detection | Minmin Zeng[42] | China | Fourth Clinical Division, School and Hospital of Stomatology, Peking University | Cephalometric radiographs | 2D | BMP | Not reported | 102 | Planmeca Promax 3D Machine |
| 102 | 2023 | Cephalometric landmark detection | Ching-Wei Wang | Taiwan | 3 Medical Centres | Cephalometric radiographs | 2D | JPG | 600 | 600 | Not reported |
| 103 | 2024 | Not reported | Abaidia Firas | Not reported | Not reported | Dental model images | 2D | JPG | Not reported | 130 | Not reported |
| 104 | 2024 | Cephalometric analysis | Carlos Andres Ferro Sanchez | Colombia | Universidad Autonoma De Occidente | Mid-sagittal CBCT views | 2D | JPG | Not reported | 200 | Not reported |
| 105 | 2024 | Video enhancement, video segmentation, motion estimation, video stabilization | Węsierski, D. | Poland | Medical University of Gdańsk | Video sequences of intra-oral scenes | 2D | MPEG-4 | 100 | 70,000 Frames from 100 Videos | Raw 10-Bit Format Through a Wide-Angle Lens |

*2D* Two-dimensional, *3D* Three-dimensional, *CBCT* Cone-Beam Computed Tomography, *MRI* magnetic resonance imaging, *JPG* Joint Photographic Experts Group, *PNG* Portable Network Graphics, *TIFF* Tagged Image File Format, *OBJ* Object files, *STL* Stereolithography, *ply* Polygon File Format, *bmp* Bitmap image, *DICOM* Digital Imaging and Communications in Medicine.

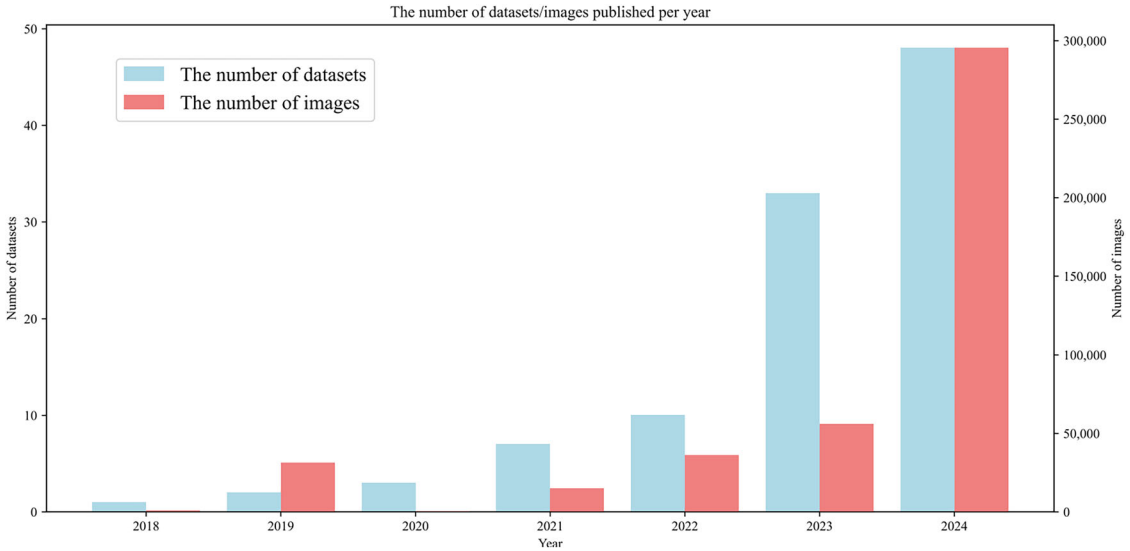


**Fig. 2 | The number of datasets and images released over years.** The bar chart illustrates the annual publication of datasets and images from 2018 to 2024. The left y-axis indicates the number of datasets, while the right y-axis indicates the number of images. Blue and red bars represent datasets and images, respectively. The chart reveals a notable upward trend, with significant increases in both datasets and images, particularly in 2023 and 2024, highlighting the growing interest and expansion in dataset and image publication during this period.

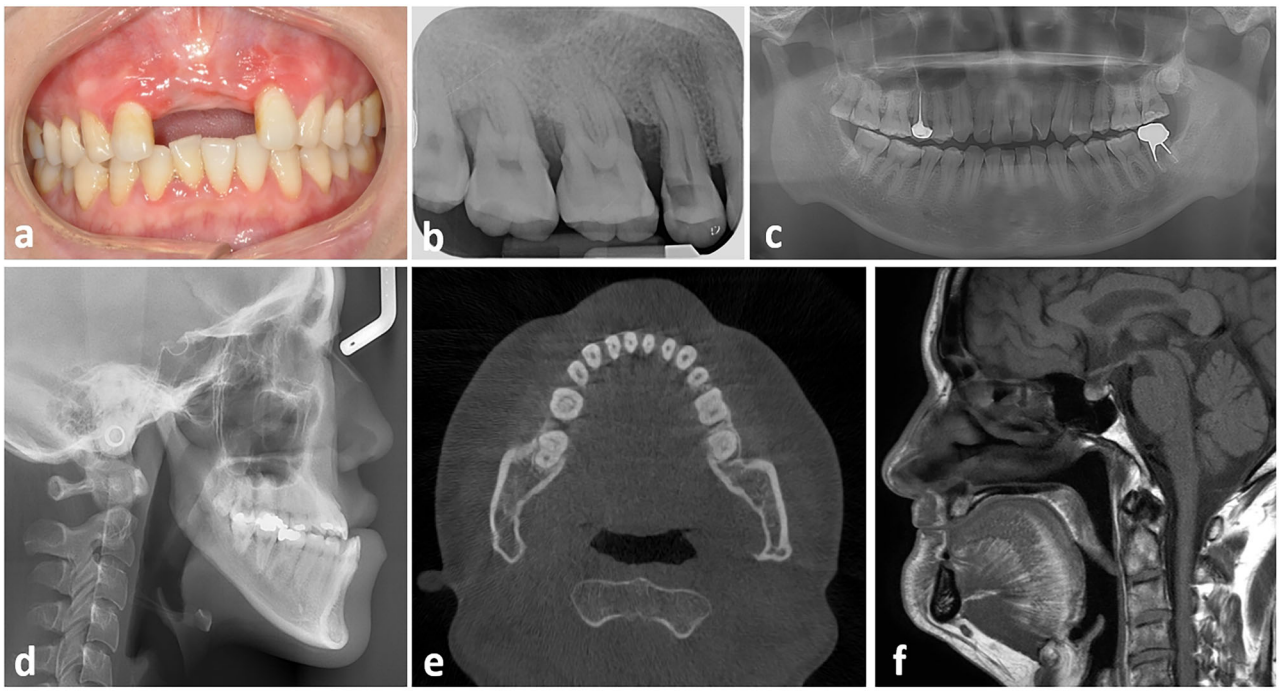


**Fig. 3 | Representative examples of image types included in the datasets.** **a** Intraoral photograph of a patient missing two maxillary central incisors. **b** Periapical radiograph of the maxillary right posterior teeth and surrounding alveolar bone, displaying moderate horizontal bone loss and severe dental caries. **c** Panoramic radiograph providing a comprehensive view of the maxillary and mandibular teeth and jaw structure. **d** Lateral cephalometric radiograph illustrating a lateral perspective of the skull, teeth, and soft tissue profile. **e** Axial cone-beam computed tomography (CBCT) scan presenting detailed cross-sectional imaging. **f** Sagittal magnetic resonance imaging (MRI) scan presenting a sagittal view of craniofacial structures.

containing over 1000 images. Only 13 (12.4%) datasets provided details about the imaging device manufacturer.

Regarding the imaging modality, 45 (43.2%) of the datasets contained panoramic radiographs, 24 (23.1%) photographs, 12 (11.5%) periapical radiographs, 8 (7.7%) histopathological images, 6 (5.8%) intra-oral/facial/model scans or images, 4 (3.9%) CBCT, with the remaining datasets including other modalities such as cephalometric radiographs, MRI, micro-CT, intraoral videos (Fig. 3). Notably, one dataset included both panoramic radiographs and CBCTs.

The image types across all datasets included 228,993 photographs, 125,975 panoramic radiographs, 29,390 histopathological images, 28,199 periapical radiographs, 12,860 intra-oral scans, 7990 head and face scans/images, 1097 model scans/images, 1031 micro-CT images, 709 CBCT scans, 392 MRI, 200 mid-sagittal CBCT, 702 cephalometric radiographs, and 100

**Table 2 | Description of the number of different image modalities in the included datasets and their corresponding purpose of dataset creation**

| Imaging modality | Number of images | Purposes of dataset creation |
|---|---|---|
| Photographs | 228,993 | Diagnosis of gingivitis, detection of multiple oral hard and soft tissue conditions, segmentation of the tongue, classification of oral cancer, segmentation of dental plaque, segmentation of teeth, etc. |
| Panoramic radiographs | 125,975 | Segmentation of teeth, detection of multiple dental conditions, detection of alveolar bone loss, segmentation of caries, segmentation of the mandible, detection of dental implants, etc. |
| Histopathological images | 29,390 | Assessment of lingual papillae patterns, classification of oral cancer cells, differentiation between oral squamous cell carcinoma and leukoplakia, classification of various nucleus, segmentation and classification of oral epithelial dysplasia, etc. |
| Periapical radiographs | 28,199 | Classification of endodontic and periodontal diseases, segmentation of caries, detection and segmentation of periapical lesions, etc. |
| Intraoral, head and face scan/images | 20,850 | Classification of the severity of craniosynostosis, automated design of complete denture metal base, localization of anatomical landmarks, localization and labelling of teeth, etc. |
| CBCT | 709 | Segmentation of teeth, dental radiology education, generation of volumetric meshes, segmentation of the mandibular canal, etc. |
| Others including model scans/images, micro-CT, MRI, mid-sagittal CBCT views, cephalometric radiographs, and intraoral videos | 3522 | Segmentation of speech MR images, detection of cephalometric landmarks, enhancement and segmentation of intraoral videos, etc. |

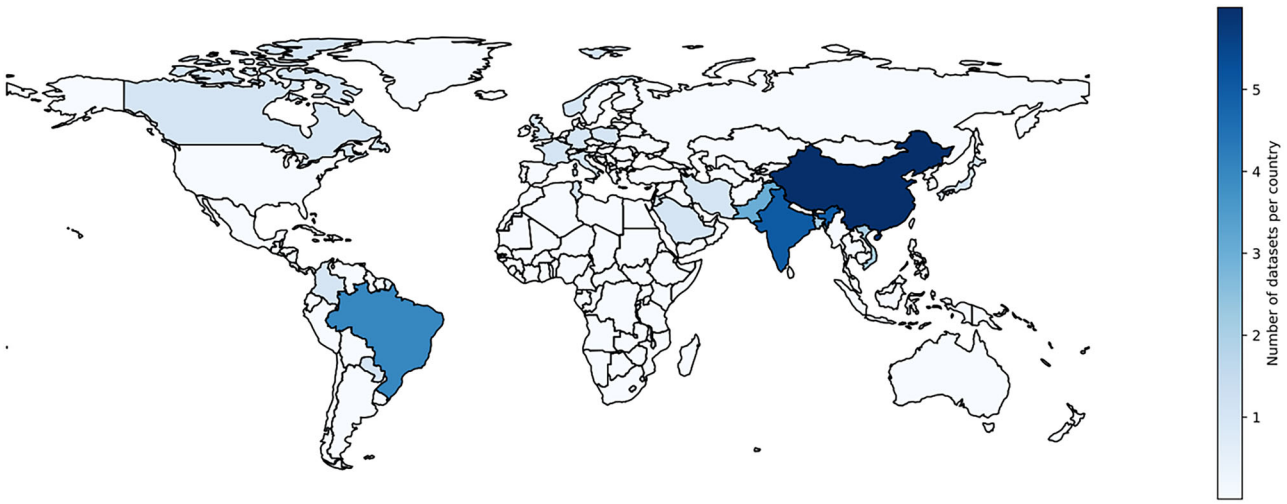


**Fig. 4 | The geographical contribution of the publicly accessible datasets included in this study.** The visual map illustrates the number of datasets sourced from different countries and regions, with darker shades representing higher contributions. Notably, countries such as China and India are highlighted in darker blue, suggesting significant dataset contributions.

intraoral videos (Table 2). All the access links to the datasets are provided in Supplementary Table S1.

### Geographical contribution and institution of origin

Out of the 105 datasets, 66 (62.9%) did not report their origin. Of the remaining, 20 originated from Asia (10 from South Asia, 7 from East Asia, and 3 from Southeast Asia), seven from Europe, six from South America, four from North Africa and the Middle East, and two from North America. The geographical distribution of the datasets with known origin is demonstrated in Fig. 4. Only 38 (36.2%) of the datasets disclosed their institutional origin, with 24 originating from university research centres and 14 from local dental clinics (Table 1).

### Purpose of dataset creation

The datasets included were created mainly for classification, segmentation, detection, and other specific tasks. For classification tasks, the datasets were designed to identify a wide range of oral conditions, including but not limited to oral cancer, oral mucosal lesions, gingivitis, calculus, ulcers, tooth discoloration, caries, missing teeth, as well as endodontic and periodontal diseases. For segmentation tasks, these datasets were used to develop AI models capable of delineating anatomical structures and pathologies, such as caries, teeth, maxilla, mandible, tongue, dental plaque, periapical lesions, mandibular canal, and oral epithelial dysplasia. Detection tasks involved the development of models to identify entities, such as dental implants, periapical lesions, alveolar bone loss, discoloured teeth, and carious lesions. The remaining datasets were created for specific tasks, such as anatomical landmark localization, volumetric mesh generation, cephalometric analysis, report generation, motion estimation, video stabilization, and the automated design of a complete denture metal base.

### Annotations and annotators

Out of the 105 datasets, 83 (79.0%) included annotations, such as the delineation of teeth, caries, and dental restorations on periapical and panoramic radiographs, the delineation of teeth, tongue, mucosal lesions on

**Table 3 | Characteristics of the annotations and the qualifications of the annotators**

| Dataset No. | Imaging modality | Image annotation details | Qualification of the annotators |
|---|---|---|---|
| Dataset No. | Panoramic radiographs | Not reported | Not reported |
| 2. | Panoramic radiographs | Two classes: mandible and background | University dentists |
| 3. | Panoramic radiographs | Five annotations: a) labelled masks, b) eye tracker generated maps (grey and quantized), c) text information describing each radiograph, d) teeth mask for each radiograph with labels, and e) maxillomandibular region-of-interest mask | An expert with unknown qualification and a fourth-year dental student |
| 4. | Panoramic radiographs | Four implant classes: BEGO, Bicon, Straumann, and others | Not reported |
| 5. | Panoramic radiographs | Thirty-two classes: one to thirty-two tooth numbering following the FDI tooth numbering system | Two dentists |
| 6. | Panoramic radiographs | Two classes: implant or not | Not reported |
| 7. | Panoramic radiographs | Two classes: caries or not | Not reported |
| 8. | Panoramic radiographs | Two classes: normal and abnormal (the presence or absence of periapical radiolucency/widening of periodontal ligament space) | Not reported |
| 9. | Panoramic radiographs | Not reported | Not reported |
| 10. | Panoramic radiographs | Not reported | Not reported |
| 11. | Panoramic radiographs | Thirty-five classes: crown, implant, root canal, and one to thirty-two tooth numbering following the FDI tooth numbering system | Not reported |
| 12. | Panoramic radiographs | Thirty-two classes: one to thirty-five tooth numbering following the FDI tooth numbering system | Not reported |
| 13. | Panoramic radiographs | Four classes: mild and severe caries, impacted tooth, and periapical lesions | Not reported |
| 14. | Panoramic radiographs | Multiple dental conditions | Not reported |
| 15. | Panoramic radiographs | Two classes: root canal treatment in the past (class 0) or is considered in need of surgery by endodontist (class 1) | Endodontist |
| 16. | Panoramic radiographs | Not reported | Not reported |
| 17. | Panoramic radiographs | Three classes: normal, tooth fillings/restoration, and caries | Not reported |
| 18. | Panoramic radiographs | Two classes: tooth and background | Not reported |
| 19. | Panoramic radiographs | Nine classes: hyper chromaticism, loss of cohesion, mitotic figures, multiple nucleoli, pleomorphic cells, altered nuclear cytoplasmic ratio, individual cell keratinization, keratin pearl, nucleo-cytoplasmic ratio | Not reported |
| 20. | Panoramic radiographs | Six classes: caries, periapical infection, pulpitis, deep sulcus, dental developmental abnormalities, and others. | Three dentists |
| 21. | Panoramic radiographs | (a) 693 x-rays labelled for quadrant detection and quadrant classes only, (b) 634 x-rays labelled for tooth detection with quadrant and tooth enumeration classes, (c) 1005 x-rays fully labelled for abnormal tooth. The diagnosis class includes four specific categories: caries, deep caries, periapical lesions, and impacted teeth. An additional 1571 unlabelled x-rays are provided for pre-training. Detection with quadrant, tooth enumeration, and diagnosis classes. | Annotations were done by a final-year dental student and verified by three dentists with over 15 years of experience |
| 22. | Panoramic radiographs | Ten classes: 'amalgam filling', 'caries', 'composite filling', 'crown', 'filling', 'implant', 'periapical lesion', 'retained root', 'root canal filling', 'root canal obturation' | Not reported |
| 23. | Panoramic radiographs | The presence of bone loss | Not reported |
| 24. | Panoramic radiographs | Not reported | Not reported |
| 25. | Panoramic radiographs | Five classes: single dental implant, two adjacent implants, crown/bridge restorations, implant(s) with crown/bridge restorations, and metallic reference spheres | Not reported |
| 26. | Panoramic radiographs | Not reported | Not reported |
| 27. | Panoramic radiographs | Five classes: molar, premolar, canine, lateral incisor, and central incisor | Not reported |
| 28. | Panoramic radiographs | Six classes: healthy tooth, caries, impacted tooth, broken crown/root, infection, fractured tooth | Not reported |
| 29. | Panoramic radiographs | Not reported | Not reported |
| 30. | Panoramic radiographs | Four classes including incisors, canines, premolars, molars | Not reported |
| 31. | Panoramic radiographs | Seven classes: 'tooth without anomalies', 'tooth with fillings', 'tooth with RCT', 'tooth with crown', 'tooth with caries', 'residual root', and 'tooth with RCT and crown' | Two postgraduate students and two senior clinicians |

**Table 3 (continued) | Characteristics of the annotations and the qualifications of the annotators**

| Dataset No. | Imaging modality | Image annotation details | Qualification of the annotators |
|---|---|---|---|
| 32. | Panoramic radiographs | Three dental issues namely, broken root, periodontally compromised tooth, and kennedy classification of partially edentulous arches | Not reported |
| 33. | Panoramic radiographs | Dental fillings, restorations, implants, and impacted teeth | Not reported |
| 34. | Panoramic radiographs | Not reported | Not reported |
| 35. | Panoramic radiographs | Two classes: male and female | Not reported |
| 36. | Panoramic radiographs | Five classes: cavity, implant, fillings, impacted tooth, and normal teeth | Not reported |
| 37. | Panoramic radiographs | Thirty-eight classes: 'amalgam filling', 'bone loss', 'caries', 'composite filling', 'crown', 'cyst', 'filling', 'fracture teeth', 'implant', 'misaligned', 'missing teeth', 'periapical lesion', 'permanent teeth', 'primary teeth', 'retained root', 'root piece', 'root canal filling', 'root canal obturation', 'root resorption', 'supra eruption', 'tad', 'unhealed socket', 'abutment', 'attrition', 'bone defect', 'cavity', 'decay', 'gingival former', 'impacted tooth', 'metal band', 'orthodontic brackets', 'permanent retainer', 'plating', 'post - core', 'rct', 'retained deciduous tooth', 'spacing', 'wire' | Not reported |
| 38. | Panoramic radiographs | Fifteen classes: none, caries, crown, filling, implant, misaligned tooth, mandibular canal, missing tooth, periapical lesion, retained root, root canal treatment, impacted tooth, maxillary sinus, 'root piece", and one unclear class. | Not reported |
| 39. | Panoramic radiographs | Fourteen classes: 'caries', 'crown', 'filling', 'implant', 'misaligned', 'mandibular canal', 'missing teeth', 'periapical lesion', 'retained root', 'root canal treatment', 'root piece', 'croen', 'impacted tooth', 'maxillary sinus' | Not reported |
| 40. | Panoramic radiographs | Twelve classes: 'caries', 'crown', 'filling', 'implant', 'misaligned', 'mandibular canal', 'missing teeth', 'periapical lesion', 'retained root', 'root canal treatment', 'root piece', 'impacted tooth' | Not reported |
| 41. | Panoramic radiographs | Thirty-two classes: one to thirty-two tooth numbering following the FDI tooth numbering system | Not reported |
| 42. | Panoramic radiographs | Two classes: tooth and background | Not reported |
| 43. | Panoramic radiographs | Thirty-two classes: one to thirty-two tooth numbering following the FDI tooth numbering system | Two dentists |
| 44. | Panoramic radiographs | Not reported | Not reported |
| 45. | Photographs (oral cavity) | Seven classes: mouth open, mouth closed, tongue up, tongue down, tongue middle, tongue left, tongue right | Not reported |
| 46. | Photographs (the mouth and oral cavity) | Two classes: oral cancer and no oral cancer | ENT doctors |
| 47. | Photographs (oral cavity) | Two classes: benign lesions and malignant lesions | Not reported |
| 48. | Photographs (the mouth and oral cavity) | Five classes: dental caries, healthy, oral cancer, periodontal, scurvy | Not reported |
| 49. | Photographs (the mouth) | Annotation of the mouth | Not reported |
| 50. | Photographs (oral cavity) | Two classes: caries or not | Not reported |
| 51. | Photographs (oral cavity) | Two classes: oral cancer and no oral cancer | Not reported |
| 52. | Photographs (oral cavity) | Two classes: discoloration or not | Not reported |
| 53. | Photographs (oral cavity) | The collection includes class labels for seven mouth and oral cavity diseases, including gingivostomatitis (gum), canker sores (cas), cold sores (cos), oral lichen planus (olp), oral thrush (ot), mouth cancer (mc), and oral cancer (oc). | Not reported |
| 54. | Photographs (oral cavity) | Seven classes: 1st molar, 1st premolar, 2nd molar, 2nd premolar, canine, central incisor, lateral incisor | Not reported |
| 55. | Photographs (oral cavity) | Not reported | Not reported |
| 56. | Photographs (oral cavity) | Three classes: grade-1, grade-2, grade-3 | Not reported |
| 57. | Photographs (the tongue) | Binary mask of the tongue region | Dental practitioners |
| 58. | Photographs (oral cavity) | Three classes: cancer, non-cancer, pre-cancer | Not reported |
| 59. | Photographs (oral cavity) | Two classes: cancer or not | Not reported |
| 60. | Photographs (oral cavity) | Diagnosis | Periodontists |
| 61. | Photographs (the mouth and oral cavity) | Two classes: cancer or not | Not reported |
| 62. | Photographs (oral cavity) | Seven classes: 'calculus', 'cavities', 'maligned', 'missing', 'plaque', 'spacing', 'stains' | Not reported |

**Table 3 (continued) | Characteristics of the annotations and the qualifications of the annotators**

| Dataset No. | Imaging modality | Image annotation details | Qualification of the annotators |
|---|---|---|---|
| 63. | Photographs (the mouth and oral cavity) | Two classes: cancer or not | Not reported |
| 64. | Photographs (oral cavity) | Five classes including calculus, mouth ulcer, tooth discoloration, caries, hypodontia | Not reported |
| 65. | Photographs (oral cavity) | Four classes: caries-free, early decay, caries, decay cavity | Not reported |
| 66. | Photographs (oral cavity) | Six classes: caries, calculus, gingivitis, tooth discolouration, ulcers, and hypodontia | Not reported |
| 67. | Photographs (oral cavity) | Tooth masks | Four orthodontists with over 6 years of experience |
| 68. | Photographs (the tongue) | Not reported | Not reported |
| 69. | Periapical radiographs | Not reported | Not reported |
| 70. | Periapical radiographs | Two classes: caries or not | Not reported |
| 71. | Periapical radiographs | Not reported | Not reported |
| 72. | Periapical radiographs | Not reported | Not reported |
| 73. | Periapical radiographs | Five classes: primary endodontic lesions, primary periodontal lesions, primary endodontic lesions with secondary periodontal lesions, primary periodontal lesions with secondary endodontic lesions, and true combined lesions | Not reported |
| 74. | Periapical radiographs | Not reported | Not reported |
| 75. | Periapical radiographs | Two classes: caries or not | Not reported |
| 76. | Periapical radiographs | Not reported | Not reported |
| 77. | Periapical radiographs | Not reported | Not reported |
| 78. | Periapical radiographs | Bounding box | Not reported |
| 79. | Periapical radiographs | Not reported | Not reported |
| 80. | Periapical radiographs | Seven classes: irreversible pulpitis, impacted tooth, improper restoration, chronic apical periodontitis, unerupted tooth, caries, and periodontitis | Not reported |
| 81. | micro-CT | Not reported | Not reported |
| 82. | CBCT | The mandible, maxilla, their associated teeth, and pdl meshes, as well as teeth principal axes | Not reported |
| 83. | CBCT | Two classes: mandibular canal and background | A team of surgeons with years of experience in maxillofacial surgery |
| 84. | Panoramic radiographs and CBCT | Two classes: tooth and background | Hospital dentists |
| 85. | Face images | Not reported | Not reported |
| 86. | Intraoral scans | Not reported | Not reported |
| 87. | Intraoral scan images | Ten classes: abutment, canine, crown, implant, implant minus, implant plus, inlay, incisor, molar, premolar | Not reported |
| 88. | Intraoral scans | 3D instance masks | Clinicians with more than 10 years of expertise in orthodontics, dental surgery, and endodontics. |
| 89. | 3D scans of complete denture boundaries on the edentulous jaws | Four classes: edentulous model, retentive mesh, tissue stop, and finishing line | A dental specialist with at least 5 years of clinical experience |
| 90. | Head and face scans | Four classes: sagittal suture fusion (scaphocephaly), metopic suture fusion (trigonocephaly), coronal suture fusion (brachycephaly and anterior plagiocephaly), and the control (normocephaly and positional plagiocephaly) | Two craniofacial surgeons |
| 91. | Histopathological images | Two classes: cancer or not | Not reported |
| 92. | Histopathological images | Not reported | Not reported |
| 93. | Histopathological images | Seven classes: “background”, “abnormal epithelial nucleus”, “healthy epithelial nucleus”, “out of focus nucleus”, “blood cell nucleus”, “reactive cell nucleus”, and “dividing nucleus” | Five specialists in pathology |
| 94. | Histopathological images | Oral squamous cell carcinoma and leukoplakia | Two or three oral pathologists that reach a histopathological diagnosis in consensus |
| 95. | Histopathological images | Normal epithelium of the oral cavity and oral squamous cell carcinoma (OSCC) | Medical doctors |
| 96. | Histopathological images | Four classes: healthy tissues, mild, moderate and severe OED | A trained specialist and validated by a pathologist. |
| 97. | Histopathological images | Three classes: cancer cells, chemoresistant cancer cells, and normal cells | Not reported |

**Table 3 (continued) | Characteristics of the annotations and the qualifications of the annotators**

| Dataset No. | Imaging modality | Image annotation details | Qualification of the annotators |
|---|---|---|---|
| 98. | Papanicolaou-stained images of the oral mucosa cells | Individual cytoplasm (orange), squamous cell (green), superficial cell nucleus (red), intermediate cell nucleus (cyan), suspicious cell nucleus (yellow), binucleate nuclei (purple), cytoplasms of cell cluster (blue). The remaining pixels were considered background (grey). | Specialists in pathology |
| 99. | Histopathological images | Normal epithelium of the oral cavity and oral squamous cell carcinoma (OSCC) | Not reported |
| 100 | MRI | Consisted of six classes, head, soft palate, jaw, tongue, vocal tract, tooth space | An MRI physicist with four years of speech MRI experience |
| 101 | Cephalometric radiographs | Nineteen landmarks | Two experienced medical doctors |
| 102 | Cephalometric radiographs | Thirty-eight craniofacial landmarks | Not reported |
| 103 | Dental model images | Text description about the occlusal view of a dental model | Not reported |
| 104 | Sagittal projection images from CBCT scans | Cephalometric parameters | Not reported |
| 105 | Video sequences of intra-oral scenes | Multi-task pseudo labels | Not reported |

photographs, the segmentation of teeth on CBCT, and the categorical label of the presence or absence of cancer cells on histopathological images (Table 3). However, only 27 (25.7%) datasets provided information about the qualification of the annotators. The annotators involved dental students, general dentists, and specialists such as endodontists, periodontists, orthodontists, radiologists, pathologists, ENT, craniofacial, and maxillofacial surgeons (Table 3).

### Ethical approval, specific terms, and licenses of the included datasets

The majority of the included datasets ($n = 88$; 83.8%) did not indicate whether they had obtained ethical approval (Table 4). Out of the 105 datasets, 65 (61.9%) specified terms or licenses for their reuse (Table 4). The licenses attached to the datasets included CC BY 4.0 ($n = 34$; 52.3%), Apache 2.0 ($n = 12$; 18.5%), CC0 1.0 ($n = 7$; 10.8%), CC BY-NC 3.0/4.0 ($n = 4$; 6.2%), CC BY-SA 3.0/4.0 ($n = 2$; 3.1%), CC BY-NC-ND ($n = 2$; 3.1%), CC BY-NC-SA 4.0 ($n = 1$; 1.5%), and MIT ($n = 1$; 1.5%). One dataset specified dual licenses (CC0 1.0 and CC-BY), while another only provided the terms of reuse.

### Applicability concerns of the included datasets and the risk of bias in annotations

The evaluation of applicability concerns for the 105 datasets and the assessment of the risk of bias in annotations are presented in Table 4. Out of these datasets, only 12 (11.4%) were rated as having a "low" applicability concern due to their documentation of ethical approval and licensing. Conversely, 36 (34.3%) datasets were deemed to have a 'high' applicability concern due to the absence of reported ethical approval and licensing. Regarding the risk of bias in the ground truth annotations, out of the 83 annotated datasets, 59 (71.1%) were rated as "high" risk due to the lack of information about the annotators. Eighteen (21.7%) datasets were rated as "low" risk, attributed to the involvement of more than one annotator with explicit medical/dental qualifications. Furthermore, six (7.2%) datasets were rated as "moderate" risk, either because they were annotated by a single qualified annotator or by multiple annotators who lacked explicit qualifications.

## Discussion

This study aimed to provide a comprehensive overview of the openly accessible oral-maxillofacial imaging datasets, their sources and characteristics of both the images and annotations. In addition, this study also investigated the ethical clearance, specific terms, and licenses concerning the reuse of these datasets. During full-text evaluation, three datasets[21,31,48] required registered access. Access to the Tufts Dental Database[21] and the dataset by Cipriano M was acquired by providing an email address, institutional affiliation, and the intended use of the data or by creating an account. However, no response was received from the owner of the dataset[48] following multiple attempts to fulfil the access requirements. The datasets by Ramakrishnan et al.[49], Chilamkurthy et al.[50], and Iosifidis et al.[51] could not be accessed as the specified download sites were not available both at the time of the initial search and at the time of manuscript submission. Access to the dataset by Ranjbar et al.[52] can only be acquired by obtaining an affiliate appointment with the institution for collaborative projects. Moreover, three datasets identified on the Kaggle platform were not available and access to a dataset by Jian[53] requires a subscription payment. Eventually, a total of 105 openly accessible datasets were identified from both electronic databases and dataset management platforms. The findings reveal a significant increase in the number of open-source datasets for oral-maxillofacial imaging since 2018.

Two previous review articles identified publicly available ophthalmological imaging datasets and skin cancer image datasets, both derived from searches on MEDLINE, Google, and Google Dataset Search[54,55]. Another study by Ni et al. identified publicly available datasets for health misinformation detection from searches on the Web of Science Core Collection and arXiv[56]. Uribe et al. identified sixteen publicly accessible dental imaging datasets, created from 2020 to 2023, containing intraoral photographs or radiographs, panoramic radiographs, cephalometric radiographs, CBCT, and intraoral 3D scans[15]. However, in contrast to their findings, this study identified a significantly higher number of datasets created between 2018 and 2024. This study identified 105 datasets containing not only dental images but also those from oral-maxillofacial regions, with a wider range of imaging modalities including intraoral and extraoral photographs, periapical radiographs, panoramic radiographs, cephalometric radiographs, histopathological images, CBCT, intraoral/facial/model scans or images, MRI, micro-CT, and intraoral videos. Moreover, this study included over fifty datasets each providing more than 1000 images while Uribe et al. reported only five datasets with over 1000 images.

Among all the datasets, panoramic radiography is the most prevalent imaging modality. The second most common imaging modality is photography, with 24 datasets consisting of images of the lips, oral cavity, teeth, buccal mucosa, and tongue. The largest dataset among those included comprised 150,000 photographic images, specifically created for tooth instance segmentation, annotated by orthodontists with the aid of a human-machine hybrid algorithm[14]. Compared to 2D images, datasets for 3D image volumes, including CBCT, MRI, intraoral, facial, and model scans, are limited and smaller probably due to the challenges associated with their acquisition, annotation, and storage. In public datasets, original 3D images are often converted into the NIfTI format to facilitate more straightforward analysis due to its superior compatibility with computational tasks.

In the literature, dental AI models were developed mainly for segmentation, detection, classification, and prediction tasks[57]. Segmentation tasks

**Table 4 | Information regarding the ethical approvals, specific terms, licenses, applicability concerns, and the risk of bias in the ground truth annotations of the included datasets**

| Dataset No. | Ethical approval | Specific terms and licenses | Applicability concern | Risk of bias in the ground truth annotations |
|---|---|---|---|---|
| 1. | Not reported | Not reported | High | – |
| 2. | Not reported | CC BY NC 3.0 | Moderate | Low |
| 3. | Ethical approval was obtained from the Tufts University Institutional Research Board (IRB ID MODCR-01-12631). | Extracts of Data Use Terms<br>• Researcher shall use the Database only for non-commercial research and educational purposes<br>• Tufts University nor Panetta's Vision and Sensing System Lab makes no representations or warranties regarding the Database<br>• Researcher accepts full responsibility for his or her use of the Database<br>• The Panetta's vision and sensing systems lab reserves the right to revise, amend, alter or delete the information provided herein at any time, but shall not be responsible for or liable in respect of any such revisions, amendments, alterations or deletions<br>• No permission is granted to reproduce the database or post into any webpage or any other storage means | Low | Moderate |
| 4. | Not reported | CC BY 4.0 | Moderate | High |
| 5. | Not reported | CC BY 4.0 | Moderate | Low |
| 6. | Not reported | CC BY 4.0 | Moderate | High |
| 7. | Not reported | Not reported | High | High |
| 8. | Not reported | CC BY 4.0 | Moderate | High |
| 9. | Not reported | CC BY 4.0 | Moderate | – |
| 10. | Not reported | Not reported | Moderate | – |
| 11. | Not reported | CC BY 4.0 | Moderate | High |
| 12. | Not reported | CC0 1.0 | Moderate | High |
| 13. | Not reported | CC BY 4.0 | Moderate | High |
| 14. | Not reported | CC BY 4.0 | Moderate | High |
| 15. | Not reported | CC BY 4.0 | Moderate | Moderate |
| 16. | Not reported | CC BY-SA 4.0 | Moderate | – |
| 17. | Not reported | CC BY 4.0 | High | High |
| 18. | Not reported | Not reported | High | High |
| 19. | Not reported | CC BY 4.0 | Moderate | High |
| 20. | Ethical approval and patients' consent has been obtained. | CC0 1.0 | Low | Low |
| 21. | All the necessary permissions have been obtained from the ethics committee. | CC0 1.0 and CC-BY | Low | Low |
| 22. | Not reported | Apache 2.0 | Moderate | High |
| 23. | Not reported | Apache 2.0 | Moderate | High |
| 24. | Not reported | Apache 2.0 | Moderate | – |
| 25. | Not reported | CC BY 4.0 | Moderate | High |
| 26. | Not reported | CC BY-SA 3.0 | Moderate | – |
| 27. | Not reported | CC BY 4.0 | Moderate | High |
| 28. | Informed Consent: All patients provided their consent in accordance with the dental ethical principles. | CC BY 4.0 | Low | High |
| 29. | Not reported | Not reported | High | – |
| 30. | Not reported | Not reported | High | High |
| 31. | Ethical approval was obtained from the Ethics Committee of China-Japan Union Hospital of Jilin University [#2024011704]. | Not reported | Moderate | Low |
| 32. | Not reported | CC BY 4.0 | Moderate | High |
| 33. | Not reported | Not reported | High | High |
| 34. | Not reported | Not reported | High | – |
| 35. | Not reported | MIT | Moderate | High |
| 36. | Not reported | CC BY 4.0 | Moderate | High |
| 37. | Not reported | Apache 2.0 | Moderate | High |

**Table 4 (continued) | Information regarding the ethical approvals, specific terms, licenses, applicability concerns, and the risk of bias in the ground truth annotations of the included datasets**

| Dataset No. | Ethical approval | Specific terms and licenses | Applicability concern | Risk of bias in the ground truth annotations |
|---|---|---|---|---|
| 38. | Not reported | Apache 2.0 | Moderate | High |
| 39. | Not reported | Apache 2.0 | Moderate | High |
| 40. | Not reported | Not reported | High | High |
| 41. | Not reported | Not reported | High | High |
| 42. | Not reported | Not reported | High | High |
| 43. | Ethical approval of the institution was obtained. | CC BY 4.0 | Low | Low |
| 44. | Ethical review and approval for this study were waived due to it being retrospective and the images were taken for reasons unrelated to this study. Informed consent was obtained from all individual participants. | CC BY 4.0 | Low | – |
| 45. | Not reported | CC0 1.0 | Moderate | High |
| 46. | Not reported | Not reported | High | Low |
| 47. | Not reported | CC BY 4.0 | Moderate | High |
| 48. | Not reported | CC0 1.0 | Moderate | High |
| 49. | Not reported | Not reported | High | High |
| 50. | Not reported | Not reported | High | High |
| 51. | Not reported | Not reported | High | High |
| 52. | Not reported | CC BY 4.0 | Moderate | High |
| 53. | Not reported | Not reported | High | High |
| 54. | Not reported | CC BY 4.0 | Moderate | High |
| 55. | Not reported | Not reported | High | – |
| 56. | Not reported | CC0 1.0 | Moderate | High |
| 57. | Not reported | Not reported | High | Low |
| 58. | Not reported | Not reported | High | High |
| 59. | Not reported | Apache 2.0 | Moderate | High |
| 60. | Not reported | CC BY 4.0 | Moderate | Low |
| 61. | Not reported | CC0 1.0 | Moderate | High |
| 62. | Not reported | Apache 2.0 | Moderate | High |
| 63. | Not reported | CC0 1.0 | Moderate | High |
| 64. | Not reported | Not reported | High | High |
| 65. | Not reported | Apache 2.0 | Moderate | High |
| 66. | Not reported | Not reported | High | High |
| 67. | Not reported | Not reported | High | Low |
| 68. | Not reported | Not reported | High | – |
| 69. | Not reported | Not reported | High | – |
| 70. | Not reported | CC BY 4.0 | Moderate | High |
| 71. | Not reported | CC BY 4.0 | Moderate | – |
| 72. | Not reported | Not reported | High | – |
| 73. | Not reported | Not reported | High | High |
| 74. | Not reported | Not reported | High | – |
| 75. | Not reported | Not reported | High | High |
| 76. | Not reported | Not reported | High | – |
| 77. | Not reported | Apache 2.0 | Moderate | – |
| 78. | Not reported | Apache 2.0 | Moderate | High |
| 79. | Not reported | CC BY 4.0 | Moderate | – |
| 80. | Not reported | CC BY-NC-SA 4.0 | Moderate | High |
| 81. | Ethical approval was obtained from the Institutional Review Board of China Medical University Hospital. (CMUH 107-REC3-092). | CC BY 4.0 | Low | – |
| 82. | Not reported | Not reported | High | High |

**Table 4 (continued) | Information regarding the ethical approvals, specific terms, licenses, applicability concerns, and the risk of bias in the ground truth annotations of the included datasets**

| Dataset No. | Ethical approval | Specific terms and licenses | Applicability concern | Risk of bias in the ground truth annotations |
|---|---|---|---|---|
| 83. | Ethical approval was obtained from Comitato Etico dell'Area Vasta Emilia Nord (Approval Number 1374/2020/OSS/ESTMO SIRER ID 1275 - NAI-CBCT-D). | Not reported | Moderate | Low |
| 84. | Ethical approval of was obtained from the Medical Ethics Committee of Sichuan Provincial People's Hospital and the University of Electronic Science and Technology Hospital Research Ethics Committee (No. 2022YR014). | CC BY-NC-ND | Low | Low |
| 85. | Ethical approval was obtained from the Ethical Committee of the Medical Faculty of the University of Heidelberg under Application No. S-039/2016. | Not reported | Moderate | – |
| 86. | Not reported | CC BY 4.0 | Moderate | – |
| 87. | Not reported | Not reported | High | High |
| 88. | Not reported | CC BY-NC-ND 4.0 | Moderate | Low |
| 89. | Not reported | CC BY 4.0 | Moderate | Moderate |
| 90. | Not reported | CC BY-NC 4.0 | Moderate | Low |
| 91. | Not reported | Not reported | High | High |
| 92. | Not reported | Not reported | High | – |
| 93. | Not reported | CC BY NC 3.0 | Moderate | Low |
| 94. | Ethical approval was obtained from the Research Ethics Committee of the Hospital Universitário Cassiano Antonio de Moraes da Universidade Federal do Espírito Santo under registration no. 5,022,438 | CC BY 4.0 | Low | Low |
| 95. | Ethical approval was obtained from the Ethical Committee of Human Studies of Institute of Advanced Study in Science and Technology, Guwahati, Assam with registration number IEC (HS)/IASST/1082/ 2014-15/2. | CC BY 4.0 | Low | Moderate |
| 96. | Ethics approval was obtained from the Ethics Committee on the Use of Animals under protocol numbers 038/09 and A016/ 21 at the Federal University of Uberlândia, Brazil. | Not reported | Moderate | Low |
| 97. | Not reported | CC BY 4.0 | Moderate | High |
| 98. | Ethical approval was obtained from the Ethics Committee (certificate number CAAE - 39212420.9.0000.5347). | Not reported | Moderate | Low |
| 99. | Not reported | Apache 2.0 | Moderate | High |
| 100 | Ethical approval was obtained from the Health Research Authority (HRA) and the Joint Research Management Ofce (JRMO) (LREC 22/PR/0058). | CC BY 4.0 | Low | Moderate |
| 101 | Not reported | CC BY 4.0 | Moderate | Moderate |
| 102 | Not reported | Not reported | High | High |
| 103 | Not reported | Not reported | High | High |
| 104 | Ethical approval was obtained from the Zgoda nr KB-14/22 wystawiona przez Bioethics Committee at the Regional Medical Chamber in Gdańsk | CC BY 4.0 | Low | High |
| 105 | Not reported | CC BY-NC | Moderate | High |

involve dividing an image into distinct sections based on variations in pixel intensity among different tissues. Detection tasks aim to localize objects within an image using class-labelled bounding boxes. Classification tasks assign a categorical label to an entire image, while prediction tasks estimate the likelihood of a certain event based on existing risk factors. Obtaining annotations for segmentation models is relatively straightforward as they can be completed through visual inspection of images[58]. On the contrary, obtaining annotations for the development of more clinically significant diagnostic models, such as models for detecting the onset of specific diseases or for diagnosing lesions that are indistinguishable from diagnostic images, are challenging. These annotations often rely on particular clinical, laboratory, or biopsy examinations[11]. Our findings reveal that the most common types of annotation from the included datasets are the mask (29%), bounding box (29%), and categorical label (20%). Notably, the annotations provided across these datasets for similar tasks differ significantly due to different labelling methods used. The diversity in annotation approaches can complicate the integration and use of annotations from datasets created for similar tasks. Moreover, the annotations for similar oral conditions differed across datasets and often lacked detailed descriptions. Thus, such annotations should be reused with caution to ensure their accuracy and precision.

While nearly 80% of the 105 datasets provided image annotations, only one-fourth of these datasets specified the annotators' qualifications. Notably, even when qualifications were mentioned, detailed information regarding the annotators' experience in dental specialties or annotation practices was rarely disclosed. The lack of this information increases the uncertainty of the annotation accuracy, affecting the reliability of open-access images and their

corresponding ground truth annotations. Even though some annotations were carried out by specialists, the accuracy of these annotations might not be guaranteed or suitable for direct use in specific AI projects. Manual adjustments or re-annotations of these annotations may be necessary to meet the requirements for certain projects. This study included nine histopathological image datasets containing various cell types, including normal oral cavity epithelium, oral cancer cells, epithelial dysplasia and Leukoplakia cells. However, only four datasets explicitly stated that the annotations were performed by pathologists or specialists in pathology. Therefore, caution is advised when reusing these annotations, especially those with unknown origins or uncertainties for the development or validation of AI models.

Unlabelled images can be effectively utilized in AI model training through self-supervised learning techniques, such as contrastive learning, mask image modelling[59,60], and semi-supervised learning[5,61]. Self-supervised learning enables models to learn data distribution without manual labels by using pretext tasks that exploit the inherent structure of the data to generate labels. This method uses large amounts of unlabelled data to learn useful representations. Subsequently, a smaller set of labelled data is employed to fine-tune the model for specific tasks. This approach minimizes the dependence on extensive manual annotations and is beneficial for utilizing large unlabelled datasets efficiently.

The majority (83.8%) of the datasets did not disclose whether they had obtained ethical approval. This finding indicates a critical area in data usage and ethics that requires further attention. Some included studies have stated that their open datasets were derived from projects with ethical approval. However, this does not automatically grant permission for others to reuse the image data. Ethical approval confirms that the initial study is in compliance with ethical standards, but it does not extend to the subsequent use of the data by third parties[62]. Sharing patient data with either internal or external teams is often essential for AI project development and validation, which may not be explicitly covered in the original ethical approvals. The Europe General Data Protection Regulation legislation highlights the necessity of strict data processing regulations, which limit health data use unless explicit consent is given, ensuring that data processing aligns with protecting individuals' vital interests[63]. However, details about patient consent are often missing in publicly accessible datasets. This situation raises serious ethical concerns about data sharing and patient consent, especially when developing AI applications in healthcare.

Public accessibility of datasets does not automatically grant unlimited usage rights, as licensing clearly defines the terms for data reuse. Dataset licenses allow creators to specify rights they reserve and those they waive. Without explicit licensing, even ethically approved datasets can still cause legal and ethical issues when reused. Common licenses include CC0-1.0 and various Creative Commons (CC) licenses, such as CC-BY, CC-BY-NC, CC-BY-SA, and CC-BY-ND[64]. The CC0-1.0 license permits creators to waive all their copyright and related rights in their works as much as legally possible. Other CC licenses provide options that retain copyright while allowing various levels of permission. For instance, CC-BY-NC allows non-commercial reuse, CC-BY permits modifications and commercial use with attribution, CC-BY-SA requires any adaptations to be shared under identical terms, and CC-BY-ND allows only unchanged and whole redistribution with proper credit. Of the datasets, 61.9% specified a license for their reuse, with over 50% licensed with CC BY, followed by Apache 2.0 (18.5%). In cases where a single dataset carries multiple licenses, such as one panoramic radiograph dataset with dual licenses (CC0 1.0 and CC-BY)[65], the strictest of the licenses is applied.

While 61.9% of the datasets specified a license for reuse, some of them might have possibly mislabelled the license on dataset platforms. This contributes to the uncertainty regarding whether the openly accessible oral-maxillofacial imaging datasets were released with valid reuse terms or license, placing them in a legal grey area. Using copyrighted datasets for training AI models can potentially lead to legal issues[62]. The common practice of creating a training dataset by repackaging existing open-source datasets can be problematic. If a dataset is protected by NoDerivatives licenses, such as CC-BY-ND, it cannot be included in a dataset to train an AI model. In such case, the trained model could be considered a derivative of the training data, violating the exclusive rights of the copyright holders. Similarly, if an AI model is trained using a dataset protected by licenses permitting only non-commercial reuse, future commercialization of the trained model might be restricted[62]. These evolving legal issues regarding dataset reuse are gaining attention from academic organizations, industry labs, and research institutions. Therefore, reusing these datasets should be cautious due to potential legal issues. Schwabe et al. introduced the METRIC-framework, which provides a systematic approach for assessing training datasets, establishing reference datasets, and designing test datasets[66]. This framework proposes fifteen awareness dimensions across five data management clusters, including measurement process (device error, human-induced error, completeness, and source credibility), timeliness (timeliness), representativeness (variety, depth of data, target class balance), informativeness (understandability, redundancy, informative missingness, feature importance), consistency (rule-based consistency, logical consistency, and distribution consistency). These dimensions could contribute to the development of clear, standardized guidelines for the ethical reuse of publicly accessible medical and dental image datasets, while strictly complying with licensing requirements.

This systematic review has limitations. First, due to the large number of images from the included datasets, it is not practical to assess the quality of all images. Since image quality is often assessed for specific clinical indications, the quality of images from the included datasets should be evaluated by interested researchers based on their intended tasks. Second, some crucial factors such as metadata completeness, identification of data reuse issues, and data traceability were not included in the assessment of the risk-of-bias for the included datasets, which might not be able to fully account for all potential biases introduced into the datasets. Furthermore, this study excluded certain large, high-quality image datasets[48] as the access could not be obtained due to a lack of response from the dataset owners, despite following the requirements for registered access. Moreover, the datasets released may be subject to continual updates without any official notification. Therefore, the changes in the number and annotations of images from the datasets should be confirmed with caution before reuse.

In conclusion, this study has systematically identified 105 public oral-maxillofacial imaging datasets and investigated their sources, characteristics, and ethical and licensing considerations. While the majority of the datasets included annotations, only some specified the annotators' qualifications. Furthermore, more than half of the datasets specified the terms or licenses for reuse, but most did not disclose whether ethical approval was obtained. These findings highlight the need for careful consideration of ethical and legal implications when reusing these datasets and suggest the need to establish clear, standardized guidelines for reusing publicly accessible image datasets.

## Methods

This systematic review was conducted in accordance with the guidance of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)[19]. The PRISMA checklist used for this review is provided in Supplementary Table S2. The study protocol has been registered on the Open Science Framework (OSF) (Registration https://doi.org/10.17605/OSF.IO/SFN5C). The focused question guiding the search was, "Which open-source datasets related to images from the oral-maxillofacial region are available?".

### Search strategy and selection criteria

The search strategy consisted of two components, including the search of two electronic scientific literature databases (PubMed and Google scholar) and three widely-used dataset management platforms (Google Dataset Search, Kaggle, and Hugging face) to identify as many publicly accessible image datasets as possible. The search was conducted in September 2024. The literature search combined free-text terms of ("dentistry" OR "dental" OR "oral" OR "maxillofacial") AND ("open source" OR "open access" OR "publicly available" OR "publicly accessible") AND ("data" OR "dataset" OR "repository") AND "images". Vocabulary and syntax were adjusted accordingly for each database. The search terms used on dataset management platforms were "dentistry" OR "dental" OR "oral-maxillofacial" OR

"dental image" OR "oral image". The specific search strategies used for all databases are provided in Supplementary Table S3.

The electronic literature database search was conducted without any restrictions on the publication period. The criteria for inclusion were:

1. Original and review articles published in English;
2. Studies that report a dataset comprised of any type of image modalities generating images from the oral-maxillofacial region, including scans of dental models from patients; and
3. Studies providing the access to the dataset.

Studies were excluded if one of the following exclusion criteria was met.

1. Studies reporting a dataset consisting of images not from human subjects;
2. Studies reporting a dataset consisting of images from cadavers or extracted teeth;
3. Studies reporting a dataset consisting of images where the oral-maxillofacial region was obscured using a blocking technique;
4. Studies reporting a dataset consisting of images that were included in the most recently updated dataset from the same source; or
5. Studies where the full text is not available or accessible.

For the dataset management platforms search, all open-source datasets consisting of images from the oral-maxillofacial region were considered eligible. The exclusion criteria were:

1. Datasets consisting of images not from human subjects;
2. Dataset consisting of images from cadavers or extracted teeth;
3. Datasets consisting of images where the oral-maxillofacial region was obscured using a blocking technique;
4. Datasets consisting of images that were included in the most recently updated datasets from the same source;
5. Datasets consisting of image files that were corrupted and could not be opened; or
6. Datasets that require payment for access.

All records retrieved from the electronic literature database search were compiled using the reference manager software (EndnoteTM Version 21, Clarivate Analytics, New York, USA). The titles were automatically checked for duplicates. Two independent reviewers (J.H. and K.F.H.) screened the titles and abstracts of each record to select studies for further full-text evaluation. Reviewer K.F.H. is a professoriate faculty member in the subdivision of Oral-Maxillofacial Radiology with over ten years of experience in conducting diagnostic imaging studies. Reviewer J.H. is a PhD candidate at the same institution with more than five years of research experience in the development of artificial intelligence algorithms and is experienced in the collection and evaluation of AI-related public datasets. Additional manual searches on the reference lists of the included studies were conducted independently by two reviewers (J.H. and K.F.H.) to further identify potentially eligible studies that met the inclusion criteria. Subsequently, the two reviewers (J.H. and K.F.H.) independently assessed the full-texts of the included studies. The two reviewers compared the studies they identified as eligible, and then discussed their reasons for considering certain studies to be included based on the defined inclusion and exclusion criteria. Agreement was reached through discussion. In cases where agreement could not be achieved, a third experienced reviewer (Q.Y.H.A) was consulted to assist in reaching a consensus. Inter-reviewer agreement was evaluated by calculating Cohen's kappa values. Eligible datasets identified from the dataset management platforms were organized using an Excel spreadsheet (Microsoft Corporation, Redmond, Washington). Any duplicates from the electronic literature database search and the dataset management platform search were eliminated.

### Extraction of dataset characteristics and outcome of interest

Details regarding the year and purpose of dataset creation, creators, country and institution of origin, imaging modality, image type and format, the number of patients and images in the dataset, the manufacturer of the imaging device, image annotation details, the qualification of the annotators, and dataset access, were extracted by two reviewers (J.H. and K.F.H.) from the included studies and the metadata of the datasets. In addition, information pertaining to the acquisition of ethical approval for image collection as well as specific terms, conditions, and licensing requirements for reusing these datasets were collected. Any discrepancies detected in the extracted data were resolved through discussion. In the case of a discrepancy between the information provided in the included studies and the dataset repository, the information from the repository was used in this study. All data were systematically tabulated using a standardized template created in an Excel spreadsheet (Microsoft Corporation, Redmond, Washington).

### Dataset accessibility

The accessibility of the datasets included in this study were divided into two categories as follows:

1. Datasets that are readily accessible and can be directly downloaded without any requirement.
2. Datasets that necessitate registered access, requiring submission of an email request or the creation of an account. Upon fulfilling these requirements, a download link for the dataset would be sent to the applicant's email. The accessibility status of these datasets was reconfirmed at the time of manuscript submission.

### Evaluation of applicability concerns of the included datasets and the risk of bias in annotations

The applicability concerns of the included datasets and the risk of bias in annotations were assessed independently by two reviewers (J.H. and K.F.H.). Any discrepancies were resolved through discussion. A dataset was deemed to have a "low" applicability concern if it reported both ethical approval as well as the terms or licensing requirements for its reuse. If only either ethical approval or terms or licenses were reported, the concern was classified as "moderate". If neither was reported, the concern was rated as "high". The assessment of the risk of bias in annotations focused on the reliability of the reference standard (i.e., ground truth annotations), which is one of the four domains proposed by the Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2), a tool widely used in diagnostic imaging studies[67]. According to the QUADAS-2, the risk of bias in the reference standard should be assessed by the signalling question "Is the reference standard likely to correctly classify the target condition?". Based on the proposed signalling question, the risk of bias in the ground truth annotations was assessed by evaluating the reliability of the reference standard used for annotation. For datasets with annotations, the "low" risk-of-bias rating was assigned to datasets where ground truth annotations are confirmed by at least two annotators with explicit medical/dental qualifications, or those supported by clinically or pathologically confirmed results. Datasets with ground truth annotations determined by a single qualified annotator, and those involving at least two annotators identified as experts but without explicit qualifications, were given a "moderate" risk-of-bias rating. All remaining datasets were categorized as having a "high" risk of bias.

## Data availability

Data sharing is not applicable to this article as no datasets were generated during the current study. All the access links to the datasets analysed in this study are provided in Supplementary Table S1.

## Code availability

Not applicable.

Received: 27 February 2025; Accepted: 20 June 2025;
Published online: 05 July 2025

## References

1. Joda, T., Yeung, A., Hung, K., Zitzmann, N. & Bornstein, M. Disruptive innovation in dentistry: what it is and what could be next. *J. Dent. Res.* **100**, 448–453 (2021).

2. Hung, K., Yeung, A. W. K., Tanaka, R. & Bornstein, M. M. Current applications, opportunities, and limitations of AI for 3D imaging in dental research and practice. *Int. J. Environ. Res. Public Health* **17**, 4424 (2020).
3. Hung, K. F. et al. Current applications of deep learning and radiomics on CT and CBCT for maxillofacial diseases. *Diagnostics* **13**, 110 (2022).
4. Hung, K., Montalvao, C., Tanaka, R., Kawai, T. & Bornstein, M. M. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: a systematic review. *DMFR* **49**, 20190107 (2020).
5. Hao, J. et al. A semi-supervised transformer-based deep learning framework for automated tooth segmentation and identification on panoramic radiographs. *Diagnostics* **14**, 1948 (2024).
6. Hao, J. et al. T-Mamba: a unified framework with long-range dependency in dual-domain for 2D & 3D tooth segmentation. Preprint at https://arxiv.org/abs/2404.01065 (2024).
7. Razaghi, M., Komleh, H. E., Dehghani, F. & Shahidi, Z. Innovative diagnosis of dental diseases using YOLO V8 deep learning model. In *IEEE 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)*, 1–5 (IEEE, 2024).
8. Xu, C. et al. TeethDreamer: 3D teeth reconstruction from five intra-oral photographs. In *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2024*, 712–721 (Springer International Publishing, Cham, 2024).
9. Mei, L. et al. DTR-net: dual-space 3D tooth model reconstruction from panoramic X-Ray images. *IEEE Trans. Med. Imaging* **43**, 517–528 (2023).
10. Gao, N. et al. Multi-level objective alignment transformer for fine-grained oral panoramic X-ray report generation. *IEEE Trans. Multimed.* **26**, 7462–7474 (2024).
11. Hung, K. F., Yeung, A. W. K., Bornstein, M. M. & Schwendicke, F. Personalized dental medicine, artificial intelligence, and their relevance for dentomaxillofacial imaging. *DMFR* **52**, 20220335 (2023).
12. Krois, J. et al. Generalizability of deep learning models for dental image analysis. *Sci. Rep.* **11**, 6102 (2021).
13. Cui, W. et al. Ctooth: a fully annotated 3d dataset and benchmark for tooth volume segmentation on cone beam computed tomography images. In *Intelligent Robotics and Applications*, 191–200 (ICIRA, 2022).
14. Zou, B. et al. Teeth-SEG: an efficient instance segmentation framework for orthodontic treatment based on multi-scale aggregation and anthropic prior knowledge. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 11601–11610 (IEEE, 2024).
15. Uribe, S. E. et al. Publicly available dental image datasets for artificial intelligence. *J. Dent. Res.* **103**, 1365–1374 (2024).
16. Ohmann, C. et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ open* **7**, e018647 (2017).
17. Meystre, S. M. et al. Clinical data reuse or secondary use: current status and potential future progress. *Yearb. Med. Inform.* **26**, 38–52 (2017).
18. Duke, C. S. & Porter, J. H. The ethics of data sharing and reuse in biology. *BioScience* **63**, 483–489 (2013).
19. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. & Group, P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int. J. Surg.* **8**, 336–341 (2010).
20. Abdi, A. H., Kasaei, S. & Mehdizadeh, M. Automatic segmentation of mandible in panoramic x-ray. *J. Med. Imaging* **2**, 044003–044003 (2015).
21. Panetta, K., Rajendran, R., Ramesh, A., Rao, S. P. & Agaian, S. Tufts dental database: a multimodal panoramic x-ray dataset for benchmarking diagnostic systems. *IEEE J. Biomed. Health Inform.* **26**, 1650–1659 (2021).
22. Zhang, Y. et al. Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. *Sci. Data* **10**, 380 (2023).
23. Hamamci, I. E. et al. Dentex: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays. Preprint at https://arxiv.org/abs/2305.19112 (2023).
24. Zhou, W. et al. A dual-labeled dataset and fusion model for automatic teeth segmentation, numbering, and state assessment on panoramic radiographs. *BMC Oral. Health* **24**, 1201 (2024).
25. Ke, W. et al. Biological gender estimation from panoramic dental x-ray images based on multiple feature fusion model. *Sens. Imaging* **21**, 1–11 (2020).
26. Murga & Stefan. RGB-D tongue state classification dataset. *Borealis* V1 https://doi.org/10.5683/SP2/5T2RD9 (2019).
27. Rashid, J. et al. Mouth and oral disease classification using InceptionResNetV2 method. *Multim. Tools Appl.* **83**, 33903–33921 (2024).
28. Kusakunniran, W. et al. Deep Upscale U-Net for automatic tongue segmentation. *MBEC* **62**, 1751–1762 (2024).
29. Liao, C. W. et al. Self-assembled micro-computed tomography for dental education. *PLoS One* **13**, e0209698 (2018).
30. Gholamalizadeh, T. et al. Open-Full-Jaw: an open-access dataset and pipeline for finite element models of human jaw. *Comput. Methods Prog. Biomed.* **224**, 107009 (2022).
31. Cipriano, M. et al. Deep segmentation of the mandibular canal: a new 3D annotated dataset of CBCT volumes. *IEEE Access* **10**, 11500–11510 (2022).
32. Wang, Y. et al. STS MICCAI 2023 challenge: grand challenge on 2D and 3D semi-supervised tooth segmentation. Preprint at https://arxiv.org/abs/2407.13246 (2024).
33. Andlauer, R. et al. 3D-guided face manipulation of 2D images for the prediction of post-operative outcome after cranio-maxillofacial surgery. *IEEE Trans. Image Process.* **30**, 7349–7363 (2021).
34. Ben-Hamadou, A. et al. Teeth3DS+: an extended benchmark for intraoral 3D scans analysis. Preprint at https://arxiv.org/abs/2210.06094v1 (2022).
35. Abdel-Alim, T. et al. Quantifying dysmorphologies of the neurocranium using artificial neural networks. *J. Anat.* **245**, 903–913 (2024).
36. Ribeiro-de-Assis, M. C. F. et al. NDB-UFES: an oral cancer and leukoplakia dataset composed of histopathological images and patient data. *Data Brief.* **48**, 109128 (2023).
37. Rahman, T. Y., Mahanta, L. B., Das, A. K. & Sarma, J. D. Automated oral squamous cell carcinoma identification using shape, texture and color features of whole image strips. *Tissue Cell* **63**, 101322 (2020).
38. Silva, A. B. et al. Oralepitheliumdb: A dataset for oral epithelial dysplasia image segmentation and classification. *J. Imaging Inf. Med.* **37**, 1691–1710 (2024).
39. Hsieh, H.-C. et al. Deep learning-based automatic image classification of oral cancer cells acquiring chemoresistance in vitro. *PloS One* **19**, e0310304 (2024).
40. Rönnau, M. M. et al. Automatic segmentation and classification of Papanicolaou-stained cells and dataset for oral cancer detection. *Comput. Biol. Med.* **180**, 108967 (2024).
41. Ruthven, M., Peplinski, A. M., Adams, D. M., King, A. P. & Miquel, M. E. Real-time speech MRI datasets with corresponding articulator ground-truth segmentations. *Sci. Data* **10**, 860 (2023).
42. Zeng, M., Yan, Z., Liu, S., Zhou, Y. & Qiu, L. Cascaded convolutional networks for automatic cephalometric landmark detection. *Med. Image Anal.* **68**, 101904 (2021).
43. Adnan, N. & Umer, F. Orthopantomogram teeth segmentation and numbering dataset. *Data Brief.* **57**, 111152 (2024).
44. Román, J. C. M. et al. Panoramic dental radiography image enhancement using multiscale mathematical morphology. *Sensors* **21**, 3110 (2021).

45. Wang, C.-W. et al. A benchmark for comparison of dental radiography analysis algorithms. *Med. Image Anal.* **31**, 63–76 (2016).
46. Wang, C.-W. et al. Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge. *IEEE Trans. Med. Imaging* **34**, 1890–1900 (2015).
47. Lindner, C. et al. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Sci. Rep.* **6**, 33581 (2016).
48. Shi, J. et al. Semantic decomposition network with contrastive and structural constraints for dental plaque segmentation. *IEEE Trans. Med. Imaging* **42**, 935–946 (2022).
49. Ramakrishnan, D. et al. A large open access dataset of brain metastasis 3D segmentations on MRI with clinical and imaging information. *Sci. Data* **11**, 254 (2024).
50. Chilamkurthy, S. et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* **392**, 2388–2396 (2018).
51. Iosifidis, A., Marami, E., Tefas, A., Pitas, I. & Lyroudia, K. The MOBISERV-AIIA eating and drinking multi-view database for vision-based assisted living. *J. Inf. Hiding Multim. Signal Process.* **6**, 254–273 (2015).
52. Ranjbar, S. et al. Weakly supervised skull stripping of magnetic resonance imaging of brain tumor patients. *Front. Neuroimaging* **1**, 832512 (2022).
53. Jian, G. Aoralscan3 tooth segmentation dataset. *IEEE Dataport*. https://doi.org/10.21227/w9mp-5w63 (2023).
54. Wen, D. et al. Characteristics of publicly available skin cancer image datasets: a systematic review. *Lancet Digit. Health* **4**, e64–e74 (2022).
55. Khan, S. M. et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit. Health* **3**, e51–e66 (2021).
56. Ni, Z., Bousquet, C., Vaillant, P. & Jaulent, M.-C. Rapid review on publicly available datasets for health misinformation detection. *Healthc. Transform. Inf. Artif. Intell.*, **305**, 123–126 (2023).
57. Hung, K. F., Ai, Q. Y. H., Leung, Y. Y. & Yeung, A. W. K. Potential and impact of artificial intelligence algorithms in dento-maxillofacial radiology. *Clin. Oral. Investig.* **26**, 5535–5555 (2022).
58. Hung, K. F. et al. Automatic detection and segmentation of morphological changes of the maxillary sinus mucosa on cone-beam computed tomography images using a three-dimensional convolutional neural network. *Clin. Oral. Investig.* **26**, 3987–3998 (2022).
59. Xie, Z. et al. Simmim: a simple framework for masked image modelling. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 9653–9663 (IEEE, 2022).
60. Chen, Z. et al. Masked image modeling advances 3d medical image analysis. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 1970–1980 (IEEE, 2023).
61. Jing, Y. et al. USCT: uncertainty-regularized symmetric consistency learning for semi-supervised teeth segmentation in CBCT. *Biomed. Signal Process. Control.* **91**, 106032 (2024).
62. Longpre, S. et al. A large-scale audit of dataset licensing and attribution in AI. *Nat. Mach. Intell.* **6**, 975–987 (2024).
63. Dziedzic, A., Issa, J., Chaurasia, A. & Tanasiewicz, M. Artificial intelligence and health-related data: the patient's best interest and data ownership dilemma. *Proc. Inst. Mech. Eng. H.* **238**, 1023–1028 (2024).
64. Kim, M. The Creative Commons and copyright protection in the digital era: Uses of Creative Commons licenses. *JCMC* **13**, 187–209 (2007).
65. Rahman, R. B. et al. Dental OPG XRAY dataset. *Mendeley Data*, V4, https://doi.org/10.17632/c4hhrkxytw.4 (2024).
66. Schwabe, D., Becker, K., Seyferth, M., Klaß, A. & Schaeffter, T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *NPJ Digit. Med.* **7**, 203 (2024).
67. Whiting, P. F. et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* **155**, 529–536 (2011).
68. Abdi, A. & Kasaei, S. Panoramic dental X-rays with segmented mandibles. *Mendeley Data V2* https://doi.org/10.17632/hxt48yk462.2 (2020).
69. Brahmi, W. & Jdey, I. Automatic tooth instance segmentation and identification from panoramic X-Ray images using deep CNN. *Multimed. Tools Appl.* **83**, 55565–55585 (2024).
70. Farzi, I., Kazemi, A. & Hosseini, M. Panoramic radiography Images with diagnosis of need for Apicoectomy surgery as label. *Mendeley Data*, V1, https://doi.org/10.17632/9d8mcyp284.1 (2023).
71. Budagam, D. et al. Instance segmentation and teeth classification in panoramic X-rays. Preprint at https://arxiv.org/abs/2406.03747 (2024).
72. Waqas, M., Hasan, S., Khurshid, Z. & Kazmi, S. OPG dataset for Kennedy classification of partially edentulous arches. *Mendeley Data*, V1, https://doi.org/10.17632/ccw5mvg69r.1 (2024).
73. Sengupta, N., Sarode, S. C., Sarode, G. S. & Ghone, U. Scarcity of publicly available oral cancer image datasets for machine learning research. *Oral. Oncol.* **126**, 105737 (2022).
74. Chandrashekar, H. S.,Geetha Kiran, A., Murali, S., Dinesh, M. S. & Nanditha, B. R. Oral images dataset. *Mendeley Data*, V2, https://doi.org/10.17632/mhjyrn35p4.2 (2021).
75. Hoang B. D. A Dental intraoral image dataset of gingivitis for image captioning. *Mendeley Data*, V1, https://doi.org/10.17632/3253gj88rr.1 (2024).
76. Nisrean, T. A dataset of dental periapical X-ray. *Mendeley Data*, V1, https://doi.org/10.17632/8ys8jssm9k.1 (2023).
77. Viet, D. Panoramic radiographs with periapical lesions Dataset. *Mendeley Data*, V3, https://doi.org/10.17632/kx52tk2ddj.3 (2024).
78. Li, C. et al. Efficient complete denture metal base design via a dental feature-driven segmentation network. *Comput. Biol. Med.* **175**, 108550 (2024).


## Acknowledgements

This study was supported by the Seed Fund for Basic Research for New Staff at the University of Hong Kong (103036002).

## Author contributions

Conceptualisation: K.F.H., Q.Y.H.A., M.M.B. and J.K.H.T.; Methodology: K.F.H., Q.Y.H.A., A.W.K.Y, M.M.B. and J.K.H.T.; Data curation and investigation: J.H., K.F.H., and Q.Y.H.A.; Writing—original draft preparation, J.H. and K.F.H.; writing—review and editing, A.N., A.W.K.Y, R.T., Q.Y.H.A., W.Y.H.L., Y.Y.L., Z.S., A.A., M.M.B., C.M. and J.K.H.T. All authors have read and approved the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01818-5.

**Correspondence** and requests for materials should be addressed to Kuo Feng Hung.