scientific data

Check for updates

OPEN

DATA DESCRIPTOR

High-resolution AI image dataset for diagnosing oral submucous fibrosis and squamous cell carcinoma

Nisha Chaudhary[1], Arpita Rai[2], Aakash Madhav Rao[3], Md Imam Faizan[1], Jeyaseelan Augustine[4], Akhilanand Chaurasia[5], Deepika Mishra[6], Akhilesh Chandra[7], Varnit Chauhan[1] & Tanveer Ahmad[1] ✉

Oral cancer is a global health challenge with a difficult histopathological diagnosis. The accurate histopathological interpretation of oral cancer tissue samples remains difficult. However, early diagnosis is very challenging due to a lack of experienced pathologists and inter- observer variability in diagnosis. The application of artificial intelligence (deep learning algorithms) for oral cancer histology images is very promising for rapid diagnosis. However, it requires a quality annotated dataset to build AI models. We present ORCHID (ORal Cancer Histology Image Database), a specialized database generated to advance research in AI-based histology image analytics of oral cancer and precancer. The ORCHID database is an extensive multicenter collection of high-resolution images captured at 1000X effective magnification (100X objective lens), encapsulating various oral cancer and precancer categories, such as oral submucous fibrosis (OSMF) and oral squamous cell carcinoma (OSCC). Additionally, it also contains grade-level sub-classifications for OSCC, such as well- differentiated (WD), moderately-differentiated (MD), and poorly-differentiated (PD). The database seeks to aid in developing innovative artificial intelligence-based rapid diagnostics for OSMF and OSCC, along with subtypes.

Background & Summary

Oral squamous cell carcinoma (OSCC) is a global cancer burden with a substantial number of individuals diagnosed each year, primarily in Southeast Asian countries where tobacco and associated products are commonly used[1,2]. Similarly, oral submucous fibrosis (OSMF) is a chronic and progressive condition that primarily affects the oral cavity[3,4]. It is characterized by the deposition of fibrous tissue in the submucosal layer, leading to restricted mouth opening, difficulty swallowing, and altered oral function[5]. OSMF also predominantly affects individuals in Southeast Asian countries where betel quid chewing is prevalent. The condition is known to have a potentially malignant nature, increasing the risk of developing oral cancer. Early detection and intervention are crucial in managing OSCC as well as OSMF and preventing its progression to malignancy.

The available diagnostic methods for OSMF and OSCC play a critical role in identifying and assessing these conditions. However, these methods have certain limitations that affect their accuracy and effectiveness. For OSMF, the diagnosis primarily relies on clinical examination and assessment of characteristic signs and symptoms[6]. The gold standard is a tissue biopsy and histopathology examination by a trained histopathologist. However, histopathological examination of the biopsy sample may not always provide a clear distinction between OSMF and early-stage OSCC, leading to diagnostic difficulties. In the case of OSCC, a combination of clinical examination, radiographic imaging, and biopsy is typically used for diagnosis. A clinical examination involves assessing the site, size, and appearance of the oral lesion. Radiographic imaging techniques such as computed tomography (CT) or magnetic resonance imaging (MRI) can help evaluate the extent of the tumor

[1]Multidisciplinary Centre for Advanced Research and Studies, Jamia Millia Islamia, New Delhi, India. [2]Rajendra Institute of Medical Sciences, Ranchi, Jharkhand, India. [3]Department of Computer Science, Ashoka University, Sonipat, Haryana, India. [4]Maulana Azad Institute of Dental Sciences, New Delhi, India. [5]King George Medical University, Lucknow, Uttar Pradesh, India. [6]All India Institute of Medical Sciences, New Delhi, India. [7]Banaras Hindu University, Banaras, Uttar Pradesh, India. ✉e-mail: tahmad7@jmi.ac.in

and identify possible metastasis[7]. Nevertheless, these imaging techniques exhibit restricted specificity when it comes to distinguishing between benign and malignant lesions.

Moreover, the lack of skilled histopathologists poses a significant obstacle, and the process of manual annotation further contributes to inter-observer discrepancies. Therefore, there is a need for further research and the development of more advanced diagnostic techniques that can improve the early detection and accurate diagnosis of these conditions, allowing for timely and appropriate management strategies to be implemented. To facilitate analysis, preprocessing of H&E images is necessary, followed by appropriate segmentation for further analysis. Computer-based algorithms have been employed to segment H&E stained images, successfully automating the process of separating the epithelial layer from the sub-epithelial layer[8]. This enables proper classification of tissue architectural changes and the extraction of relevant features for machine learning. However, despite these advancements, the application of these tools to human tissue samples has not yielded definitive results due to a lack of comprehensive histopathology databases.

To build deep learning algorithms, we need well-annotated H&E by expert histopathologists, but for oral cancer, we don't have enough large datasets on H&E. The lack of a publicly accessible histology image database for oral diseases presents a formidable obstacle. These databases, along with digital pathology databases, play a vital role in advancing healthcare by facilitating the development of more precise AI-based diagnostic tools. They serve as valuable resources for training and refining AI models tailored specifically for healthcare applications. Several publicly available databases have been established, housing distinct image datasets for various medical conditions, thereby aiding in the training and enhancement of AI algorithms[9,10]. However, in the realm of oral cancer, the availability of image data is noticeably limited compared to other cancer types like breast, lung, and skin cancer. Existing histology image databases primarily consist of tissue slide images related to OSCC, with none specifically including OSMF. Furthermore, there is a dearth of databases containing patch-level annotated images of OSCC. While certain research groups offer low-magnification image databases, these images fail to capture intricate nuclear features, making them unsuitable for training machine learning algorithms.

While whole slide imaging (WSI) offers advantages in generating large amounts of data and capturing comprehensive tissue information, challenges such as high computational requirements, software restrictions, and costs hinder its widespread use[11,12]. Further, issues related to image quality and uniformity in WSI datasets further complicate the integration of AI-powered algorithms effectively. Moreover, the lack of publicly accessible histology image databases specifically dedicated to oral diseases poses a significant challenge[13]. There is also a conspicuous absence of high-magnification images that comprehensively represent other oral diseases. Notably, oral conditions like OSMF lack adequate representation in these databases.

Addressing this gap necessitates that we present the ORCHID database for oral cancer, with specific emphasis on conditions like OSMF and OSCC. We believe the ORCHID database will aid the scientific community in building and harnessing AI technologies to enhance the accuracy and effectiveness of AI-based diagnostic tools, ultimately improving patient care and outcomes in the field of oral healthcare.

Methods

Human ethical clearance. Tissue slides were collected with the approval of an ethical committee from the participating hospitals and research institutions, (1) Jamia Millia Islamia, New Delhi(Proposal No.: 6(25/7/241/JMI/IEC/2021), (2) Maulana Azad Institute of Dental Sciences, New Delhi(Proposal No.: F./18/81/MAIDS/Ethical Committee/2016/8099); (3) Rajendra Institute of Medical Sciences, Jharkhand(Proposal No:. ECR/769/INST/JH/2015/RR-18/236), (4) Banaras Hindu University, Banaras(Proposal No.: Dean/2021/EC/2662), and (5) All India Institute of Medical Sciences, New Delhi(Proposal No.: IEC-828/03.12.2021, RP-33/2022), India. The buccal mucosa tissue samples were collected for three classes, normal, OSMF, and OSCC, with grade-wise annotation from the pathologists at each hospital. Data collection for the study was conducted with the explicit consent of the patients involved, following a rigorous ethical review and approval process carried out by relevant committees. Informed consent was obtained from all participants, ensuring they were fully aware of the study's purpose, procedures, potential risks, and benefits. They were given the opportunity to ask questions and seek clarification before providing their consent to participate. Participants willingly agreed to the open publication of their data, understanding that their identities would be protected and their information anonymized. The manuscript includes specific references to ethical approval granted by different institutions, indicating their compliance with ethical guidelines and regulations. These references serve as a means of tracking and verifying the study's adherence to ethical standards.

Haematoxylin and eosin staining (H&E). Biopsy samples of normal, OSMF and OSCC tissues underwent H&E staining. The staining procedure was conducted either in-house or outsourced to different laboratories. To eliminate staining variations across different laboratories, the preparation of H&E slides involved five histopathology labs, each utilizing their own independently developed and optimized protocols for the staining process. Following staining, the samples were examined under a microscope by a skilled histopathologist to assess cellular morphology, and tissue architecture, and identify any distinctive features or abnormalities specific to each sample type. This evaluation by the histopathologist involved grading the tissue slides for OSCC and OSMF, as well as differentiating between normal and diseased tissue sections. Subsequently, the annotated and validated images were utilized for further analysis.

Image acquisition. Images were acquired using a 1000X magnification (100X objective) lens from Leedz microimaging (LMI) bright field microscopy. To capture the images consistently, we utilized ToupView imaging software, which was configured for automatic adjustments. This setting applies to both white balance and camera settings, thereby standardizing the image acquisition process across different slides. The images of the H&E stained slides were captured at 1000X magnification(100X objective lens). By setting the ToupView software

to automatically adjust white balance and camera settings, we aimed to minimize human intervention and the variability it introduces. This approach ensures that the images are not only consistent but also replicable in different laboratory settings, provided similar equipment and software settings are used. We collected approximately 100–150 images per tissue slide, which were stored in PNG file format.

Expert annotation and validation. The data included in the ORCHID database underwent rigorous expert annotation and validation to ensure a high level of quality and accuracy. In our expert validation process, 'sufficient detail' for an image to be qualified was determined based on several key criteria. Firstly, the clarity of histological features which depict the necessary histological structures, such as cellular details and tissue architecture. Images should be free from artifacts that could interfere with accurate interpretation (e.g., folds, tears, excessive staining). The image must be in focus, with appropriate contrast and resolution to discern pathological features. Our team of pathologists and histopathology experts independently assessed each image against these criteria to ensure only high-quality images were included in our study. Images that were blurry or lacked sufficient detail were dismissed as they would not provide accurate or reliable information. Next, the experts evaluated the annotations that accompany the images. These annotations were scrutinized for consistency and accuracy, to ensure that they accurately represented the disease conditions depicted in the images. The process of labeling the slides was conducted manually by trained pathology experts. This involved a careful review of each slide to identify and label the specific disease conditions present. This procedure was crucial to ensure that the slides were correctly categorized. Furthermore, the slides that showed staining artifacts were also rejected. Staining artifacts can occur during the preparation of the slides and can alter the appearance of the tissue, potentially leading to misinterpretation or incorrect diagnosis. As such, only slides that were free from such errors and provided a clear and accurate representation of the oral pathology were included in the database. These standardization processes ensure that AI models are trained and validated on data that consistently represent the true pathological features. Standardized and validated data enhance the model's ability to generalize findings across different datasets and real-world scenarios.

Stain normalization. The handling of the samples at each hospital during the collection of the tissue samples led to staining problems that persisted even after following the established H&E staining protocol. To address and minimize the variations in staining appearance across different sites in the H&E images, a stain normalization method was implemented, specifically the Reinhard stain normalization technique[14], as shown in Fig. 1b. This approach, described in the study, involves a series of steps to standardize the color properties of the images to a desired standard. The first step is scaling the input image to match the target image statistics. This involves adjusting the intensity values of the input image to align with the desired color distribution of the target image. The scaling ensures that the overall brightness and contrast of the input image are consistent with the target image. The next step involves transforming the image from the RGB color space to the LAB color space proposed by Ruderman. The LAB color space separates the image into three channels: L (lightness), A (green-red color component), and B (blue-yellow color component). By performing the transformation, the image is represented in a color space that better captures the perceptual differences in human vision. Finally, Reinhard color normalization is applied to the LAB image. Reinhard color normalization adjusts the color properties of the image to align with a desired standard. It achieves this by equalizing the mean and standard deviation of the LAB channels across the image.

If the LAB statistics for the input image are not provided, they are derived from the input image itself. This ensures that the normalization process is tailored to each individual image. Below is the equation for the same:

$$I_n = I_o * (1 + k_1 * (L_o - \mu_L) + k_2 * (S_o - \mu_S))$$

where:

I_n is the normalized image
I_o is the original image
k_1 and k_2 are constants that are chosen to optimize the appearance of the normalized image
L_o is the average brightness of the original image
S_o is the average saturation of the original image
μ_L and μ_S are the average brightness and saturation of a reference image.

By minimizing variations in staining and image quality, AI models can focus on learning relevant pathological patterns rather than adapting to artifacts or inconsistencies.

Patch generation. After normalization, we generated image patches of size 512 by 512 pixels from 1000X magnified images (Fig. 1d). The patches were generated by left-to-right sequential cropping (overlapping 256 pixels) in the original images. Below is the procedure for automatically creating overlapping patches from original large-size image:

Let W and H represent the width and height of the input image, respectively.
P_w and P_h as the width and height of each patch. Here, $P_w = P_h = 512$.
Let O denote the overlap between adjacent patches, with $O = 256$.

The number of patches along the width (N_x) and height (N_y) of the image can be calculated using the formula:

$$N_x = \lfloor W - P_w/P_w - O \rfloor + 1, N_y = \lfloor H - P_h/P_h - O \rfloor + 1$$

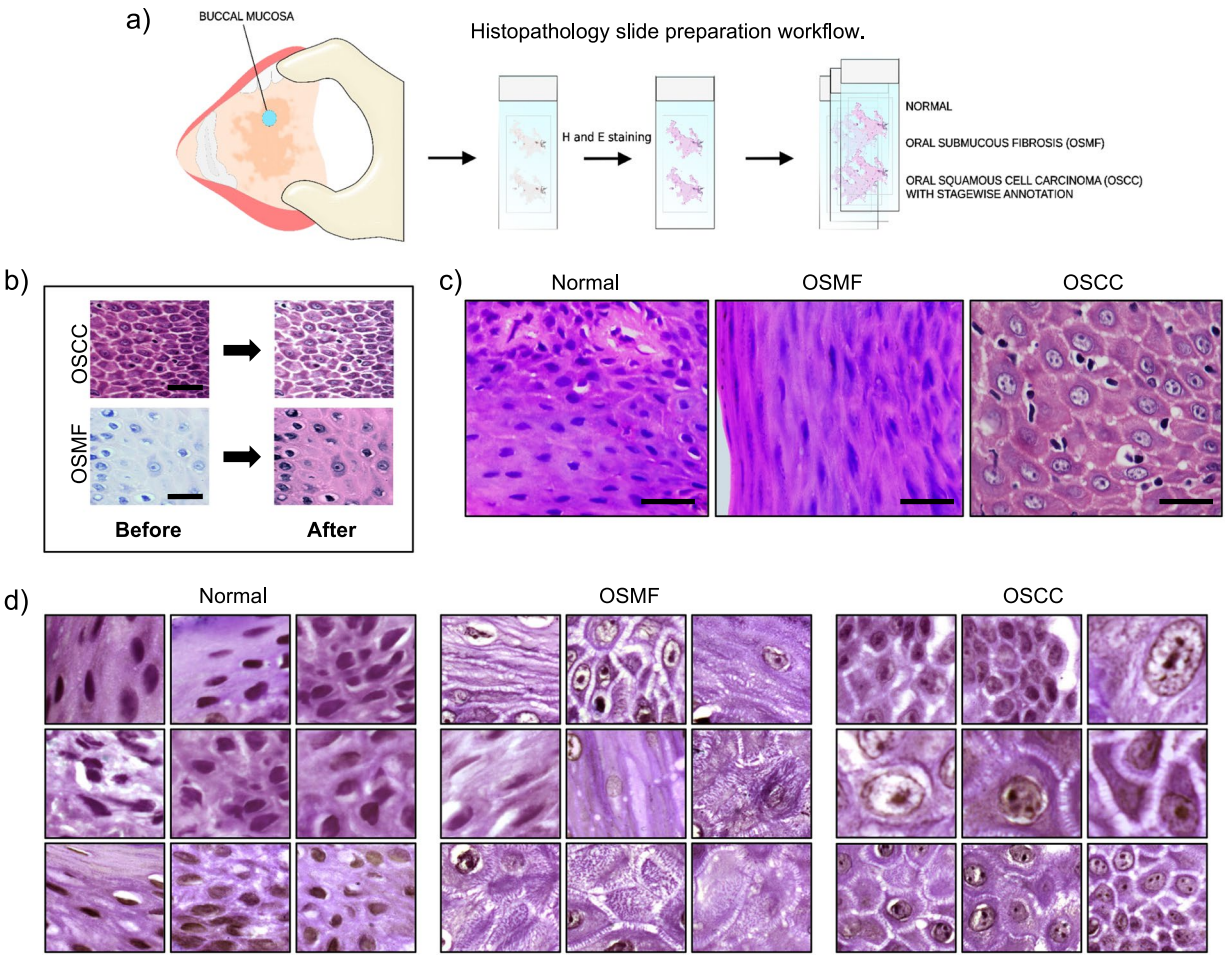

Fig. 1 Workflow, Image Analysis, and Stain Normalization. (**a**) The workflow for preparing oral histopathology slides involves a series of steps, from the collection of tissue samples to slide preparation and staining. (**b**) Stain normalization is performed to standardize the stain appearance in the images. The Reinhard stain normalization method is utilized for this purpose, ensuring consistent and comparable staining across the images. The scale bar is 10 μm. (**c**) Representative images captured at a magnification of 1000X exhibit normal tissue, cases of OSMF, and cases of OSCC. These images were digitized using bright field microscopy, providing a visual depiction of the different stages involved in the preparation and staining of tissue slides. The scale bar is 10 μm. (**d**) Image patches, measuring 512 by 512, are generated from the 1000X images of normal tissue, OSMF cases, and OSCC cases. These patches serve as representative examples of specific regions within the larger images, offering focused insights into the characteristics of normal tissue as well as OSMF and OSCC conditions.

For a given patch at position, (i,j), where i ranges from 0 to $N_x - 1$ and j ranges from 0 to $N_y - 1$, the top-left corner (x_{tl}, y_{tl}) and bottom-right corner (x_{br}, y_{br}) of the patch are determined as:

$$x_{tl} = i \times (P_w - O)$$

$$y_{tl} = j \times (P_h - O)$$

$$x_{br} = x_{tl} + P_w$$

$$y_{br} = y_{tl} + P_h$$

The condition $x_{br} \leq W$ and $y_{br} \leq H$ ensures the patch is within the image bounds.

Baseline model development and fine-tuning. We performed benchmarking of ten deep Convolutional Neural Network (DCNN) through pre-training and fine-tuning our models aimed at classifying oral cancer from non-cancerous samples (Fig. 3). The study focuses on three class classification tasks: Normal vs. Oral Submucous Fibrosis (OSMF) vs. Oral Squamous Cell Carcinoma (OSCC).

ORCHID

Preparatory data		Test data
Train [70%]	Validation [20%]	Test [10%]

Split	Classes	Image Count(100X)	Patches(512)
train	Normal	1,045	30,122
	OSMF	2,095	53,453
	WDOSCC	2,790	49,968
	MDOSCC	2,699	56,138
	PDOSCC	1,599	44,256
val	Normal	294	8,481
	OSMF	592	15,104
	WDOSCC	788	14,126
	MDOSCC	760	15,878
	PDOSCC	451	12,532
test	Normal	163	4,706
	OSMF	328	8,267
	WDOSCC	433	7,692
	MDOSCC	421	8,566
	PDOSCC	247	6,760
	Total	**14,705**	**3,36,049**

WD= **W**ell **D**ifferentiated, **MD**=**M**oderately **D**ifferentiated and **PD**= **P**oorly **D**ifferentiated

Fig. 2 Details statistics of the ORCHID database. This figure provides a distribution of the ORCHID dataset images and patient cases across five categories: normal samples, and samples with varying degrees of differentiation in Oral Squamous Cell Carcinoma (OSMF, PDOSCC, MDOSCC, WDOSCC). The entire dataset was split into 70% train, 20% validation and 10% test set.

The OSCC class aggregates three distinct stages: Well-Differentiated (WD), Moderately Differentiated (MD), and Poorly Differentiated (PD) OSCC, treating them as a unified class due to their common pathological origin.

The InceptionV3[15] model was pre-trained on the ImageNet dataset, providing a strong initial set of learned features. It not only offers a robust balance between accuracy and overfitting but also exceeds in computational efficiency. The architecture of InceptionV3 is uniquely suited to handle the complexity and variability in the ORCHID dataset, making it an optimal choice for ensuring both high performance and applicability in a clinical setting. The model's top layers were excluded to allow for customization. The model was then fine-tuned by setting all layers in the InceptionV3 model as trainable. This process allows the model to adapt to the specific dataset being used in the study. A flattened layer was added to convert the output of the InceptionV3 model into a 1-dimensional tensor. This was followed by a global average pooling 2D layer, a dense layer with 128 units and a ReLU activation and L2 regularization (penalty = 0.01) function, facilitating feature extraction and non-linear transformations. Finally, a dense layer with 3 units and a softmax activation function was employed to produce the output probabilities for the three classes in the classification task. The model was compiled using the RMSprop optimizer with a learning rate of 0.0000001 (or 10e-7) and trained with the categorical cross-entropy loss function. The training process was executed for 50 epochs, with performance evaluation across three-fold dataset.

The same settings were used for both the classification models that are; the first model(model-1) which classifies the image patched into normal, OSMF, and OSCC, and the second model(model-2) which classifies the image patches into WD, MD, and PD grades of OSCC. This baseline architecture aimed to capture local and global patterns within the cellular graphs indicative of cancerous transformations.

To ensure the reproducibility of results, random seeds = 42 were set during dataset splitting and model initialization phases. This practice guaranteed consistent data shuffling and initialization patterns across experiments. The refined InceptionV3 model demonstrated improved classification performance across all tasks, with notable gains in precision and recall. The model effectively captured the nuclear structures, distinguishing between normal, OSMF, and OSCC conditions with high accuracy. The development and fine-tuning of the InceptionV3 model for oral cancer classification exemplify the potential of DCNN in biomedical applications.

Data Records

The ORCHID data used in the current study has been made available at Zenodo under the CC-BY 4.0 license[16,17].

The data consists of digitized slides that were collected from 150 patient samples and stained for analysis (Table S1). The digitization process involved capturing images using a 1000X magnification (100X objective) lens, as depicted in Fig. 1a. The dataset encompasses images from three distinct classes: normal, OSMF, and OSCC. Each class folder within the dataset contains image tiles generated at a 1000X magnification level. The visual representation of these images is depicted in Fig. 1c which showcase the appearance and characteristics of the different classes. To provide a quantitative overview of the ORCHID dataset. Figure 2 presents a summary of the number of images available in each of the five classes(folders), which are as follows, Normal, OSMF, WDOSCC, MDOSCC, and PDOSCC.

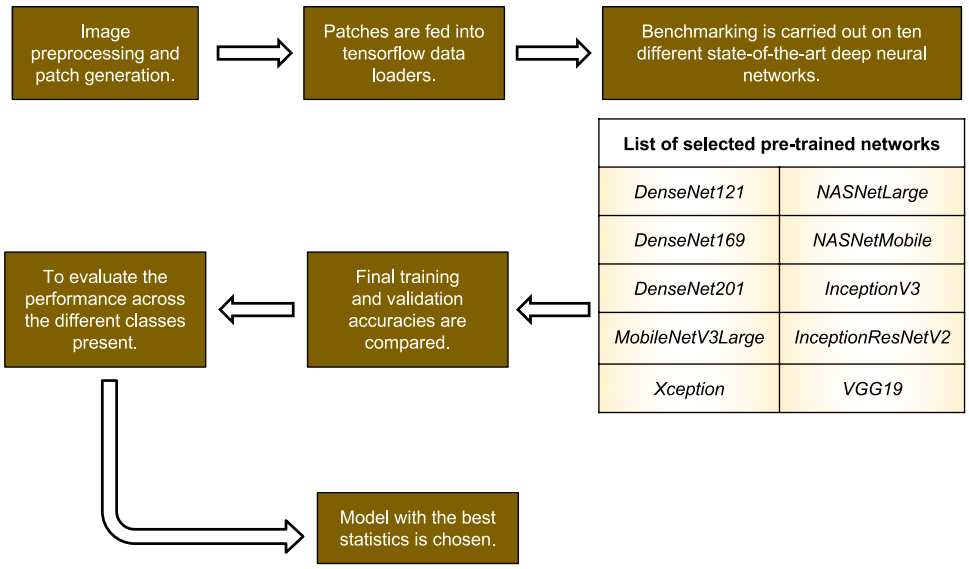

Fig. 3 Flowchart describing the process to benchmark which pre-trained model to choose. The flowchart serves as a visual representation of the process involved in selecting an appropriate pre-trained model for a specific task. It outlines the steps and criteria to consider when evaluating different models. Further, the flowchart provides a systematic approach to benchmarking various pre-trained models, taking into account factors such as model architecture, training data, performance metrics, and compatibility with the classification task at hand.

Each class folder consists of subfolders representing different tissue slides collected from different patients. The naming of these folders start from the initials of dataset name, 'o', followed by class ID and source ID from where the sample has been collected and lastly the sample ID itself. All the images are stored inside these subfolders as per the tissue slide, they belong to. The naming of images is done in such a way, that each label represents, first the dataset name-'o', followed by class ID, source ID, then patient-ID and lastly the image-ID. The tabulated information helps to understand the distribution and proportion of images within each split and class, aiding in the analysis and utilization of the ORCHID dataset for the study or related research endeavors.

Technical Validation

The histology images in the ORCHID database involved a rigorous and systematic approach to ensure the reliability and accuracy of the dataset. To validate the dataset, it was used as input for DCNN algorithms for the classification of pathology images within the ORCHID dataset. Technical validation involves a series of experiments and analyses to ensure the robustness, reliability, and generalizability of the model on the ORCHID dataset (Fig. S1).

This initial dataset breakdown is critical for ensuring a balanced representation of classes within the model training and validation processes. The dataset is clearly partitioned into training (70%), validation (20%), and testing (10%) sets, with an appropriate distribution of images (Fig. 2). The patches of size 512 by 512 were generated consecutively. This was essential for avoiding bias in the model's predictive performance. Specifically, the chosen architecture for the DCNN models was InceptionV3, and the details of its configuration and implementation can be found in the methods section. For the training of DCNN, we first took the patches from the train and validation set and carried out a three-fold cross-validation experiment. Both the models were evaluated across three folds to ensure consistency in their predictive capabilities.

The results demonstrated that the model successfully distinguished between the normal, OSMF, and OSCC classes, with a training accuracy of 99.18% and a validation accuracy of 98.25%, as illustrated in Fig. 4a. However, when classifying the three different grades of OSCC (WD, MD, and PD), the classification accuracy was slightly lower. The model achieved a training accuracy of 92.81% and validation accuracy of 91.53%, as shown in Fig. 5a. Thus proving the model competency in inheriting the spatial structure within tissue samples, providing a novel approach to oral cancer diagnosis.

They also underscore the importance of comprehensive metrics beyond just accuracy, including precision (the model's ability to avoid false positives), recall (the model's ability to find all the positive samples), and the F1 score (a balance between precision and recall), for evaluating the clinical utility of such diagnostic models.

On the test set, the model-1 maintained high performance with accuracy ranging from 97.51% to 97.69%, precision from 97.53% to 97.71%, recall from 97.51% to 97.69%, and F1-score from 97.51% to 97.70%, evaluated over 35,991 data points (Fig. 4b,c and Table S2). These results indicate that the models are not only robust during validation but also maintain their predictive power on unseen data. However, model-2 shows the accuracies ranged from 89.24% to 89.78%, with precision, recall, and F1-scores closely aligned, across 23,018 data points (Fig. 5b,c and Table S2). Overall, the results demonstrate that the InceptionV3-based DCNN models are highly capable of distinguishing between Normal, OSMF, and OSCC, as well as the subtypes within OSCC.

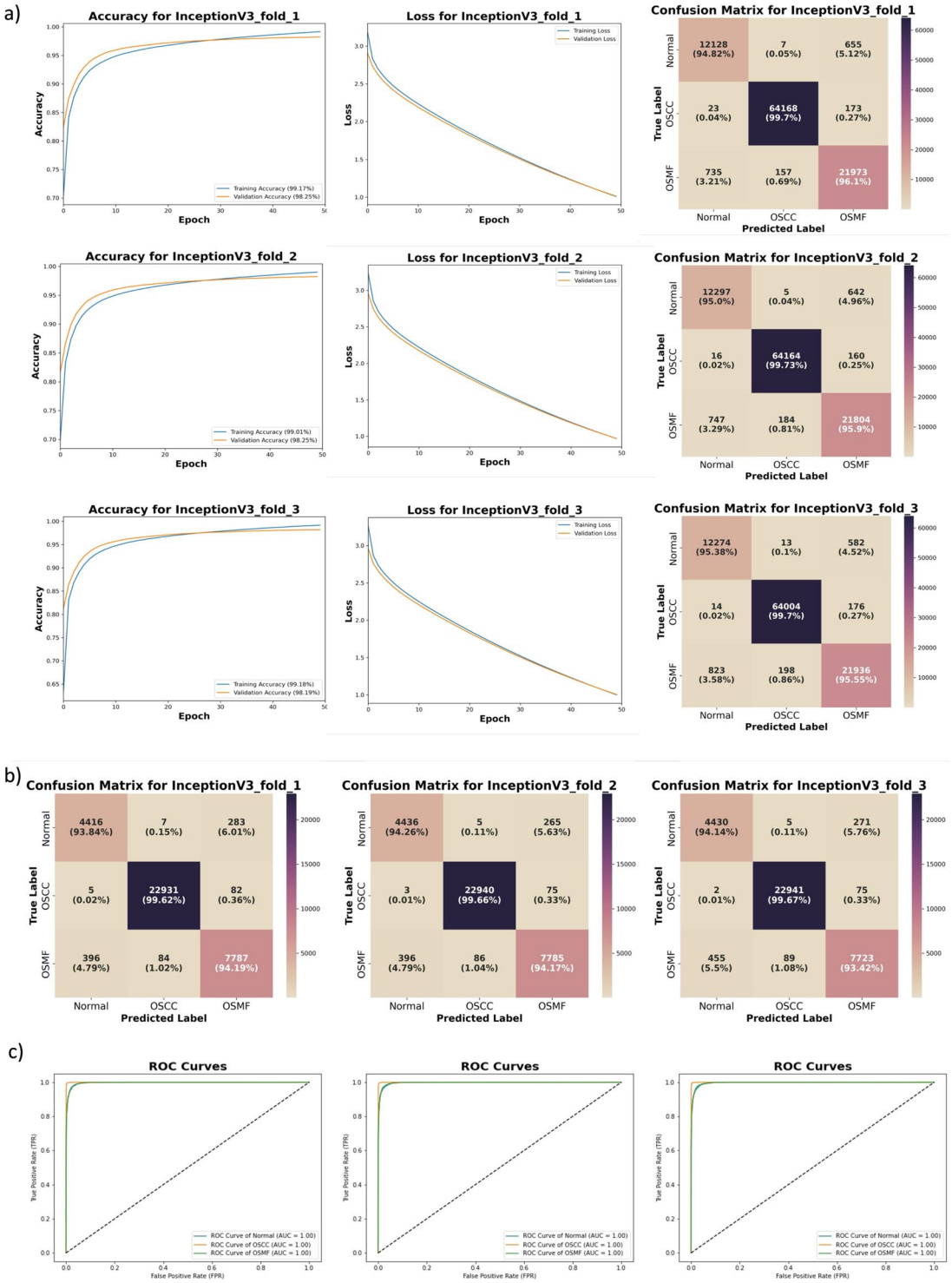

Fig. 4 Model-1 performance evaluation metrics. (**a**) Classification performance of the InceptionV3 model for normal, OSMF and OSCC. This model focus on assessing the effectiveness of classification algorithm InceptionV3 in distinguishing between different oral tissue conditions: Normal, OSMF, and OSCC. The metrics provide a quantitative measurement of the accuracy, loss, and prediction performance (confusion matrix) across three folds data for training and validation set. (**b**) Testing of classification model-1 on 35,991 test images belonging to 3 classes. (**c**) AUC-ROC curve for each class.

The technical validation process ensures that the ORCHID database is reliable and suitable for subsequent analysis and research. We have further supplemented the consistent performance of the models on the external data as well analysis of which are incorporated in Tables S7 and S8 (Fig. S2). The performance of the InceptionV3 model on the dataset demonstrates its capability to accurately classify normal, OSMF, and OSCC cases.

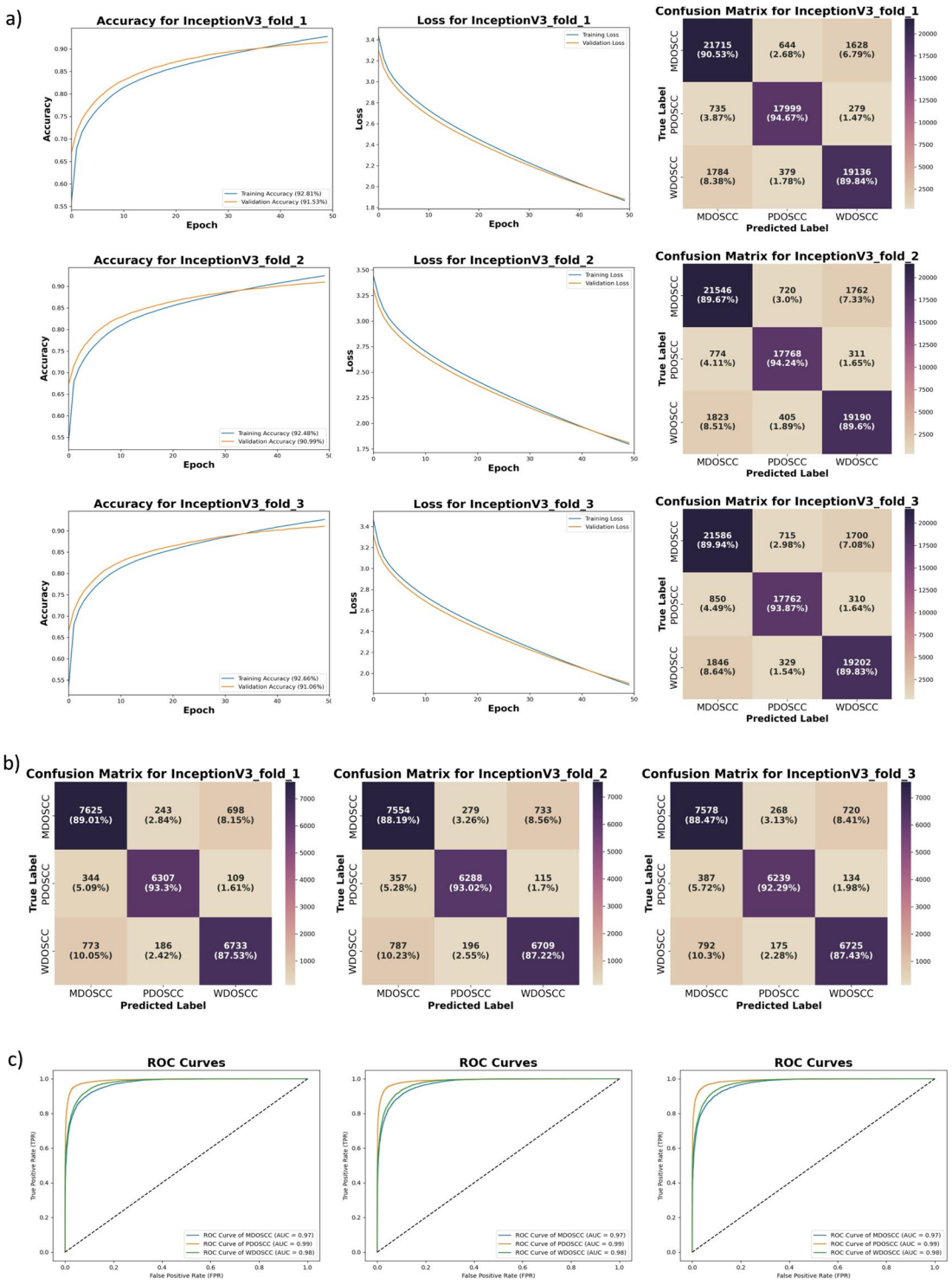

Fig. 5 Model-2 performance evaluation metrics. (**a**) Classification performance of the InceptionV3 model for well-differentiated(WD) OSCC, moderately-differentiated(MD) OSCC, and poorly-differentiated(PD) OSCC. This model specifically focus on the classification of different grades of OSCC: well-differentiated(WD) OSCC, moderately-differentiated(MD) OSCC, and poorly-differentiated(PD) OSCC. The metrics measure the performance of classification algorithm, InceptionV3 in correctly classifying and differentiating between these different grades of OSCC. (**b**) Testing of classification model-2 on 23,018 test images belonging to 3 classes. (**c**) AUC-ROC curve for each class.

Nevertheless, the performance in classifying the different grades of OSCC indicates the potential for improvement and calls for additional investigation and refinement. Furthermore, the need for a substantially larger image dataset is recognized, and efforts toward expanding the dataset are currently underway as part of ongoing work.

Limitations. While the ORCHID dataset offers high-resolution images, certain limitations must be acknowledged, especially in comparison to whole slide images (WSIs). High-resolution images, while providing detailed cellular and subcellular structures, are limited in their field of view compared to WSIs. This can lead to a potential loss of contextual information that is often crucial for comprehensive pathological analysis. High-resolution images represent specific areas of interest within a tissue slide, potentially missing other diagnostically significant regions. On the other hand, WSIs covers the entire tissue section, thus reducing the risk of missing critical diagnostic features. While high-resolution images reduce computational demands compared to WSIs, they still require substantial processing power for analysis due to their size and detail level. This could limit the accessibility of advanced AI-driven tools for some research and clinical settings with less computational infrastructure. By focusing on specific tissue areas, this may pose integration challenges with other diagnostic modalities or broader histopathological datasets. The precise annotation of high-resolution images demands significant expertise and time, especially given their detailed nature. This can lead to variability in annotations, affecting the consistency and reproducibility of AI model training. AI models trained solely on high-resolution patches might not generalize well to full slide contexts or to different magnifications and preparations, potentially limiting their applicability across diverse clinical settings. Addressing these limitations will require ongoing research, including the development of hybrid approaches that combine the strengths of high-resolution imaging with the comprehensive perspective offered by WSIs, and efforts to enhance computational methods for managing large-scale high-resolution data efficiently. Even though there are difficulties, ORCHID provides very clear images that allow us to see tiny abnormalities. Hence, our primary goal with this dataset is to establish a foundation for future research and to offer a high-quality resource that can be expanded upon.

In summary, we have made an initial attempt to provide a comprehensive image database for two of the most prominent oral conditions, OSCC and OSMF. We believe that more such databases will be made publicly available in the near future. These comprehensive image databases will facilitate the development of accurate AI-based diagnostic tools for oral diseases, ultimately improving patient care and outcomes in the field of oral healthcare. In future, integration of databases comprising molecular markers, transcriptome, metabolome, and other biomarkers, combined with oral histological image through advanced AI-driven imaging techniques, holds great promise in improving diagnostic accuracy and precision. This potential has already been observed in the diagnosis of lung and breast cancers[18]. This expansion will aid in developing a more comprehensive AI-driven diagnostic tool.

By making this dataset openly accessible, we aim to encourage other researchers to contribute additional data and annotations, thus facilitating the growth of a more extensive and diverse dataset over time.

Code availability

The ORCHID data is uploaded via Zenodo platform in.zip file format. The tools to generate the data for training of the DCNN are provided in the Python scripts updated on the GitHub account. https://github.com/NishaChaudhary23/ORCHID/.

Received: 21 May 2024; Accepted: 22 August 2024;
Published online: 27 September 2024

References

1. Sung, H. *et al*. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).
2. Mailankody, S. *et al*. Epidemiology of rare cancers in India and South Asian countries–remembering the forgotten. *Lancet Reg. Health-Southeast Asia* **12** (2023).
3. Prabhu, S. R., Wilson, D., Daftary, D. K. & Johnson, N. W. Oral diseases in the tropics. in *Oral diseases in the tropics* xxv–794 (1991).
4. Rao, N. R. *et al*. Oral submucous fibrosis: A contemporary narrative review with a proposed inter-professional approach for an early diagnosis and clinical management. *J. Otolaryngol. - Head Neck Surg.* **49**, 3 (2020).
5. Tekade, S. A. *et al*. Early stage oral submucous fibrosis is characterized by increased vascularity as opposed to advanced stages. *J. Clin. Diagn. Res. JCDR* **11**, ZC92 (2017).
6. El-Naggar, A. K. WHO classification of head and neck tumours. (2017).
7. Thoenissen, P. *et al*. The role of magnetic resonance imaging and computed tomography in oral squamous cell carcinoma patients' preoperative staging. *Front. Oncol.* **13**, 972042 (2023).
8. Rathore, A. S., Gupta, A., Shetty, D. C., Kumar, K. & Dhanapal, R. Redefining epithelial characterization in oral submucous fibrosis using morphometric analysis. *J. Oral Maxillofac. Pathol.* **21**, 36–40 (2017).
9. Willemink, M. J. *et al*. Preparing Medical Imaging Data for Machine Learning. *Radiology* **295**, 4–15 (2020).
10. Prior, F. *et al*. Open access image repositories: high-quality data to enable machine learning research. *Clin. Radiol.* **75**, 7–12 (2020).
11. Dimitriou, N., Arandjelović, O. & Caie, P. D. Deep learning for whole slide image analysis: an overview. *Front. Med.* **6**, 264 (2019).
12. Smith, B., Hermsen, M., Lesser, E., Ravichandar, D. & Kremers, W. Developing image analysis pipelines of whole-slide images: Pre- and post-processing. *J. Clin. Transl. Sci.* **5**, e38 (2021).
13. Rahman, T. Y., Mahanta, L. B., Das, A. K. & Sarma, J. D. Histopathological imaging database for oral cancer analysis. *Data Brief* **29**, 105114 (2020).
14. Reinhard, E., Adhikhmin, M., Gooch, B. & Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **21**, 34–41 (2001).
15. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 2818–2826 (2016).
16. Chaudhary, N. *et al*. High-resolution AI image dataset for diagnosing oral submucous fibrosis and squamous cell carcinoma *Zenodo*. https://doi.org/10.5281/zenodo.12636426 (2024)

17. Chaudhary, N., & Ahmad, T. Validation and Test Datasets for "High-resolution AI image dataset for diagnosing oral submucous fibrosis and squamous cell carcinoma" *Zenodo*. https://doi.org/10.5281/zenodo.12646943 (2024).
18. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit. Med.* **5**, 48 (2022).

Acknowledgements

N.C. is the recipient of a senior research fellowship from the Indian Council of Medical Research(3/1/2(1)/Oral/2021-NCD-II), New Delhi, India. This work was also supported by the Science and Engineering Research Board (CRG/2020/002294), and the Indian Council of Medical Research (ICMR) (GIA/2019/000274/PRCGIA (Ver-1)), New Delhi, India. We also acknowledge the computing support from the Mphasis F1 Foundation and the Center for Bioinformatics and Computational Biology (B.I.C.) (BT/PR40220/BTIS/137/22/2021) facility at Ashoka University. We are thanking Farhat Zeba and Sumra Khan for helping out with the imaging.

Author contributions

N.C. and T.A. conceptualized the study. A.R. and M.I.F. collected and processed tissue slides. N.C., M.I.F., and V.C. performed microscopy imaging. N.C. and A.M.R. performed image processing, DCNN training, designed the ORCHID, and analyzed the results. A.R., J.A., AA.C., D.M., and A.C. annotated the oral cases and provided their pathology expertise and guidance. T.A. supervised the study. All the co-authors contributed feedback and suggestions towards the preparation of 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/s41597-024-03836-6.

Correspondence and requests for materials should be addressed to T.A.