MELBA Journal Machine Learning for Biomedical Imaging

A Histopathological Image Dataset for Oral Squamous Cell Carcinoma and Normal Tissue Analysis

Salma Bahaa [1]

Oral Pathology, Faculty of Dentistry, Alexandria University, Egypt Champollion, Elazarita, 21563, Egypt

Abstract

Oral squamous cell carcinoma accounts for nearly 90% of all oral malignancies worldwide. While early detection is critical to improve the disappointed 35%–45% advanced-stage survival rate, manual histopathological evaluation remains a laborious process prone to intra-observer variability and physical strain. Artificial intelligence offers a promising decision-support solution to optimize diagnostic accuracy; however, existing open-source datasets frequently suffer from distinct limitations, including a lack of true normal tissue controls, the confounding grouping of dysplasia with invasive cancer, or a restriction to single-magnification views that fail to mimic a pathologist's multi-scale workflow. To address these gaps, we introduce a curated, expert-labeled, AI-ready histopathological dataset consisting of 1,714 images obtained from 115 patients, explicitly designed to differentiate between OSCC and completely healthy oral tissue. For clinical usage, the repository provides paired images at low-power magnification to evaluate overall tissue architecture and high-power magnification to distinguish cellular malignancy criteria. Crucially, the dataset is partitioned using a strict patient-wise split rather than an image-level split, successfully eliminating the risk of data leakage during the training, validation, and testing phases of machine learning models. By offering a clean binary baseline with dual-magnification features, this publicly available dataset provides a robust framework for developing scalable deep learning tools optimized for early oral cancer screening. Link of dataset: Bahaa, Salma (2026), "Histopathological images of oral squamous cell carcinoma and normal tissue", Mendeley Data, V1, doi: 10.17632/7m9zkcx539.1.163

Keywords

Oral Squamous Cell Carcinoma, Histopathology, AI-Ready Dataset

Article informations

https://doi.org/10.59275/j.melba.2026-273d

Volume 2026, Received: 2026-06-19, Published 2026-09-31

Corresponding author: s.halawa164183.na@alexu.edu.eg

Special issue: MICCAI Open Data 2026 x MELBA

Guest editors: AT, MS et al.

1. Background

Oral and lip cancer ranked as the 16th most frequent cancer and the 15th leading cause of cancer mortality worldwide in 2022, with 389,846 new cases detected according to the World Health Organization (WHO). In Egypt, the disease ranks 18th in both occurrence and mortality rates, accounting for 1,578 newly diagnosed cases (Ferlay et al., 2024). Accounting for nearly 90% of all oral malignancies, oral squamous cell carcinoma (OSCC) represents the most common type of cancer found within the oral cavity (Muralidharan et al., 2025). Although advancements have been made in OSCC treatment modalities, the prognosis remains poor. The 5-year overall survival rate still drops significantly from 70%–80% for early-stage diagnoses down to just 35%–45% when the disease reaches an advanced stage (Xu et al., 2025). Therefore, improving capabilities for early detection serves as a pivotal intervention for reducing cancer morbidity and securing better prognosis (Huang, 2026).

For the reliable diagnosis of OSCC, histopathological examination remains the definitive gold standard among available diagnostic methods (Li, 2025). Routine microscope work predisposes pathologists to dizziness and cervical strain, Moreover, as cancer cases rapidly escalate, reviewing massive volumes of tissue slides becomes an increasingly laborious process. Furthermore, manual analysis is prone to observer variabilities, leading to a critical lack

of diagnostic accuracy (Li and Fevens, 2020).

Therefore, A decision-support tool is required to confirm a pathologist's diagnosis, Hence Artificial intelligence (AI) can play a critical role in optimizing diagnostic accuracy (Sukegawa et al., 2023).

To realize this objective in this specific context, focuses on gathering an extensive, multi-center global collection of OSCC samples and normal tissue controls is the initiative to help in training, deployment and improving of different AI models. To date, research in this domain has relied on a limited number of existing cohorts. In (Freire et al., 2022) dataset, they extracted 1,020 image patches, each measuring 640 x 640 pixels, from the scanned whole slide images. Then a specialist manually outlined the tumor boundaries in every patch for segmentation then verified by a senior pathologist. For (Rahman et al., 2023) work, Their dataset contains 1,224 H&E-stained histopathological images from 230 patients, expertly curated using a Leica ICC50 HD microscope and split across two magnification levels. The 100x magnification set includes 528 images (89 normal oral epithelium and 439 OSCC), while the 400x magnification set consists of 696 images (201 normal epithelium and 495 OSCC). In (Falcão Ribeiro de Assis et al., 2023) work, The NDB-UFES dataset, collected between 2010 and 2021 at the Federal University of Espírito Santo (UFES) in Brazil, pairs full histopathological images and image patches with comprehensive clinical and sociodemographic metadata from patients. Notably, the repository focuses entirely on abnormal conditions and contains no normal tissue controls. However, it does not explicitly differentiate between OSCC and dysplasia samples, electing to group them together. For (Hammad et al., 2026) work, The SinaiU-OMD dataset is a curated collection of 300 anonymized histopathological images but expands beyond normal tissue and OSCC to include other oral and maxillofacial lesions such as odontogenic keratocyst, Ameloblastoma (plexiform type), and Pleomorphic adenoma. For (Guan et al., 2025), The Multi-OSCC dataset features a large cohort of 1,325 patients, with 6 high-resolution images captured per individual across $\times$200, $\times$400, and $\times$1000 magnifications. For each magnification, the images cover both the core and the invasive edge of the tumor. Moving beyond basic classifications, the dataset is richly labeled across six clinical tasks tracking crucial diagnostic and prognostic outcomes.

Although the presence of previous work in this field, many repositories lack true "normal" tissue samples, meaning that AI models only learn what cancer looks like without ever seeing what healthy oral tissue looks like for comparison. Second, when datasets do include pre-cancerous conditions like dysplasia, they often put them together with actual invasive cancer under the same category. This blurring of lines makes it impossible for an AI to learn the exact boundaries needed for precise, early-stage diagnosis. Finally, most datasets only provide images at a single magnification level. In real life, pathologists diagnose cancer by constantly shifting between low power to see the overall tissue structure and high power to see individual cell details. Training an AI on single-magnification images limits its ability to understand these multi-scale features. Because of these limitations, current AI models often struggle to perform accurately when tested in real-world hospital workflows.

To fill these gaps, we introduce a new, carefully organized dataset that shows the clear difference between healthy and cancerous tissue. By keeping normal tissue and oral cancer strictly separated, it removes the confusing labels found in older datasets. It also includes images at both low and high magnification levels, allowing AI models to learn exactly like a human pathologist does.

2. Summary

The main goal of this dataset is to help researchers build better AI tools for detecting oral cancer by providing clear, high-quality images that show the exact difference between healthy tissue and cancer. The dataset is completely AI-ready, meaning it is already organized, labeled by experts, and split into low-power ($100\times$) and high-power ($400\times$) magnification levels so machine learning models can start training on it immediately, it was split by cases to help in dividing data into training, validation and testing without the risk of data leakage. To use this dataset, researchers just need a standard computer setup for deep learning, which includes a modern graphics card (GPU), Python, and common AI frameworks like PyTorch or TensorFlow to process the images.

3. Discussion

The biggest strength of this dataset is situated in organization and different magnifications. By capturing images at both magnification levels, it allows AI models to look at both the overall tissue structure like epithelial pearls and individual cell malignant criteria, just like a real pathologist does. Also, because we kept a strict, expert-checked line between completely healthy tissue and actual cancer, there are no confusing or mixed labels. The images were divided by cases to simplify the procecess of traning, validation and testing to prevent further overfitting. Lastly, getting 1,714 images from 115 different patients means that data has a lot of natural variety, which stops the AI from just memorizing the patterns of a few individuals.

Compared to other recent datasets, our work fills a unique gap. For example, the NDB-UFES dataset (Falcão Ribeiro de Assis et al., 2023) has great patient information

but lacks healthy tissue samples and puts pre-cancer and actual cancer together. Another dataset, Multi-OSCC,(Guan et al., 2025) has a massive number of patients but focuses only on advanced cancer at very high magnification levels with other pathological lesions. Our dataset bridges the gap by providing a clean, easy-to-use setup that includes true healthy tissue and two practical magnifications levels for early screening specifically designed to differentiate between OSCC and normal tissue without any other lesion.

However, there are limitations to this work. First, all tissue slides were collected from a single institution and captured using the same microscope . This creates a "single-center bias," meaning an AI trained exclusively on these images might struggle with slides from other hospitals that use different staining techniques or camera systems. Second, to maintain a clean binary baseline, we deliberately excluded pre-cancerous conditions like dysplasia. While this makes the dataset ideal for initial screening tests, real-world clinical workflows are inherently more complex and frequently include these pre-cancerous cases.

4. Resource Availability

4.1 Summary statement

This publicly available dataset contains histopathological images of OSCC and normal oral tissue from an Egyptian population, addressing the scarcity of geographically diverse oral pathology datasets. It supports the development and evaluation of machine and deep learning models for AI-based oral cancer classification and contributes to improve representation of underserved populations in computational pathology research.

4.2 Data availability

Bahaa, Salma (2026), "Histopathological images of oral squamous cell carcinoma and normal tissue", Mendeley Data, V1, doi: 10.17632/7m9zkcx539.1.

4.3 Potential Use Cases

This dataset supports development and evaluation of machine learning and deep learning models for histopathological classification of OSCC and normal tissue. It enables automated cancer detection, computer-aided diagnosis, feature extraction, transfer learning, and benchmarking across magnifications. With annotation, it can also be used for segmentation tasks and other advanced computational analyses. Additionally, it supports digital pathology research and oral cancer screening systems, relevant to oral pathology, head and neck oncology, and computational pathology.

Misuse risks include unvalidated clinical deployment or use as a standalone diagnostic system without professional medical oversight.

4.4 Licensing

5. Methods

5.1 Data Details

This dataset was retrospectively collected between early 2023 and 2025.

Inclusion Criteria

Squamous cell carcinoma slides: Histopathological slides of OSCC exhibiting clear morphological features of malignancy and meeting acceptable quality standards were included.

Normal tissue slides: Histopathological slides of normal oral epithelium with high image quality and no evidence of dysplasia were included.

Exclusion criteria

Slides containing artifacts (e.g., staining errors, blur, tissue folds, or imaging defects) that could adversely affect the performance of machine learning models were excluded.

To highlight the scope of this study, relevant clinical and demographic information—including age, sex, and lesion anatomical location—was collected for each OSCC and control case (Radosavovic et al., 2020; Steiner et al., 2021) as summarized in Table 1.

5.2 Methods Used for the Data Creation

Serial 4–5 µm sections were prepared from paraffin-embedded tissue blocks of OSCC and normal tissues, and stained with hematoxylin and eosin (H&E). The samples were obtained from the archives of the Oral Pathology Department, Faculty of Dentistry, Alexandria University, Egypt. The dataset comprises 115 histopathological cases, including 49 cases of OSCC and 66 cases of normal tissue. The OSCC cases span three histological grades: 37 intermediate (75.5%), 9 high (18.4%), and 3 low (6.1%). Image acquisition was performed using an Olympus DP20 digital

Table 1: Distribution of OSCC and Normal samples by gender, age group, and anatomical site on image basis

Category	Subcategory	OSCC	Normal	P-value
Gender	Male	221	274	0.00022
	Female	203	170	
Age Group	Group 0 (< 40 years)	4	170	< 0.00001
	Group 1 (40–60 years)	225	270	
	Group 2 (> 60 years)	195	4	
Anatomical Site	Buccal / Posterior buccal mucosa	25	30	< 0.00001
	Mandibular alveolar ridge	83	94	
	Lateral surface of tongue	173	38	
	Ventral surface of tongue	40	0	
	Upper lip	12	0	
	Maxilla / Maxillary alveolar ridge	46	0	
	Hard palate / Palatal maxilla	36	0	
	Retromolar area	61	0	
	Zygoma / Condyle	23	0	
	Dorsal surface of tongue	0	17	
	Anterior gingiva	0	265	

camera mounted on an Olympus BX41-P polarizing light microscope (Olympus, USA), with CellSens Standard software. Images were captured at two magnification levels. At 400× magnification (40× objective), 427 OSCC images and 437 normal images were obtained, focusing on cellular-level malignant features, including hyperchromatism, pleomorphism, altered nuclear-to-cytoplasmic ratio, prominent nucleoli, and mitotic activity.Fig.1 **(A,B)**. At 100× magnification (10× objective), 421 OSCC images and 429 normal images were acquired, emphasizing tissue architecture such as keratin pearls, epithelial pearls, and tumor cell nests, along with malignancy-related features observable at lower magnification (Neville et al., 2015). Fig.1 **(C,D)**. All images were stored in JPEG format with a resolution of 1600 × 1200 pixels. The images were subsequently standardized to a resolution of 1100 × 1200 pixels. White mask normalization was then applied using a binary mask. Finally, color normalization was performed on the RGB channels prior to recombining them into the final image. These preprocessing steps were essential to ensure computational efficiency and maintain consistency across the dataset (Albalawi et al., 2024). Our study was designed to encompass multiple diagnostic fields in a Zigzag method (Sancheti et al., 2018), with all cases independently verified by three pathologists. Patient anonymity was maintained through the use of serial hospital identification numbers assigned to each case. Since the primary objective of this dataset was the classification of squamous cell carcinoma versus normal tissue, no manual annotation or segmentation labels were generated. The dataset was intended for image-level classification tasks rather than localization-based analyses. Data collection and preparation are illustrated in Fig.2

6. Validation

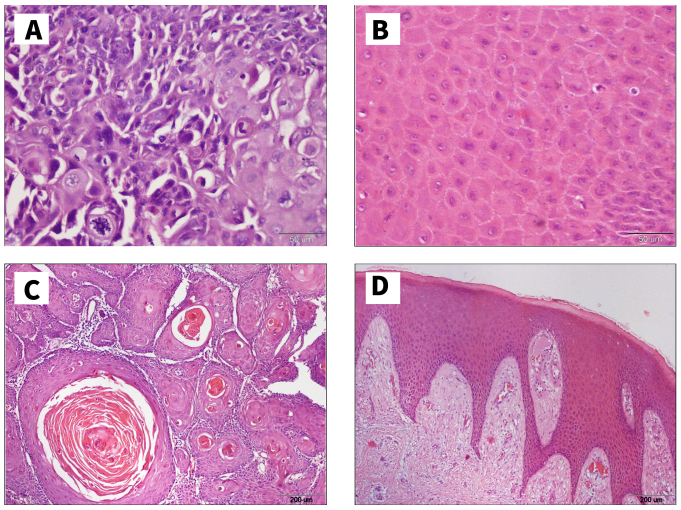

Figure 1: Histological images samples show **(A)** OSCC using objectives of 40 × **(B)** Normal epithelium using objectives of 40 × **(C)** OSCC using objectives of 10 × **(D)** Normal epithelium using objectives of 10 ×

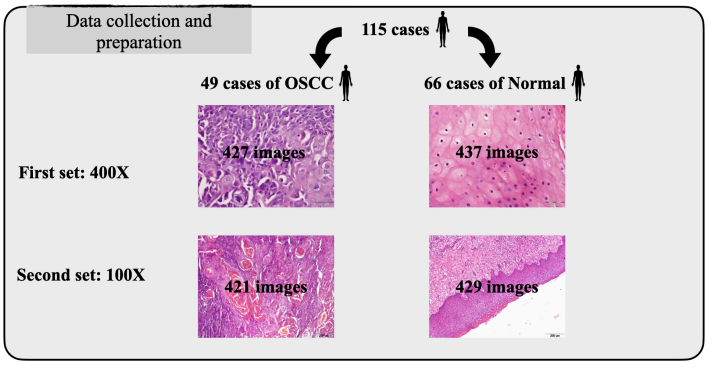

Figure 2: Data collection and preparation workflow

6.1 Description of approaches ensuring quality of data

Histopathological images were reviewed for image clarity, staining quality, focus consistency, and labeling accuracy prior to inclusion in the dataset. Images with severe artifacts, blur, incomplete tissue regions, or inadequate staining quality were excluded. Class labels (OSCC and normal tissue) were verified by three independent pathologists.

Quality control metrics included visual assessment of image sharpness, staining consistency, tissue integrity, following Zigzag method in capturing images to ensure non-repetition of different images (Sancheti et al., 2018)

This dataset was previously utilized in our earlier study (Bahaa et al., 2026) for the development and evaluation of several machine learning and deep learning models, including Fine Tree (FT), Medium Tree (MT), Linear Discriminant (LD), K-Nearest Neighbors (KNN), Linear Support Vector Machine (SVM), and a Neural Network model based on a multilayer perceptron (MLP) architecture. This is the graphical abstract for our work.Fig.3

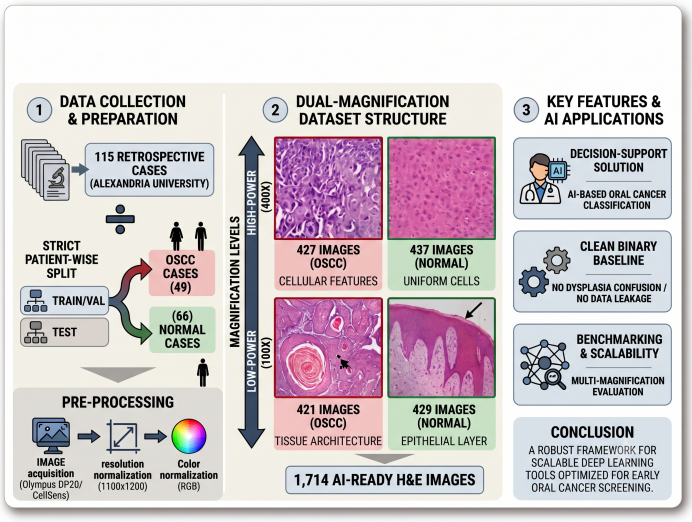

Figure 3: Graphical abstract

7. Acknowledgements

This study was self-funded by the authors, and no financial support or grants were received from any funding agency or institution.

8. Ethical Standards

Our study utilized retrospective tissue samples from the archive of Oral Pathology Department at Alexandria University. Informed consents for the use of Patients' data and material in research were obtained from all patients before their initial surgery, as part of the standard institutional procedure in Cranio Maxillo-facial and Plastic Surgery Department in Alexandria main hospital. Additionally, this retrospective study protocol was reviewed and approved by Research Ethics Committee, Faculty of Dentistry, Alexandria University, Institutional Review Board (IRB) under approval number (001056- IORG0008839). All procedures performed in our study involving human data and material were in accordance with the Declaration of Helsinki.

The dataset is considered non-sensitive, as it contains anonymized histopathological images without personally identifiable patient information.

9. Conflicts of interest

We declare we do not have conflicts of interest.

References

Eid Albalawi, Arastu Thakur, Mahesh Thyluru Ramakrishna, Surbhi Bhatia Khan, Suresh SankaraNarayanan, Badar Almarri, and Theyazn Hassn Hadi. Oral squamous cell carcinoma detection using efficientnet on histopathological images. *Frontiers in Medicine*, 10:1349336, 2024.

Salma Bahaa, Omneya Ramadan, Zeinab Elsayed Darwish, Eslam El-Fiky, and Mai M Saleh. Texture-based image feature analysis for the classification of oral squamous cell carcinoma using machine learning approach. *BMC Oral Health*, 26(1):105, 2026.

Maria Clara Falcão Ribeiro de Assis, Leandro Muniz de Lima, Liliana Aparecida Pimenta de Barros, Tânia Regina Velloso, Renato Krohling, and Danielle Camisasca. Ndb-ufes: An oral cancer and leukoplakia dataset composed of histopathological images and patient data, 2023.

J. Ferlay, M. Ervik, F. Lam, et al. Global cancer observatory: Cancer today (version 1.0), 2024. URL `https://gco.iarc.who.int/today`.

Dalí Freire, Paulo Faria, Adriano Loyola, Sergio Cardoso, Bruno Travencolo, and Marcelo do Nascimento. H&E-stained oral squamous cell carcinoma histological images dataset. Mendeley Data, V1, 2022. URL `https://doi.org/10.17632/9bsc36jyrt.1`.

Jinquan Guan, Junhong Guo, Qi Chen, Jian Chen, Yongkang Cai, Yilin He, Zhiquan Huang, Yan Wang, and Yutong Xie. A high magnifications histopathology image dataset for oral squamous cell carcinoma diagnosis and prognosis. `https://github.com/guanjinquan/OSCC-PathologyImageDataset`, 2025.

Mohamed Hammad, Mohamed Ghoneim, Waleed Shaker, Souham Meshoul, and Mohamed Arafa. Sinaiu-omp dataset: A curated histopathological image dataset for oral and maxillofacial pathology to support ai research, 2026.

W. Huang. Artificial intelligence and its application in early oral cancer screening: A systematic review. *Frontiers in Oncology*, 16:1789708, 2026. . URL `https://doi.org/10.3389/fonc.2026.1789708`.

Jiaxing Li. Fusion feature-based hybrid methods for diagnosing oral squamous cell carcinoma in histopathological images. *Frontiers in Oncology*, 15:1551876, 2025.

S. Li and T. Fevens. The detection of cancer cells in histopathology based on machine vision: A review. *Artificial Intelligence in Medicine*, 108:101931, 2020. . URL `https://doi.org/10.1016/j.artmed.2020.101931`.

Shrikanth Muralidharan, Monali Nikalje, Tulsi Subramaniam, Judy Annie Koshy, Ajit V Koshy, and Deepak Bangera. A narrative review on oral squamous cell carcinoma. *Journal of Pharmacy and Bioallied Sciences*, 17 (Suppl 1):S204–S206, 2025.

Brad W Neville, Douglas D Damm, Carl M Allen, and Angela C Chi. *Oral and maxillofacial pathology*. Elsevier Health Sciences, 2015.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020.

Tabassum Yesmin Rahman, Lipi B. Mahanta, Anup K. Das, and Jagannath D. Sarma. Histopathological imaging database for Oral Cancer analysis. Mendeley Data, V2, 2023. URL `https://doi.org/10.17632/ftmp4cvtmb.2`.

Sankalp Sancheti, Shivani Jain, Rashmi GS Phulari, Rajendrasinh Rathore, and Arpan K Shah. Quantitative assessment of tumor-associated tissue eosinophilia and mast cells in tumor proper and lymph nodes of oral squamous cell carcinoma. *Journal of Oral and Maxillofacial Pathology*, 22(1):145, 2018. . URL `https://doi.org/10.4103/jomfp.jomfp_24_17`.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Shintaro Sukegawa, Sawako Ono, Futa Tanaka, Yuta Inoue, Takeshi Hara, Kazumasa Yoshii, Keisuke Nakano, Kiyofumi Takabatake, Hotaka Kawai, Shimada Katsumitsu, et al. Effectiveness of deep learning classifiers in histopathological diagnosis of oral squamous cell carcinoma by pathologists. *Scientific reports*, 13(1):11676, 2023.

Zi-Ling Xu, Manbin Xu, Zhichen Sun, Qin Feng, Shaowei Xu, and Hanwei Peng. A nomogram for predicting overall survival in oral squamous cell carcinoma: a seer database and external validation study. *Frontiers in Oncology*, 15, 2025. .