Scientific Data

https://doi.org/10.1038/s41597-026-06954-5

Article in Press

A Smartphone-based Comprehensive Dataset of Annotated Oral Cavity Images for Enhanced Oral Disease Diagnosis

Received: 10 October 2025

Accepted: 23 February 2026

Cite this article as: Madan Kumar, P.D., Ranganathan, K., Lavanya, C. *et al.* A Smartphone-based Comprehensive Dataset of Annotated Oral Cavity Images for Enhanced Oral Disease Diagnosis. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-06954-5

P. D. Madan Kumar, K. Ranganathan, C. Lavanya, S. Rajeshwari, Anwesh Nayak, Ramesh Kestur, Raghuram Bharadwaj Diddigi & Sushree S. Behera

A Smartphone-based Comprehensive Dataset of Annotated Oral Cavity Images for Enhanced Oral Disease Diagnosis

P. D. Madan Kumar[1], K Ranganathan[1], C Lavanya[1], S. Rajeshwari[1], Anwesh Nayak[2], Ramesh Kestur[2], Raghuram Bharadwaj Diddigi[2], and Corresponding Author: Sushree Behera[2*]

[1]Ragas Dental College and Hospital Chennai, India, 600119

[2]International Institute of Information Technology Bangalore, India, 560100

[*]Corresponding Author: Sushree Behera (Sushree.behera@iiitb.ac.in)

ABSTRACT

This study introduces a SMARTphone-based, expert annotated dataset of Oral Mucosa images (SMART-OM), collected to facilitate the development of Artificial Intelligence and Machine Learning (AI/ML) technologies for automated diagnosis of Oral Cancer (OC) and Oral Potentially Malignant Disorders (OPMD). The dataset consists of 2,469 images from 331 subjects from four distinct classes: healthy/normal, variations from normal, OPMD, and OC. The images are captured using Android and iOS smartphone cameras under real-world clinical conditions in visible light. Each image is annotated by expert dental surgeons using the open-source VGG image annotator. Elaborate patient metadata, including clinical diagnosis, age, sex, and lifestyle-based risk indicators such as smoking, smokeless tobacco usage, alcohol consumption, and areca nut chewing, are recorded via a customized Jotform. The data collection and handling procedures are adhered to the ethical guidelines outlined in the Declaration of Helsinki and its amendments for research involving human subjects, with informed consent obtained from each subject. The SMART-OM dataset is intended to advance research and development of AI/ML algorithms for automated oral lesion detection.

INTRODUCTION

According to the International Agency for Research on Cancer (IARC), lip and oral cavity cancers accounted for 177,757 deaths, with 377,713 new cases worldwide in 2020, ranking as the 16th most common cancer globally[1]. The primary risk factors for Oral Cancer (OC) include being male, older age, tobacco use, excessive alcohol consumption, betel quid chewing, and lower socioeconomic status[2]. Successful management of OC depends on early diagnosis; however, low- and middle-income countries lack sufficient resources for screening and early detection, leading to delayed diagnosis[3]. Conventional oral examination relies heavily on visual inspection by clinicians, making the diagnostic process inherently subjective and prone to inter- and intra-observer variability. This issue is further compounded by the limited availability of trained oral pathology specialists, particularly in resource-constrained and geographically underserved settings[6]. The use of Artificial Intelligence (AI) and Machine Learning (ML) techniques on oral cavity images provides an interesting option for developing automated systems for screening and early detection of OC and pre-malignant conditions. The automated systems aim to improve oral health care access in resource constrained settings. Over the past two decades, most of the oral lesion detection frameworks based on ML and AI techniques have been developed using

histopathological images from publicly available datasets[5]. However, acquiring histopathological images requires specialized equipment that is expensive, limiting their applicability in community-based environments including primary healthcare centers. An emerging alternative to this expensive imaging setup involves leveraging visible light images, or RGB images, captured through smartphones without requiring any additional accessories[4]. This approach is particularly useful in regions, where the prevalence of OC is elevated and the availability of experts is low. However, a major challenge is the limited availability of open-source datasets containing oral cavity images in visible light to support model training and validation.

Several studies have developed ML and deep learning-based frameworks for OC diagnosis using large-scale oral datasets collected from multiple hospitals and web-based interfaces[7,9,10]. For instance, Fu et al.[10] collected 44,409 clinical oral images from 11 hospitals between April 12, 2006, and Nov 25, 2019. Song et al.[9] acquired 2,350 intraoral cheek mucosa images from three hospitals, captured using a smartphone-based intraoral screening device. However, these datasets are not publicly available due to several reasons, including institutional restrictions. A recent review[8] revealed that out of 332 OC image datasets, only one dataset is publicly accessible. This open-source dataset [18], compiled from various ENT hospitals, is limited to cancerous lesions found on the lips and tongue and contains only 131 images, out of which 87 are categorized as cancerous and 44 as non-cancerous. Hence, this dataset is not suitable for broader, generalized classification tasks[8]. Recently, with the advances in image collection and annotation methods, Piyarathne et al.[13] introduced a dataset of oral cavity images captured using smartphone cameras containing oral cavity images along with overlapping polygon-shaped annotations corresponding to lesion regions.

In this work, we introduce the SMART-OM dataset, a SMARTphone-based expert-annotated dataset of Oral Mucosa images, collected to facilitate the development of AI/ML technologies for automated diagnosis of oral disorders such as OPMD and OC. It contains 2,469 images from 331 subjects from four distinct classes: healthy/normal, variations from normal, OPMD, and OC. The images are captured using Android and iOS smartphone cameras under real-world clinical conditions in visible light. Each image is annotated by expert dental surgeons using the open-source VGG image annotator[19] (VIA) tool. Unlike the dataset introduced by Piyarathne et al.[13], the SMART-OM dataset contains region-based annotations as well as full annotations superimposed on the original images. The region-based annotations highlight the presence of anomalies such as lesions in images from the OC category, whereas the full annotation covers polygonal annotations corresponding to each object present in the oral cavity images. These comprehensive annotations enrich the dataset by offering precise spatial information, useful for the development of efficient AI models for disease diagnosis and classification. Additionally, the dataset contains detailed patient metadata, including clinical diagnosis, age, sex, and lifestyle-based risk indicators such as smoking, smokeless tobacco usage, alcohol consumption, and areca nut chewing, via a customized Jotform. The images and annotations are provided in JPEG and JavaScript Object Notation (JSON) formats, respectively, ensuring compatibility with contemporary ML and AI frameworks. The clinical data are descriptive and capture the detailed insight behind each diagnosis and are validated by experts. This comprehensive integration of

visual and clinical information enables the development of deep learning-based classification and detection algorithms as well as advanced vision-language models, contributing to the development of accurate and interpretable diagnostic tools.

METHODS

Ethical considerations

This study was conducted in accordance with the Declaration of Helsinki and its amendments and was approved by the Institutional Ethics Review Committee of Ragas Dental College and Hospital (IRB No. RIEC/20231021/PHD). Written, informed consent was obtained from all participants prior to clinical examination and intraoral image acquisition, including consent for the use and dissemination of anonymized images and associated metadata for research and public data sharing purposes. Only adult participants were enrolled in this process; therefore, parental or guardian consent was not required. The released dataset contains no direct or indirect identifiers; facial features were not photographed, and all metadata fields were fully anonymized prior to deposition.

Data collection

Participants for this study were recruited through two primary approaches. First, a door-to-door survey was conducted in Ranipet district in the state of Tamilnadu in India. This was carried out in collaboration with community health workers from Thirumalai Mission Hospital. Second, additional participants aged 18–85 years were enrolled from dental camps organized across various locations in Chennai over a six-month period, between May and October 2024. Individuals who declined to undergo clinical assessment were excluded from participation.

Method of recording data and taking intraoral images

Following enrollment, the participants were provided comfortable seating in a chair positioned against a wall for adequate head support, as the procedures were conducted in a community-based setting. To assess the risk of developing OC or OPMD, community health workers first collected basic demographic information. Personal and habit histories, including smoking, betel chewing, and alcohol consumption, were also recorded. Subsequently, a dental surgeon performed clinical examinations, following which the intraoral images were captured. The images were captured using two smartphones: a Samsung Galaxy M15 and an Apple iPhone14. The Samsung smartphone had a rear camera system comprising a 50 MP (f/1.8) main wide-angle camera, a 5 MP (f/2.2) ultra-wide-angle camera, and a 2 MP (f/2.4) macro camera. Prior to each session, the camera lens was cleaned with a microfiber cloth, and the resolution was set to the highest available level (aperture: f/20–f/32, ISO 100–200, shutter speed: 1/25–1/200, focus: autofocus). The iPhone's rear camera system had a 12 MP (f/1.5) main wide-angle camera and a 12 MP (f/2.4) ultra-wide-angle camera with a 120° field of view. The device employs autofocus with sensor-shift optical image stabilization, while exposure parameters such as aperture, ISO (typically ranging from 50 to 2000), and shutter speed (up to 1/8000 s) are automatically regulated by the in-built imaging system.

The area of interest was focused on the centre of the grid in the mobile phone screen, and the following eight intraoral images were acquired.

- Dorsal surface of tongue
- Ventral surface of tongue
- Right side of the buccal mucosa
- Left side of the buccal mucosa
- Labial surface of upper lip
- Labial surface of lower lip
- Maxillary or upper arch
- Mandibular or lower arch

For best results, image acquisition was primarily conducted in the presence of natural light with the camera positioned approximately 4–5 cm from the oral cavity. Additional light sources (such as a torch or flashlight) were employed as per the requirement. In addition to that, retraction aids, including mouth mirrors and wooden or plastic retractors, were used to ensure proper visualization. FIGURE 1 illustrates the process of capturing each region along with the corresponding sample images. The first two columns corresponding to each region depict the procedure of image capturing using smartphones. It is important to note that correct measures were taken to capture the images with minimal external intervention. To maintain subject anonymity, the original images were converted into sketch art presentations. The images were visually examined for quality based on the parameters and assessment criteria outlined in TABLE 2. Images that were out of focus or blurry were reacquired. When appropriate, a punch biopsy was performed to enable histological investigation before rendering a final diagnosis.

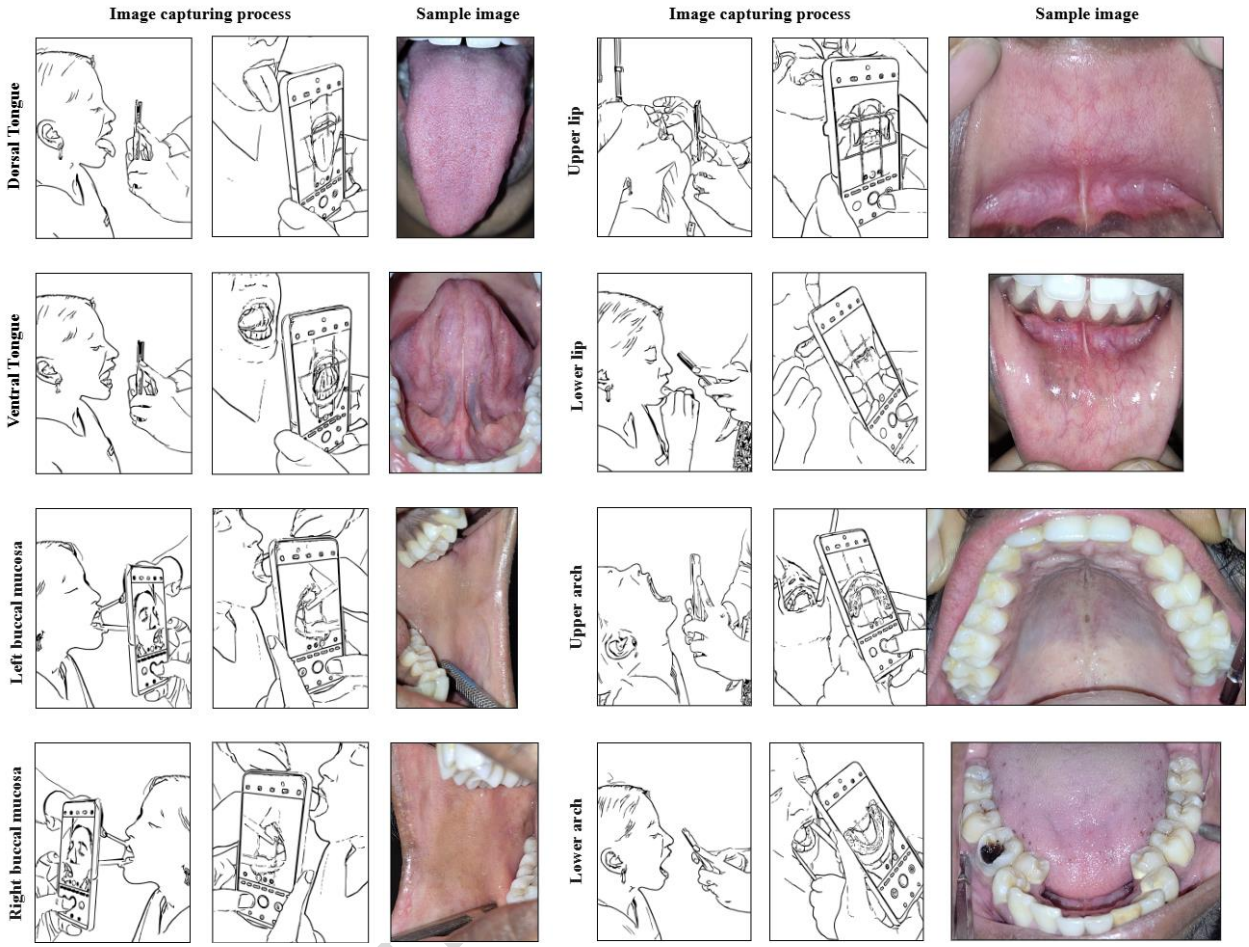

FIGURE 1: The process of image capturing for each intraoral region and the corresponding sample images

DATA RECORDS

The dataset has been uploaded to the Figshare repository[20] and is organized into four directories. Each directory corresponds to a specific class label of oral tissue pathology and contains associated image data as well as detailed annotations. The description of class labels is as follows:

- Normal – representing healthy oral tissues,
- Variation from Normal – comprising tissues that exhibit deviations from typical morphology but are not classified as pathological,
- OPMD – encompassing Oral Potentially Malignant Disorders, and
- Oral Cancer – comprising clinically and histologically diagnosed malignant oral lesions.

The OC category consisted of oral squamous cell carcinoma cases. The OPMD category included homogeneous leukoplakia, non-homogenous leukoplakia, erythroplakia, erythro-leukoplakia, oral submucous fibrosis, lichen planus, lichenoid lesions, proliferative verrucous leukoplakia, and speckled leukoplakia. Images that did not meet the criteria for either normal/healthy or OC/OPMD

were categorized as variations from normal, whereas intraoral images without distinct mucosal changes were considered normal/healthy. Each image was annotated by expert dental surgeons to demarcate the relevant intraoral regions corresponding to the assigned diagnostic class. FIGURE 2 shows sample images from each category along with the corresponding detailed annotations. In cases where multiple lesions were present within the same category, each lesion was annotated separately. However, if an image contained lesions from different categories, only the lesion consistent with the image's assigned category was annotated, while others were disregarded. For example, if an image was categorized as "homogeneous leukoplakia," only that lesion was annotated, even if other types of lesions were present. Each of the four class directory is further divided into four sub-directories based on the level of annotation available:

- Unannotated – contains raw images without any annotation.
- Region Annotation – includes images with annotations highlighting specific sites of interest.
- Full Annotation – comprises images with comprehensive annotations covering all intraoral soft and hard tissues.
- Lesion Annotation – focuses on lesions annotated within the images.

It can be noted that the Normal directory does not include the Lesion Annotation subdirectory, as normal tissues do not contain lesions. Within all subdirectories, including Unannotated, images are categorized based on eight anatomical sites or intraoral regions:

- Dorsal tongue
- Ventral tongue
- Left buccal mucosa
- Right buccal mucosa
- Upper lip
- Lower lip
- Upper arch
- Lower arch

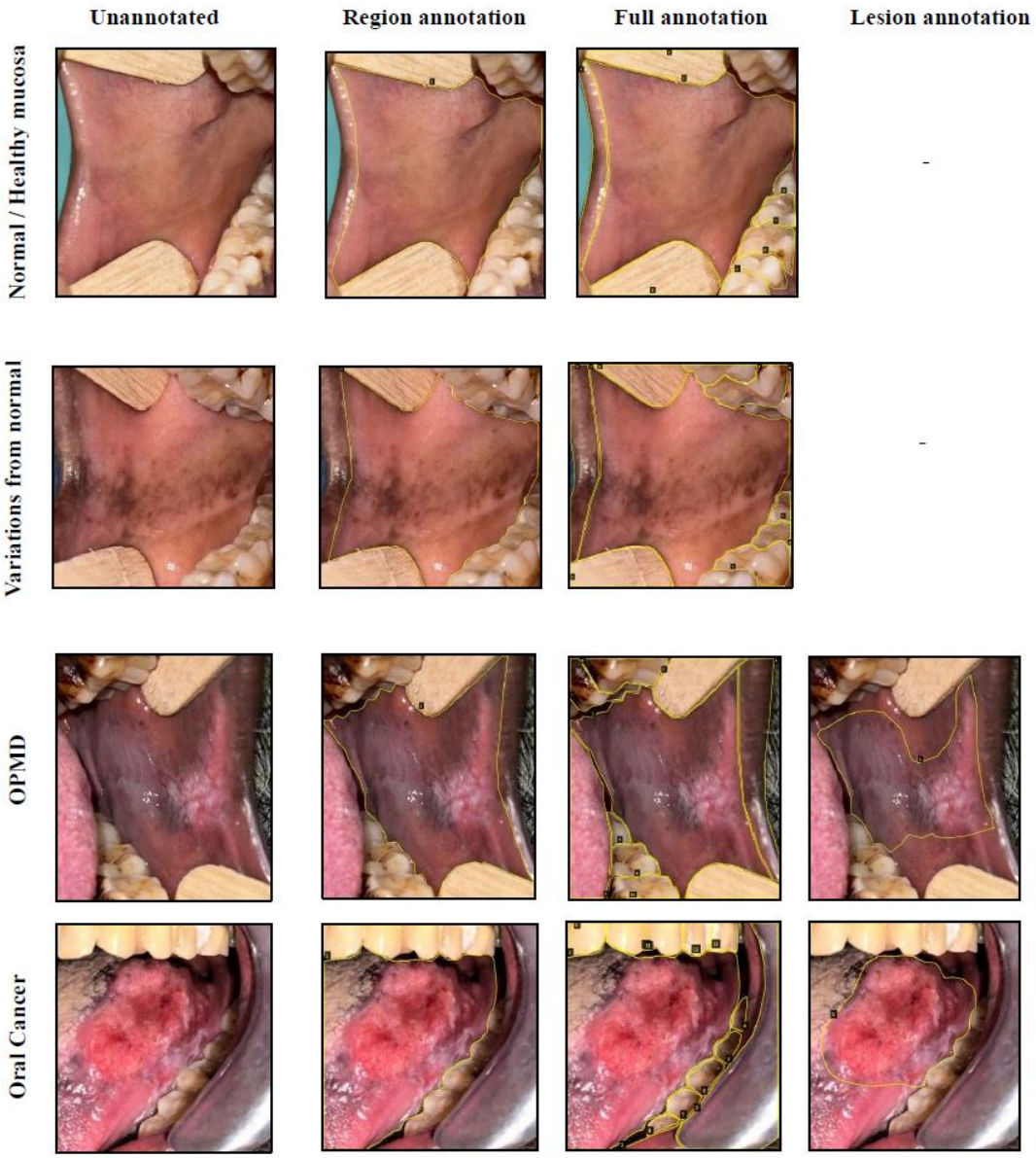

FIGURE 2: Sample images in each category along with the corresponding annotations

The intraoral images are archived in JPEG format, with filenames following the convention A_B_C.JPG. Each image was annotated using the open-source VGG Image Annotator (VIA) tool[19] (version 3.0.13) by two expert annotators. Once images were loaded from the server into the VIA interface, regions of interest (ROI) were delineated using the region shape panel, followed by boundary markings for additional areas of interest. The corresponding annotations were stored in JSON files with filenames structured as A_B_D.json, containing structured information outlining labeled regions, typically in the form of polygon-shaped annotations. In the filename structure "A" denotes a unique anonymized alphanumeric patient ID, "B" indicates the location where the images were captured (R for Ranipet and W for the World Vision camp), "C" specifies the intraoral

site using a two-letter code (DT: dorsal tongue, VT: ventral tongue, LB: left buccal mucosa, RB: right buccal mucosa, UL: upper lip, LL: lower lip, UA: upper arch, LA: lower arch), and "D" indicates the level of annotation (region, full, and lesion annotation).

To assess inter-annotator reliability, two annotators independently evaluated a subset of the dataset containing 100 intraoral images, classifying each as either "Normal" or "Abnormal." Agreement between annotators was quantified using Cohen's Kappa (κ), a statistical measure that accounts for chance agreement. Out of the 100 cases, the annotators concurred on 45 "Normal' and 50 "Abnormal" images, yielding 95 agreements in total. The resulting κ value was approximately 0.90, reflecting an excellent level of agreement. This high concordance underscores the reliability and consistency of the annotation process. Any disagreements in image categorization were resolved through discussion and, when required, with the input of a third and fourth reviewer. To ensure structured labelling and facilitate diverse use cases, the expert annotations were organized into three categories: lesion annotation, which focused solely on the lesion area (when present); full annotation, which encompassed all anatomical landmarks visible in the image; and regional annotation, which highlighted only the area of interest. Key features of the annotation files are as follows:

- Each JSON file encodes spatial coordinates that define the annotated regions.
- Annotations can be visualized as overlays on the corresponding images using the VIA tool (FIGURE 2).
- The JSON files are available as a separate download, enabling flexible dataset usage.

The Unannotated subdirectory does not contain JSON files, as it consists only of raw images. The Descriptors directory contains XLSX files corresponding to each annotation level. Each file includes separate Excel sheets for different tissue classes. Images are annotated using the VIA tool, where relevant regions are annotated using polygons, and each polygon is associated with a numeric identifier. The labels of these regions can be mapped using the appropriate Descriptor XLSX files. Within each sheet, the columns represent the following:

- S.No: Serial number of the entry
- File_name: Name of the annotated image file
- Label_ID: Expert labels that map to the corresponding polygonal annotation number in the image (1, 2, 3, ...).

The descriptors, along with the annotations, provide features that can be used to develop machine learning models for early diagnosis in oral pathology. In the Metadata directory, an Excel file that contains demographics and clinical data is provided. It comprises two sheets, 'Demographics & personal history' and 'Clinical findings.' Each row provides the data corresponding to one patient with SMITA_ID as the key. The objective of providing metadata along with image annotations is to facilitate the development of multi-modal learning frameworks for disease diagnosis and classification. The dataset has been uploaded to the Figshare repository and can be accessed using the link provided here[20].

Data overview

The SMART-OM dataset comprises 2,469 intraoral images collected from the eight intraoral regions. Out of the 331 subjects, 151 subjects presented with oral lesions, while the remaining 180 subjects did not exhibit any detectable lesions. TABLE 1 presents the distribution of image counts across the eight intraoral regions, stratified by the four diagnostic categories: normal/healthy, variations from normal, OPMD, and OC. As shown in FIGURE 3, 86.9% of the total images are Normal/Healthy, 7.2% depict Variations from Normal, 5.1% correspond to OPMD, and only 0.8% represent OC. Furthermore, FIGURE 4 presents the detailed distribution of age, gender and habits within the dataset. FIGURE 4(a) illustrates the gender-wise distribution of subjects across different age groups. Out of the 331 subjects included in the dataset, 177 (53.48%) were male and 154 (46.52%) were female. The age of the subjects ranged from early adolescence to late adulthood (approximately 18–85 years). The overall density curve demonstrated a near-normal bell-shaped distribution, and the Kernel density estimation (KDE) revealed that the highest concentration of cases observed in the 30–60-year age group. As shown in the gender-wise distribution of habits in FIGURE 4(b), tobacco use, both in the form of smoking and chewing, was the predominant habit and accounted for the largest subgroup. Arecanut use, although the least common habit, displayed a relatively higher representation among females compared to other habits. The age-wise distribution of habits (FIGURE 4(c)) further revealed that individuals reporting multiple habits were generally older, with an interquartile range spanning the fifth and sixth decades and outliers extending beyond 70 years. In contrast, the no-habit group exhibited broad and heterogeneous age distribution, with a median of ~45 years and outliers into the eighth decade, suggesting that individuals without risk behaviours were distributed across all age groups.

TABLE 1: Number of images in normal, variations from normal, oral potentially malignant disorders (OPMD), and oral cancer (OC) categories

Region	Number of images			
	Normal	Variations from Normal	Oral Potentially Malignant Disorders (OPMD)	Oral Cancer (OC)
Dorsal tongue	298	29	00	00
Ventral tongue	295	05	01	09
Left buccal mucosa	218	49	61	02
Right buccal mucosa	230	48	52	04
Upper lip	274	00	01	00
Lower lip	277	29	08	01
Upper arch	264	19	02	02
Lower arch	289	00	00	02
Total	2145	179	125	20
Grand Total	2469			

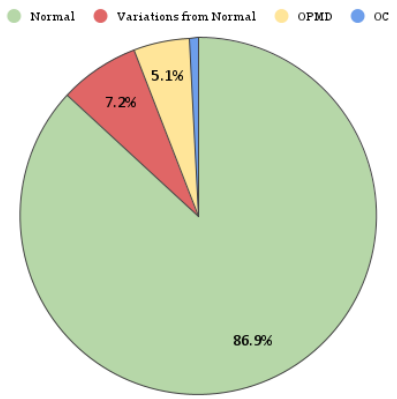

FIGURE 3: Percentage of images corresponding to each diagnostic category

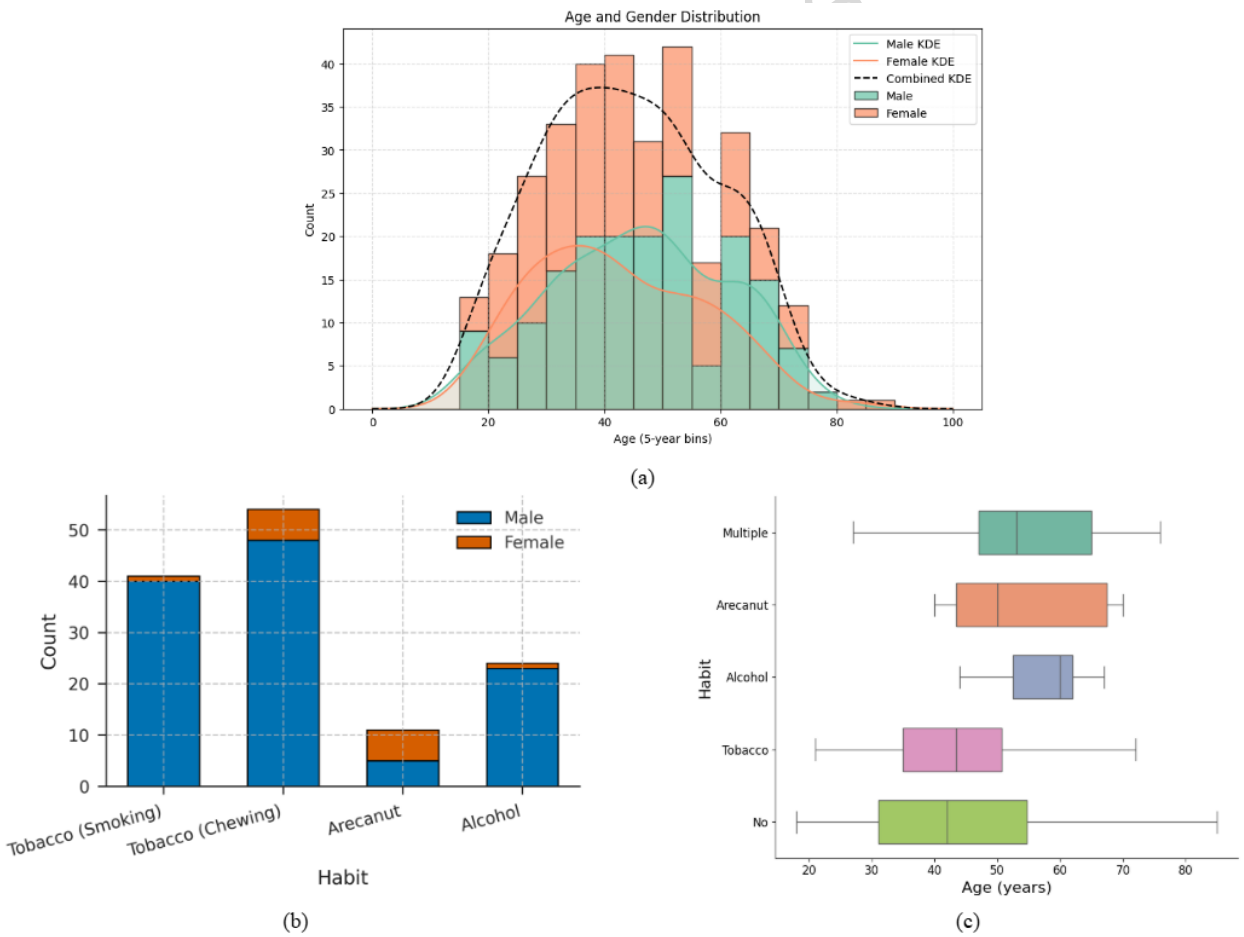

FIGURE 4: Distributions corresponding to subject metadata; (a) age and gender distribution of the subjects, (b) gender-wise distribution of habits and (c) habit-wise distribution of age.

A key distinguishing feature of the SMART-OM dataset lies in its site-specific categorization. Unlike conventional oral health datasets that rely on broad or generalized classifications, this resource is meticulously organized by individual intraoral regions. Such a detailed structure facilitates granular analysis of site-specific conditions, enhancing diagnostic precision and providing researchers with a valuable framework for targeted investigations. By bridging the gap left by less specific datasets, it substantially enriches the scope of oral health research. More importantly, the dataset excludes any personal identifiers, ensuring that all metadata and the images remain fully anonymized.

Furthermore, the dataset emphasizes the documentation of normal tissues and potentially malignant conditions, supported by a range of annotation types. Notably, the buccal mucosa regions contain a significant number of lesion-specific annotations, reflecting their clinical importance in the early detection of precancerous changes. However, the dataset shows limited representation of oral cancer cases, with fewer fully annotated or lesion-specific cancer images. This underrepresentation could constrain the development of highly robust AI models for malignant lesion detection. Nevertheless, given ongoing and future trends in medical image analysis toward data-efficient and foundation-model-driven learning[29,30,31,32], we believe this dataset remains highly relevant and useful for both current and emerging algorithms, particularly in realistic low-data clinical settings.

TECHNICAL VALIDATION

All intraoral images underwent systematic visual quality assessment using predefined parameters and evaluation criteria (TABLE 2). The selected parameters reflect established practices in clinical dental photography and digital imaging and were chosen to ensure diagnostic usability, reproducibility, and consistency across images acquired under heterogeneous conditions[22,23]. Composition and framing criteria were applied to ensure that diagnostically relevant regions occupied the majority of the image frame. Central positioning of the area of interest with more than 60% coverage minimizes background interference and enhances visual focus, while standardized cropping and angulation reduce geometric distortion and support reliable anatomical comparison across images[22,24]. Magnification and resolution requirements were defined to preserve fine structural detail critical for identifying subtle mucosal changes and early pathological features[24,25]. Optical zoom was permitted to maintain native spatial resolution, whereas digital zoom was excluded due to its tendency to introduce interpolation artifacts and degrade image fidelity[24]. Parameters related to patient and instrumentation artifacts were included to minimize confounding visual elements. Images exhibiting motion blur, extraneous objects, or unintended obstructions were excluded to ensure unobstructed visualization of anatomical structures. Hygiene-related criteria, including the absence of saliva, debris, and lens contamination, were applied to prevent reflectance artifacts and loss of contrast that could compromise visual assessment[22,23].

Consistency and standardization metrics were incorporated to support reproducibility and comparability. Images meeting these criteria could be reliably reproduced under similar

acquisition conditions and were suitable for longitudinal or cross-sectional comparison of the same anatomical regions. Technical specifications were selected to balance image quality with practical feasibility. A minimum resolution of 8–12 megapixels ensured sufficient spatial detail without excessive storage requirements, and accepted file formats (RAW, TIFF, or high-quality JPEG) preserved image integrity. Exposure and illumination ranges were defined to maintain uniform brightness and avoid information loss due to overexposure or underexposure under both daylight and low-light conditions[26,27]. Additional quantitative thresholds were applied to control contrast, color fidelity and noise. Standardized limits for highlights, shadows, white balance, vibrance, saturation, and hue deviation were used to maintain consistent color representation and prevent artificial enhancement[25]. Controlled ranges for texture, clarity, sharpening, and smoothness improved feature visibility while minimizing the introduction of synthetic artifacts. Noise levels were constrained by restricting ISO values to below 800 to preserve structural detail[22,28].

Image quality assessment was performed through a structured manual visual evaluation by trained investigators using the predefined criteria listed in TABLE 2. Each image was independently reviewed. Quantitative technical attributes such as resolution, ISO value, exposure parameters, and file format were verified using embedded image metadata extracted with ExifTool (v12.73). Only images satisfying all technical and qualitative criteria were retained in the final dataset, ensuring standardized image quality and suitability for downstream clinical evaluation and computational analysis.

TABLE 2: Criteria for image quality assessment

Parameter		**Assessment criteria**
Composition and Framing	Centering	Over 60% of the area is covered by the area of interest, which is concentrated in the centre grid.
	Cropping	● Minimal unnecessary background ● Include all necessary anatomical structures relevant to the clinical evaluation
	Angle and Perspective	Photo captured to prevent distortion at a constant, standard angle
Magnification and Resolution	Detail Visibility	Capture fine details, such as small lesions, cracks, or the texture of the gingiva
	Zoom	Preferred only optical zooming not the digital zooming
Patient and Instrumentation Artifacts	Minimization of Artifacts	Free from extraneous objects
	Patient Movement	Absence of motion blur
Hygiene and Cleanliness	Absence of Debris	Free of debris and saliva
	Surface Cleanliness	Camera lens free from dust
Consistency and Standardization	Reproducibility	Consistently reproducible
	Comparison Ability	Comparable to previous or subsequent images of the same area
Technical Specifications	Resolution	At least 8-12 megapixels
	File Format	RAW, TIFF, or high-quality JPEG
Exposure	Proper Brightness	Uniform lightening, avoiding over/under exposure
	Balanced Illumination	Bright daylight: -0.3 to -0.7, Low light: +0.3 to +0.7
Others	Contrast	1000:1
	Highlights	< 245

	Shadows	>15
	Whites	Color temperature: 5500k
	Blacks	> 10
	Vibrance	20-30%
	Saturation	Within ±10% from the original image
	Hue	< ±5 degrees on the color wheel
	Texture	+10 to +30
	Clarity	+10 to +40
	Sharpening	+30 to +70
	Noise	ISO < 800
	Smoothness	+10 to +30

To establish a baseline and assess the performance of state-of-the-art classification models, two distinct sets of evaluations were conducted on the SMART-OM dataset. The first set of experiments focused on abnormality detection, while the second set addressed disease classification. Both tasks are clinically significant, as abnormality detection aids in the early identification of atypical findings, whereas disease classification provides more specific diagnostic insights, thereby supporting accurate medical decision-making and treatment planning.

Experimental evaluations were carried out to perform abnormality detection from oral pathology images. For this task, the dataset was grouped into two categories: normal and abnormal. The normal class consisted exclusively of healthy oral images, while the abnormal class encompassed all images exhibiting deviations from normal tissues, including those categorized as variations from normal, OPMD, and OC. The task of disease classification was framed as a multi-class classification problem, in which each image was assigned to one of four categories: normal, variation from normal, OPMD, or OC. Six state-of-the-art deep learning models were employed in this study, namely, ResNet18, ResNet34, ResNet50, VGG16, EfficientNet-b0, and Vision Transformer (ViT)[17]. ResNet[14] is a deep CNN architecture that incorporates residual connections to mitigate the vanishing gradient problem during training. The ResNet variants, ResNet18, ResNet34, and ResNet50, differ primarily in network depth, with 18, 34, and 50 layers, respectively. VGGNet[15] is characterized by its deep and uniform structure, utilizing small convolutional filters for effective image classification; VGG16 represents a 16-layer variant of this family. EfficientNet[16] employs a compound scaling method that jointly optimizes depth, width, and resolution to achieve a balance between accuracy and efficiency, with EfficientNet-b0 serving as one of its lightweight variants. In contrast to these CNN-based models, ViT is a transformer-based architecture that treats images as sequences of patches and employs self-attention mechanisms to capture long-range dependencies and achieve competitive performance. For both abnormality detection and disease classification tasks, pre-trained versions of these models were fine-tuned on the SMART-OM dataset.

The dataset was split into training and inference sets in an 80:20 ratio, ensuring that 80% of the images from each class were randomly selected for training, while the remaining 20% were reserved for inference. Figures 5 and 6 present the class-wise distribution of images across different intraoral sites in both the training and test sets for binary and multi-class classifications,

respectively. During model training, hyperparameters were optimized using 5-fold cross-validation (k=5), wherein 20% of the training images were held out for validation at each epoch. Each fold was run for 100 epochs, with early stopping applied to prevent overfitting and ensure convergence. After identifying the optimal set of hyperparameters through grid search, the model was re-trained (fine-tuned) on the entire training set without a validation split. This approach was adopted because the number of images in each class was relatively small, and utilizing all available training data helped maximize the model's learning capacity. For binary classification, images were labeled as 0 (*Normal*) or 1 (*Abnormal*). For multi-class classification, images were assigned labels 0, 1, 2, and 3, corresponding to the classes: *Normal*, *Variation from normal*, *OPMD*, and *OC*, respectively. One-hot encoding of labels was employed as a preprocessing step prior to training and inference. Binary cross-entropy loss was used for binary classification, while categorical cross-entropy loss was applied for multi-class classification. As shown in figures 5 and 6, the distribution is skewed toward the Normal class, highlighting the significant imbalance in the dataset. To overcome this, we have used weighted cross-entropy loss functions instead for both binary and multi-class classifications.

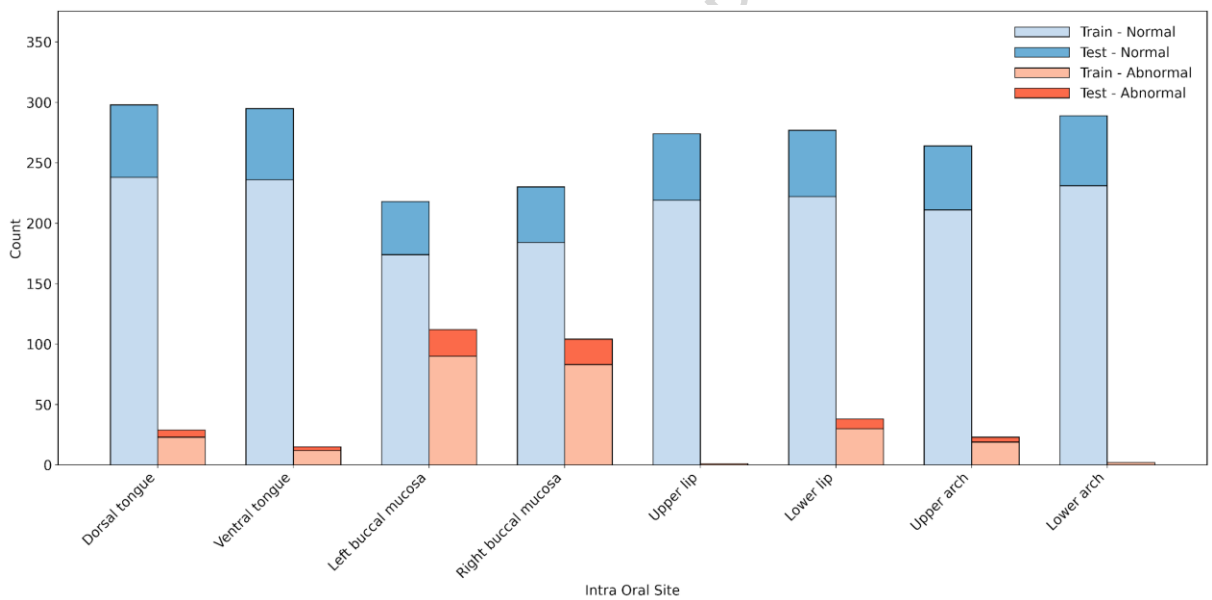

FIGURE 5: Class-wise train and test distribution for binary classification

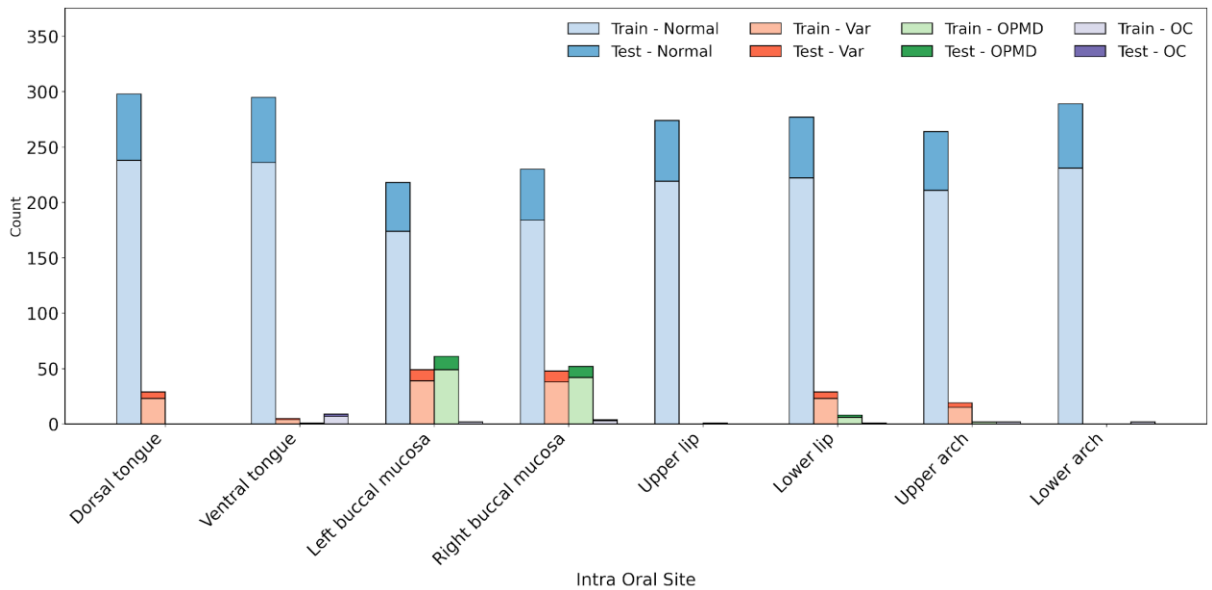

FIGURE 6: Class-wise train and test distribution for multi-class classification

The performance of the models was evaluated using standard classification metrics, including accuracy, precision, sensitivity, specificity, and F1-score. To compute these metrics, a confusion matrix was constructed, summarizing the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). In our experiments, the abnormal class was treated as the positive class, while the normal class was treated as the negative class. Accordingly, a test image classified as FP corresponds to a normal image incorrectly predicted as abnormal, whereas an FN corresponds to an abnormal image incorrectly predicted as normal. Classification accuracy represents the proportion of correctly classified samples across both normal and abnormal categories, reflecting the model's overall ability to distinguish between the two classes. Precision quantifies the reliability of positive predictions by measuring the proportion of correctly identified abnormal samples among all samples predicted as abnormal. Sensitivity (or recall) is defined as the proportion of correctly classified abnormal images out of all actual abnormal images in the test data, reflecting the model's ability to detect abnormalities. Specificity, on the other hand, measures the proportion of correctly classified normal images out of all actual normal images in the test data. In the context of medical screening for oral disorders, both sensitivity and specificity are critical indicators of performance. However, sensitivity is particularly important, as failing to detect an abnormality may result in serious consequences for the patient. Mathematically, these performance metrics are expressed as follows:

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN} \times 100\right)\%$$

$$Precision = \left(\frac{TP}{TP + FP} \times 100\right)\%$$

$$Sensitivity = \left(\frac{TP}{TP + FN} \times 100\right)\%$$

$$Specificity = \left(\frac{TN}{TN + FP} \times 100\right)\%$$

$$F1 - Score = \left(2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}\right)\%$$

TABLE 3 and TABLE 4 present the results of binary and multi-class classification tasks across all six models. To address the class imbalance during inference, both micro- and macro-averaged metrics were computed to evaluate the model performance. Additionally, the class-specific metrics corresponding to all models were reported. In both tables, the numbers in **Bold** indicate the best performance achieved in each case. As reported in TABLE 3, for the binary classification task, the ResNet18 model demonstrates superior performance across most metrics, while ViT, ResNet34, and ResNet50 achieve the highest performance in some specific cases. As observed, the highest test accuracy achieved in this case is 90.89%. Considering macro averaging, the ResNet18 model achieves a precision of 79.74%, a recall (or sensitivity) of 80.14%, a specificity of 80.14%, and an F1-score of 79.94%. Using micro averaging, the model achieves equal precision, sensitivity, specificity, and F1-score of 90.89%. This superior classification performance of ResNet18 is likely attributed to its effective residual connections, which enable the model to capture rich discriminative information. In contrast, the lighter EfficientNet-b0 model exhibits comparatively inferior recognition performance. This can be due to the inherent trade-off between computational efficiency and representational capacity, although EfficientNet-b0 offers reduced complexity, its limited discriminative power results in noticeably lower classification performance.

TABLE 3: Binary classification performance on the SMART-OM dataset

Performance Measures		**Models**					
		Resnet18	**ResNet34**	**ResNet50**	**EfficientNet-b0**	**VGG16**	**ViT**
Accuracy (%)		**90.89**	88.66	90.48	89.43	89.27	84.82
Macro Average	**Precision (%)**	**79.74**	75.00	79.54	76.66	76.39	69.35
	Sensitivity (%)	**80.14**	78.86	75.91	76.66	73.89	76.65
	Specificity (%)	**80.14**	78.86	75.91	76.66	73.89	76.65
	F1-Score (%)	**79.94**	76.70	77.55	76.66	75.04	71.89
Micro Average	**Precision (%)**	**90.89**	88.66	90.48	89.47	89.27	84.82
	Sensitivity (%)	**90.89**	88.66	90.48	89.47	89.27	84.82
	Specificity (%)	**90.89**	88.66	90.48	89.47	89.27	84.82
	F1-Score (%)	**90.89**	88.66	90.48	89.47	89.27	84.82
Class-wise Metrics (w.r.t. *Normal*)	**Precision (%)**	**94.87**	94.74	93.62	93.95	93.13	94.49
	Sensitivity (%)	**94.65**	92.09	95.58	93.95	94.65	87.67
	Specificity (%)	**65.62**	**65.62**	56.26	59.37	53.12	**65.62**
	F1-Score (%)	**94.76**	93.40	94.59	93.95	93.89	90.95
Class-wise Metrics	**Precision (%)**	64.61	55.26	**65.45**	59.37	59.65	44.21
	Sensitivity (%)	**65.62**	**65.62**	56.25	59.37	53.12	**65.62**

(w.r.t. *Abnormal*)	**Specificity (%)**	94.65	92.09	**95.58**	93.95	94.65	87.67
	F1-Score (%)	**65.12**	60.00	60.50	59.37	56.20	52.83
AUC		**0.91**	0.90	0.90	0.90	0.87	0.88

Interestingly, ResNet18 also achieves the best class-wise metrics for the *Normal* class among all convolutional and transformer-based models, with a precision of 94.87%, sensitivity of 94.65%, specificity of 65.62%, and F1-score of 94.76%, highlighting its effectiveness on the relatively small SMART-OM dataset. More importantly, the relatively lower architectural complexity of ResNet18 compared to deeper models like ResNet34 and ResNet50 likely helps mitigate overfitting, thereby improving the generalization performance on the test set. In contrast, when examining class-specific metrics for the *Abnormal* class, the ResNet50 model achieves the highest precision and specificity. Nevertheless, ResNet18 attains the best sensitivity and F1-score. The top performance metrics for the abnormal class include precision of 65.45%, sensitivity of 65.62%, specificity of 95.58%, and F1-score of 65.12%. The combination of high precision and sensitivity for the *Normal* class, alongside the high specificity for the *Abnormal* class, suggests that the model performs better at accurately identifying healthy cases than abnormal ones. This behavior can be directly attributed to the significant class imbalance between the normal and abnormal classes, as clearly depicted in the training and test distributions in FIGURE 5.

We observe a similar trend in the classification metrics for the multi-class case as shown in TABLE 4. The ResNet18 model outperforms other architectures across most performance metrics, including both macro- and micro-averaged measures. In terms of class-specific metrics, ResNet50 achieves superior performance in several cases, while ViT performs best for a few specific metrics. The highest test accuracy achieved for multi-class classification is 89.07%. The best macro-averaged performance metrics include a precision of 73.43%, sensitivity of 53.43%, specificity of 89.03%, and F1-score of 59.02%. Both ResNet18 and ResNet50 achieve the top precision, sensitivity, and F1-score of 89.07%, along with a specificity of 96.36% while using micro averaging.

TABLE 4: Multi-class classification performance on the SMART-OM dataset

Performance Measures		**Models**					
		ResNet18	**ResNet34**	**ResNet50**	**EfficientNet-b0**	**VGG16**	**ViT**
Accuracy (%)		**89.07**	88.26	89.07	86.23	85.42	80.36
Macro Average	**Precision (%)**	**73.66**	58.02	74.62	66.82	40.67	54.56
	Sensitivity (%)	**53.43**	52.58	61.57	49.42	36.72	62.30
	Specificity (%)	86.98	87.44	84.63	84.88	81.29	**89.03**
	F1-Score (%)	**59.07**	54.61	65.99	53.84	38.27	56.95
	AUC	**0.92**	0.91	0.91	0.90	0.77	0.90
Micro Average	**Precision (%)**	**89.07**	88.26	**89.07**	86.23	85.42	80.36
	Sensitivity (%)	**89.07**	88.26	**89.07**	86.23	85.42	80.36
	Specificity (%)	**96.36**	96.09	**96.36**	95.41	95.14	93.45

	F1-Score (%)	**89.07**	88.26	**89.07**	86.23	85.42	80.36
	AUC	**0.97**	**0.97**	**0.97**	**0.97**	0.96	0.96
Class-wise Metrics (w.r.t. *Normal*)	**Precision (%)**	93.26	93.64	91.83	92.25	90.24	**95.56**
	Sensitivity (%)	96.51	95.81	**96.74**	94.19	94.65	85.12
	Specificity (%)	53.12	56.25	42.19	46.87	31.25	**73.44**
	F1-Score (%)	**94.86**	94.71	94.22	93.21	92.39	90.04
Class-wise Metrics (w.r.t. *Variation from normal*)	**Precision (%)**	**72.31**	38.46	40.00	31.03	28.00	23.53
	Sensitivity (%)	29.73	27.03	16.22	24.32	18.92	**43.24**
	Specificity (%)	**96.72**	96.50	98.03	95.62	96.06	88.62
	F1-Score (%)	**34.92**	31.75	23.08	27.27	22.58	30.48
Class-wise Metrics (w.r.t. *OPMD*)	**Precision (%)**	59.09	50.00	**66.67**	44.00	44.44	32.50
	Sensitivity (%)	54.17	54.17	**66.67**	45.83	33.33	54.17
	Specificity (%)	98.08	97.23	**98.30**	97.02	97.87	94.25
	F1-Score (%)	56.52	52.00	**66.67**	44.90	38.09	40.62
Class-wise Metrics (w.r.t. *OC*)	**Precision (%)**	**100.00**	50.00	**100.00**	**100.00**	0.00	66.67
	Sensitivity (%)	33.33	33.33	**66.67**	33.33	0.00	**66.67**
	Specificity (%)	**100.00**	99.80	**100.00**	**100.00**	**100.00**	99.80
	F1-Score (%)	50.00	40.00	**80.00**	50.00	0.00	66.67

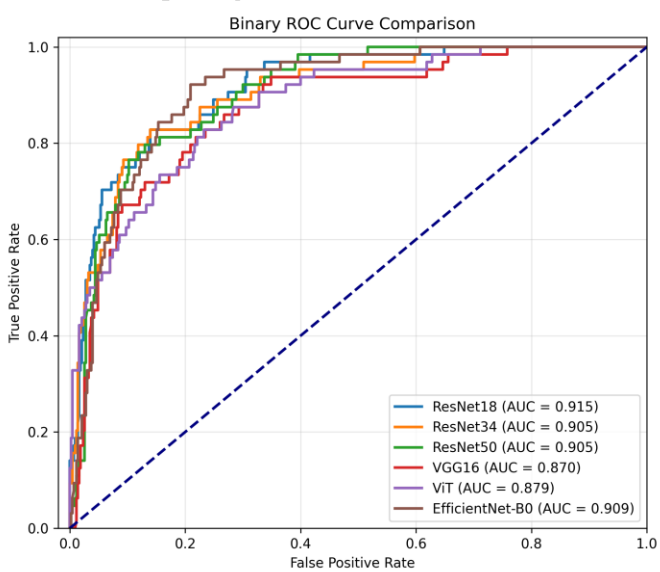

FIGURE 7: Receiver operating characteristics (ROC) curve corresponding to binary classification

Considering the class-wise metrics, the *Normal* class achieves the highest performance with a precision of 95.56%, sensitivity of 96.74%, and F1-score of 94.86%, although the specificity

drops to 73.44%. The combination of high precision and sensitivity with relatively lower specificity indicates that the models are better at identifying TP than TN, meaning they classify *Normal* cases more accurately than the other three categories. For the *Variation from Normal* class, the performance exhibits an opposite pattern. The model achieves a specificity of 96.72%, while the precision, sensitivity, and F1-score are substantially lower, at 72.31%, 43.24%, and 34.92%, respectively. This reduced performance is primarily due to the low TP for this class, as the images in this category closely resemble those of the *Normal* class, often leading to misclassification. The limited number of samples in the *OPMD* and *OC* classes further exacerbates this imbalance, which is reflected in the inconsistent and often random nature of their performance metrics. Specifically, for the *OPMD* class, the ResNet50 model achieves a precision of 66.67%, sensitivity of 66.67%, specificity of 98.30%, and F1-score of 66.67%. Similarly, for the *OC* class, the model attains precision of 100.00%, sensitivity of 66.67%, specificity of 100.00%, and F1-score of 80.00%. It is important to note that these results are based on five-fold cross-validation, and the reported test set metrics correspond to the best-performing validation parameters. Consequently, the performance observed for the *OPMD* and *OC* classes is likely influenced by the limited sample size and inherent class imbalance. However, continued expansion of the dataset with more images across all diagnostic categories is anticipated to further improve model generalization, enhance robustness, and boost classification performance, particularly for the underrepresented OPMD and OC classes.

Figures 7 and 8 show the receiver operating characteristics (ROC) curves for all six models corresponding to the binary and multi-class classification tasks, respectively. As observed from FIGURE 7 as well as TABLE 3, the ResNet18 model achieves the highest area under the ROC curve (AUC) of 0.91 for binary classification, indicating strong discriminative performance between normal and abnormal classes. For the multi-class classification task, FIGURE 8 illustrates the class-specific ROC curves for all models. Among them, the ResNet34 model achieves the highest AUC of 0.91 for the *Normal* class and 0.86 for the *Variation from normal* class. For the *OPMD* and *OC* classes, the ResNet18 model achieves the best performance, with AUC values of 0.95 and 1.00, respectively. Additionally, as reported in TABLE 4, the macro-averaged AUC reaches 0.92, while the micro-averaged AUC is 0.97, further demonstrating the overall effectiveness of the proposed models. These results are comparable to state-of-the-art performance observed in deep learning-based medical image analysis tasks, highlighting not only the feasibility but also the strong potential of deep neural networks for automated analysis of oral diseases using intraoral images.

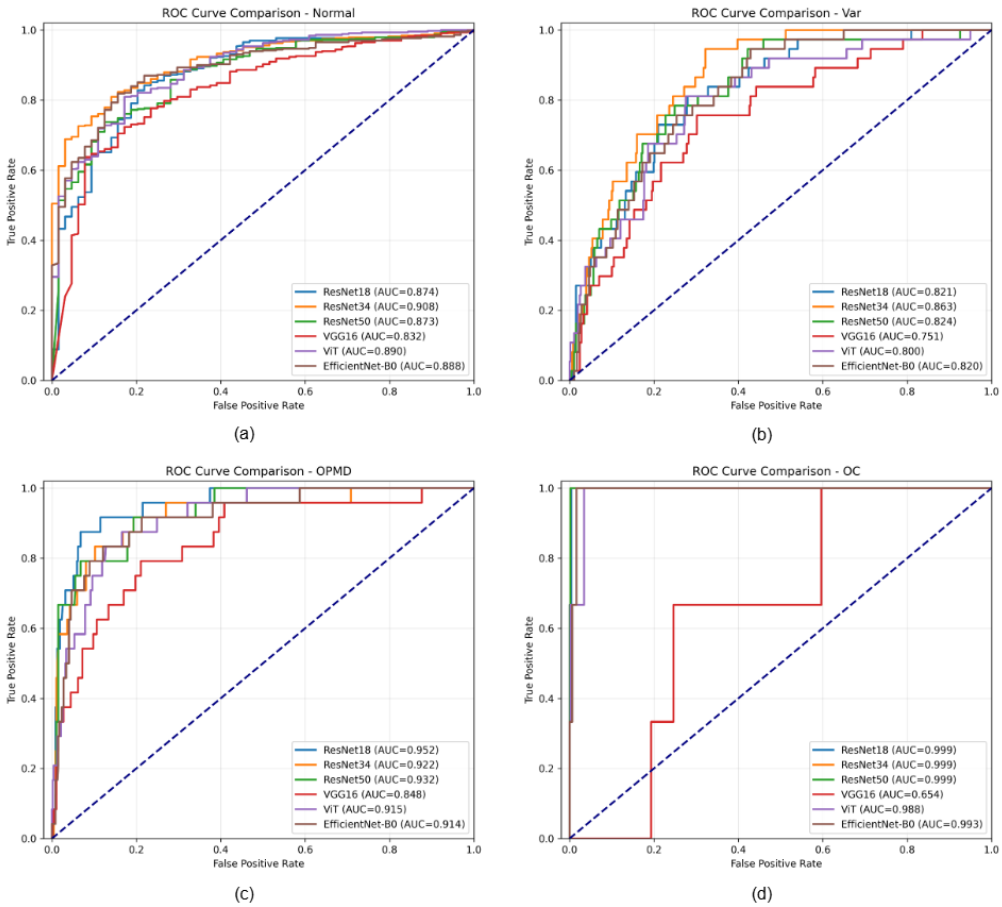

FIGURE 8: Class-wise receiver operating characteristics (ROC) curves corresponding to (a) Normal, (b) Variation from normal, (c) OPMD, and (d) OC classes.

In this study, we introduced SMART-OM, a newly curated intraoral image dataset designed to advance automated analysis of oral health conditions, including variations from normal, oral potentially malignant disorders (OPMD), and oral cancer (OC). The dataset is unique in its site-specific organization, detailed annotations, and inclusion of comprehensive patient metadata, which together enhance its utility for developing and evaluating AI-driven diagnostic models. We further performed an extensive technical validation by benchmarking several state-of-the-art deep learning models for both abnormality detection and disease classification tasks, demonstrating that these models can achieve performance levels comparable to existing state-of-the-art results in medical image analysis. The dataset presently contains a predominance of healthy images and relatively few OPMD and OC cases, which reflects community-level prevalence and results in class imbalance, thereby limiting statistical generalizability. However, it is important to contextualise this dataset within the rapidly evolving landscape of medical image analysis such as few-shot learning techniques. The dataset acquisition is a work-in-progress, and future efforts will focus on increasing the number of samples across all diagnostic categories, incorporating additional patient data, and extending coverage to a wider range of intraoral regions to improve its diversity as well as representativeness. Future extensions of the dataset may also enable

stratification of the OC class into labels corresponding to different stages of lesion development. With these advancements, we expect the SMART-OM dataset would serve as a valuable resource for the development of robust, scalable, and clinically applicable tools for early detection and diagnosis of oral diseases.

CODE AVAILABILITY

The GitHub repository containing the codes for technical analysis, model training and inference, as well as hyperparameter tuning, can be accessed at https://github.com/Anwesh2000/SMART_OM_Dataset_Technical_Validation[21].

DATA AVAILABILITY

The SMART-OM dataset has been deposited in the Figshare repository and can be accessed here[20].

REFERENCES

1. Rai, P., Ng, A., Intekhab, I., Sim, Y. F., Lai, C. W. M., & Loh, J. P., Oral Cancer in Asia-a systematic review. Advances in Oral and Maxillofacial Surgery, 8, 100366, (2022).
2. Sankaranarayanan, R., Ramadas, K., Amarasinghe, H., Subramanian, S., & Johnson, N., Oral cancer: prevention, early detection, and treatment. Cancer: disease control priorities. 3rd ed. Washington, DC: The International Bank for Reconstruction and Development/The World Bank, 3, 85-99 (2015).
3. Jain, A. K., Oral cancer screening: insights into epidemiology, risk factors, and screening programs for improved early detection. Cancer Screening and Prevention, 3(2), 97-105 (2024).
4. Talwar, V., Singh, P., Mukhia, N., Shetty, A., Birur, P., Desai, K. M., ... & Vinod, P. K., AI-assisted screening of oral potentially malignant disorders using smartphone-based photographic images. Cancers, 15(16), 4120 (2023).
5. Chaudhary, N., Rai, A., Rao, A. M., Faizan, M. I., Augustine, J., Chaurasia, A., ... & Ahmad, T., High-resolution AI image dataset for diagnosing oral submucous fibrosis and squamous cell carcinoma. Scientific Data, 11(1), 1050 (2024).
6. Mira, E. S., Saaduddin Sapri, A. M., Aljehanı, R. F., Jambı, B. S., Bashir, T., El-Kenawy, E. S. M., & Saber, M., Early diagnosis of oral cancer using image processing and Artificial intelligence. Fusion: Practice & Applications, 14(1) (2024).
7. Dixit, S., Kumar, A., & Srinivasan, K., A current review of machine learning and deep learning models in oral cancer diagnosis: recent technologies, open challenges, and future research directions. Diagnostics, 13(7), 1353 (2023).
8. Sengupta, N., Sarode, S. C., Sarode, G. S., & Ghone, U., Scarcity of publicly available oral cancer image datasets for machine learning research. Oral Oncology, 126, 105737 (2022).
9. Song, B., Sunny, S., Li, S., Gurushanth, K., Mendonca, P., Mukhia, N., ... & Liang, R., Bayesian deep learning for reliable oral cancer image classification. Biomedical Optics Express, 12(10), 6422-6430 (2021).

10. Fu, Q., Chen, Y., Li, Z., Jing, Q., Hu, C., Liu, H., ... & Xiong, X., A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study. EClinicalMedicine, 27 (2020).
11. Di Fede, O., Panzarella, V., Buttacavoli, F., La Mantia, G., & Campisi, G., Doctoral: A smartphone-based decision support tool for the early detection of oral potentially malignant disorders. Digital Health, 9, 20552076231177141 (2023).
12. Rajendran, S., Lim, J. H., Yogalingam, K., Kallarakkal, T. G., Zain, R. B., Jayasinghe, R. D., ... & Cheong, S. C., Image collection and annotation platforms to establish a multi-source database of oral lesions. Oral Diseases, 29(5), 2230-2238 (2023).
13. Piyarathne, N. S., Liyanage, S. N., Rasnayaka, R. M. S. G. K., Hettiarachchi, P. V. K. S., Devindi, G. A. I., Francis, F. B. A. H., ... & Jayasinghe, R. D., A comprehensive dataset of annotated oral cavity images for diagnosis of oral cancer and oral potentially malignant disorders. Oral Oncology, 156, 106946 (2024).
14. He, K., Zhang, X., Ren, S., & Sun, J., Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778 (2016).
15. Simonyan, K., & Zisserman, A., Very deep convolutional networks for large-scale image recognition. Preprint at arXiv:1409.1556 (2014).
16. Tan, M., & Le, Q., Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, 6105-6114. PMLR (2019).
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at arXiv:2010.11929 (2020).
18. S. Barot. Oral cancer (lips and tongue) images. https://www.kaggle.com/datasets/shivam17299/oral-cancer-lips-and-tongue-images (2020)
19. Dutta, A., & Zisserman, A., The VIA annotation software for images, audio and video. In Proceedings of the 27th ACM international conference on multimedia. 2276-2279 (2019).
20. P D, Dr Madan Kumar; K, Dr Ranganathan; S, Dr Rajeshwari; Pavithra, Dr. M; Lavanya C, Dr; Nayak, Anwesh; et al. (2026). SMART-OM: A SMARTphone based expert annotated dataset of Oral Mucosa images. figshare. Dataset. https://doi.org/10.6084/m9.figshare.31341790.v1
21. SMART-OM Dataset Technical Validation, https://github.com/Anwesh2000/SMART_OM_Dataset_Technical_Validation Accessed on 08-10-2025
22. Ito FA, Fernandes DT, Fernandes CI, Scarini JF, Soares LE, de Pauli Paglioni M, Carrard VC, Standardization in Oral Photography. In Clinical Decision-Making in Oral Medicine: A Concise Guide to Diagnosis and Treatment Jun 8. pp. 11-16. Cham: Springer International Publishing (2023).

23. Lin, I., Datta, M., Laronde, D. M., Rosin, M. P., & Chan, B., Intraoral photography recommendations for remote risk assessment and monitoring of oral mucosal lesions. international dental journal, 71(5), 384-389 (2021).
24. Casaglia, A., De Dominicis, P., Arcuri, L., Gargari, M., & Ottria, L., Dental photography today. Part 1: basic concepts. ORAL & implantology, 8(4), 122 (2016).
25. Momin, S., Sarfaraz, H., Sengottuvel, K., Shetty, R., Amreen, S., & Shetty, M., Comparison of image quality, color accuracy, and resolution in intraoral photography using digital single lens reflex camera and smartphone cameras: A pilot study. Journal of Dental Sciences. 2025 Oct 7: 01-09 (2025).
26. Piemonte, E. D., Gilligan, G. M., Costa, M. F. G., & Lazos, J. P., How to improve photographs with smartphones for oral telemedicine. Exploration of Digital Health Technologies, 2(5), 249-258 (2024).
27. Shahrul, A. I., Shukor, N., & Norman, N. H., Technique for orthodontic clinical photographs using a smartphone. International Journal of Dentistry, 2022(1), 2811684 (2022).
28. Ferreira, C. D. A. P., Pereira, F. D. A. V., Rodrigues, M. V. B., Cabral, J. L. D. O. A., & de Campos Tuña, I. T., Art and science of dental photography: suggested photographic protocol with cellular device. Observatório De La Economía Latinoamericana, 22(8), e6167-e6167 (2024).
29. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I., Learning transferable visual models from natural language supervision. In International conference on machine learning. 8748-8763. PmLR (2021).
30. Wiggins, W. F., & Tejani, A. S., On the opportunities and risks of foundation models for natural language processing in radiology. Radiology: Artificial Intelligence, 4(4), e220119 (2022).
31. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R., Segment anything. In Proceedings of the IEEE/CVF international conference on computer vision. 4015-4026 (2023).
32. Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., & Merhof, D., Foundational models in medical imaging: A comprehensive survey and future vision. arXiv preprint arXiv:2310.18689 (2023).

AUTHOR CONTRIBUTIONS

P.D.M.K., K.R., C.L., and S.R. contributed to data acquisition, data interpretation, image annotation, drafting, expert validation, and revision of the manuscript. A.N. performed technical validation and was responsible for model development, training, and evaluation of the deep learning models. R.K., R.B.D., and S.B. contributed to the conceptualization of the technical framework and study design, provided overall mentoring, and were involved in drafting and critical revision of the manuscript. All authors reviewed and approved the final manuscript.

ACKNOWLEDGEMENT

The authors acknowledge the use of AI-assisted tools to aid in rephrasing sections of the manuscript.

FUNDING DECLARATION

This study is part of an Indian Council of Medical Research (ICMR), India project (Project ID IIRP-2023-1049) funded by Small Extramural Grants – 2023.

COMPETING INTERESTS

The authors declare no competing interests.