scientific reports

Check for updates

OPEN

Attention-guided convolutional network for bias-mitigated and interpretable oral lesion classification

Adeetya Patel[1], Camille Besombes[1], Theerthika Dillibabu[1], Mridul Sharma[1], Faleh Tamimi[2], Maxime Ducret[3], Peter Chauvin[1] & Sreenath Madathil[1]✉

Accurate diagnosis of oral lesions, early indicators of oral cancer, is a complex clinical challenge. Recent advances in deep learning have demonstrated potential in supporting clinical decisions. This paper introduces a deep learning model for classifying oral lesions, focusing on accuracy, interpretability, and reducing dataset bias. The model integrates three components: (i) a Classification Stream, utilizing a CNN to categorize images into 16 lesion types (baseline model), (ii) a Guidance Stream, which aligns class activation maps with clinically relevant areas using ground truth segmentation masks (GAIN model), and (iii) an Anatomical Site Prediction Stream, improving interpretability by predicting lesion location (GAIN+ASP model). The development dataset comprised 2765 intra-oral digital images of 16 lesion types from 1079 patients seen at an oral pathology clinic between 1999 and 2021. The GAIN model demonstrated a 7.2% relative improvement in accuracy over the baseline for 16-class classification, with superior class-specific balanced accuracy and AUC scores. Additionally, the GAIN model enhanced lesion localization and improved the alignment between attention maps and ground truth. The proposed models also exhibited greater robustness against dataset bias, as shown in ablation studies.

Keywords Oral lesion diagnosis, Interpretability, Guided attention inference network, Bias mitigation, CNN

More than 300,000 oral cancers (OC) are diagnosed annually worldwide.[1] These cancers have a much lower probability of survival ($< 50\%$) than common cancers (e.g., breast, prostate)[2], and one of the highest morbidity and suicide rates of all cancers. Importantly, more than half of all patients present with an advanced-stage disease; this proportion has not decreased in the past 40 years[3,4]. An advanced stage at diagnosis is associated with a lower probability of survival, higher health care cost, and increased risk of significant impairment and deformity[5,6]. Hence, early diagnosis plays a key role in reducing the burden of this disease.

An abnormal change in the soft tissue of the oral cavity (oral lesion) can be one of the earliest signs of OC. Hence, early diagnosis of OC can be achieved by correctly situating any oral soft tissue lesion in the continuum from normal mucosa to malignancy. For example, screening for abnormal oral lesions, even based on visual examination alone, has been shown to significantly reduce the morbidity and mortality associated with OC in high-risk populations[7–10]. However, many general dentists do not appear to adopt a systematic approach for evaluating oral lesions[11,12] and also report major challenges in diagnosing oral lesions[12–14]. For example, half of the general dentists surveyed in a study felt that their knowledge and training were not up to date in diagnosing oral lesions[14]. Furthermore, there is a wide variety of oral lesions with similar clinical appearances, and many systemic conditions have oral manifestations, which make them more difficult to diagnose[13,15–17]. These difficulties, combined with limited clinical training in oral pathology, lead to high diagnostic uncertainty and delay[16,18,19].

Deep learning algorithms have demonstrated remarkable success in various image recognition tasks[20]. In the context of oral lesion classification, there have been several attempts to develop CNN-based models to aid in clinical decision-making[21–27] Almost all of these previous attempts have been focused on the classification task of cancer vs no-cancer. This approach may miss a large group of potentially malignant oral lesions, where an early and accurate diagnosis and excision could prevent a malignant transformation. Furthermore, two key

[1]Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montreal, Canada. [2]College of Dental Medicine, QU Health, Qatar University, Doha, Qatar. [3]Faculté d'Odontologie, Université Claude Bernard Lyon 1, Lyon, France. ✉email: sreenath.madathil@mcgill.ca

challenges need to be addressed to enhance the practicality and reliability of these models in clinical practice: interpretability and dataset bias[28–30].

Interpretability refers to the ability to understand and explain the model's decision-making process, enabling clinicians to trust and interpret the predictions from the model effectively. While CNNs have demonstrated impressive performance in various domains, they are often considered black-box models[31], making it difficult to ascertain the underlying features and reasoning behind their predictions[32]. This lack of interpretability hinders their adoption in clinical settings, where explanations and insights are crucial for informed decision-making and building trust between the model and healthcare professionals[33,34].

Furthermore, dataset bias, including oral lesion classification, is common in medical image analysis[35]. Biases can arise due to variations in imaging conditions, equipment, patient demographics, and lesion characteristics, among other factors[28,36]. These biases can introduce unwanted confounding factors and affect the generalization ability of deep learning models[37]. Consequently, the models may exhibit poor performance when presented with data that significantly deviates from the training distribution, thereby limiting their clinical utility and potentially leading to misdiagnosis or inadequate treatment plans[38,39].

To address these challenges, we propose a deep learning model for oral lesion classification that emphasizes interpretability and robustness against dataset bias. Our method draws inspiration from the existing GAIN (Guided Attention Inference Network) framework, a weakly supervised semantic segmentation approach[40]. While GAIN has been previously proposed and utilized in other domains[41,42], our contribution lies in its adaptation and application specifically for the domain of oral lesions classification task. Moreover, our approach also aims to address the issue of dataset bias[43] by leveraging segmented masks to guide the attention of a convolutional neural network (CNN) during training. These masks highlight the relevant regions within oral cavity images, enabling the model to focus on the informative areas while suppressing the influence of irrelevant or confounding factors. By incorporating this guidance mechanism, our approach reduces the impact of variations in imaging conditions, patient demographics, and lesion characteristics that contribute to dataset bias[44]. Furthermore, we augment the network with the task of predicting the anatomical location of the lesion within the oral cavity image, considering eight potential anatomical locations. This augmentation serves a dual purpose: It enriches the model's interpretability and capitalizes on the standard set of CNN features shared between the anatomical site prediction and lesion classification tasks. Our underlying hypothesis is rooted in the idea that capturing the inherent relationship between lesion types and their corresponding anatomical sites can yield performance improvements for the overall model, thereby amplifying its clinical utility[45].

Here, we present a comprehensive evaluation of our proposed approach on a dataset of oral cavity images encompassing 16 different types of lesions. We compare the performance of our method with state-of-the-art approaches and demonstrate its effectiveness in achieving accurate and interpretable oral lesions classification. Additionally, we assess the robustness of our model to dataset bias by conducting experiments on diverse data subsets.

Results

Our approach of incorporating guided attention considerably improved the performance of the model compared to the standard fine-tuning of EfficientNet-B5 for the oral lesion classification task, with the added advantage of improved interpretability and robustness against dataset bias.

Oral lesion classification performance

The class-specific and overall performance metrics are presented in Table 1. The `GAIN` model showed 7.2% relative improvement in balanced accuracy, increasing from 73.4 to 78.7%. Additionally, the model showed an improved brier score and comparable AUC values. Although the `GAIN+ASP` model exhibited modest incremental performance gains compared to the `GAIN` model, it still surpassed the baseline performance by 2.8% relative improvement.

Class-specific metrics revealed that the `GAIN` model achieved clear improvements in balanced accuracy for all classes except two, compared to the baseline model. However, for those two classes the `GAIN+ASP` model showed comparable accuracy to the baseline model. Notably, adding anatomical site classification improved the discriminatory power of the model for the majority (11 out of 16) classes, with AUC values ranging from 83.7 to 99.7%. The `GAIN` model showed significant improvements in balanced accuracy for Squamous Cell Carcinoma (oral cancer) and Mucocele lesion classes, with 10% and 16% relative improvements over the baseline, respectively. Moreover, the `GAIN` and `GAIN+ASP` models improved balances accuracy and F1 scores when considering broader class labels based on prognosis of lesions (non-benign vs benign vs premalignant vs malignant), suggesting a improvement in potential clinical utility (Appendix Table A1).

Anatomical site classification performance

The `GAIN+ASP` model showed exceptional performance in the classification of anatomical sites, achieving an overall balanced accuracy of 91.6%. Site-specific accuracies ranged from 76.9 to 94.9%, as detailed in Table 1. Furthermore, the model showed strong discriminatory abilities between different anatomical sites, with AUC values spanning from 92.3 to 98.78%. The results underscore the model's robustness and effectiveness in distinguishing between various anatomical locations.

Improvement in interpretability

The interpretability of our proposed methodology was evaluated by comparing the Intersection over Union (IoU), and Dice-Sørensen coefficient (DSC) between the baseline model and our methods, `GAIN` and `GAIN+ASP`. The IoU measures the overlap between the attention maps and ground truth annotations, indicating the alignment between the generated maps and the actual lesion regions. Our `GAIN+ASP` model showed a 6.5% relative

	Baseline			GAIN			GAIN+ASP		
	BS (↓)	BA (↑)	AUC (↑)	BS (↓)	BA (↑)	AUC (↑)	BS (↓)	BA (↑)	AUC (↑)
Oral lesion classes									
Actinic cheilitis solar	0.022	0.953	0.981	0.024	**0.973**	**0.991**	**0.020**	0.926	0.981
Aphthous ulcers	0.053	0.648	0.855	0.054	**0.685**	0.844	**0.051**	0.656	**0.881**
Cheek lip tongue chewing	0.057	0.522	0.837	0.057	**0.578**	0.700	**0.056**	0.576	**0.837**
Denture stomatitis	**0.012**	**0.896**	**0.986**	0.013	0.778	0.935	0.013	0.890	0.955
Fordyce granules	0.017	0.732	0.879	0.015	**0.787**	**0.937**	**0.014**	0.699	0.920
Geographic tongue	0.152	0.780	0.960	**0.137**	**0.866**	0.968	0.150	0.799	**0.973**
Gingival hyperplasia	**0.027**	0.767	0.851	0.030	0.702	0.835	0.028	**0.789**	**0.862**
Gingival cyst	0.017	0.788	0.910	**0.016**	0.793	**0.912**	0.018	**0.810**	0.800
Gingivitis	0.029	0.773	0.694	0.031	**0.774**	**0.810**	**0.029**	0.736	0.686
Hairy tongue	0.037	**0.913**	0.966	0.037	0.870	0.952	**0.035**	0.900	**0.972**
Leukoedema	0.016	0.781	0.913	**0.015**	0.797	0.934	0.016	**0.857**	**0.933**
Lymphoepithelial cyst	0.026	0.661	**0.831**	**0.024**	**0.716**	0.772	0.024	0.616	0.744
Lymphoid tissue	0.028	0.814	0.835	**0.025**	**0.844**	0.869	0.026	0.826	**0.935**
Mucocele	0.052	0.599	0.827	0.049	**0.752**	0.841	**0.048**	0.695	**0.846**
Palatal papillomatosis	0.008	0.890	0.996	0.011	**0.918**	0.974	**0.008**	0.891	**0.997**
Squamous cell carcinoma	0.126	0.596	0.885	**0.117**	**0.696**	0.850	0.123	0.642	**0.924**
Lesions overall	0.339	0.734	0.902	**0.327**	**0.787**	0.893	0.330	0.755	**0.914**
Anatomical sites									
Buccal-mucosa							0.034	0.853	0.923
Floor-of-mouth							0.017	0.769	0.947
Gingiva							0.043	0.870	0.939
Inner-lip							0.028	0.910	0.983
Outer-lip							0.019	0.920	0.951
Palate							0.025	0.893	0.985
Tongue							0.101	0.949	0.987
Sites overall							0.136	0.916	0.972

Table 1. Comparison of baseline and our methods. Best metrics in each row is highlighted. GAIN=the model with classification and guidance streams, GAIN+ASP = model with additional anatomical site classification stream, BS = Brier Score, BA = Balanced accuracy, AUC = Area under ROC curve

Metric (↑)	Baseline	GAIN	GAIN+ASP
IoU	0.2454	0.2582	**0.2614**
DSC	0.3470	0.3594	**0.3647**

Table 2. Metrics for interpretability. Significant values are in bold. IoU = intersection over union, DSC = dice-sørensen coefficient

improvement in IoU compared to the baseline model, suggesting improved agreement. Similarly, the DSC, which measures the similarity between the attention maps and annotations, showed a 5% relative enhancement with the GAIN+ASP method compared to the baseline. We note that these improvements were also observed in the GAIN model, which employs guided attention alone. Improving the overall agreement between the generated attention maps and ground truth segmentation maps may not fully capture the dimensions of performance gain. It is particularly important for the guidance approach to enhance the attention on those test images where the baseline model performs poorly. As shown in Tables 2 and 3 , our models, GAIN and GAIN+ASP, significantly improved the IoU for test images falling into the lowest quartile of attention scores in the baseline model, with relative improvements exceeding 100%. The gradient of relative improvement across quartiles at baseline may indicate the inherent performance limitations of the baseline model. However, given that our primary task is image-level classification rather than segmentation of oral lesions, the implications of these improvements in attention are only evaluated as a secondary outcome.

Furthermore, the visual comparison of attention maps between the baseline and GAIN, as shown in Fig. 1, reveals the improved alignment of attention with ground truth regions in our approach. The attention maps clearly indicate the regions of interest that the model focused on when making predictions. These regions aligned well with known characteristics and diagnostic indicators of the respective oral lesions, enabling clinicians to gain insights into the model's decision-making process and identify key features contributing to oral lesion

	Average IoU (% change from baseline) (↑)		
Quartiles of IoU (Based on baseline)	Baseline	GAIN	GAIN+ASP
First	0.015	0.042 (+ 182.6%)	0.040 (+ 171.9%)
Second	0.106	0.121 (+ 14.4%)	0.132 (+ 24.7 %)
Third	0.293	0.298 (+ 1.6%)	0.301 (+ 2.8%)
Fourth	0.567	0.571 (+ 0.7%)	0.571 (+ 0.7%)

Table 3. Improvement in IoU with GAIN and GAIN+ASP models by quartiles of IoU of the baseline model.

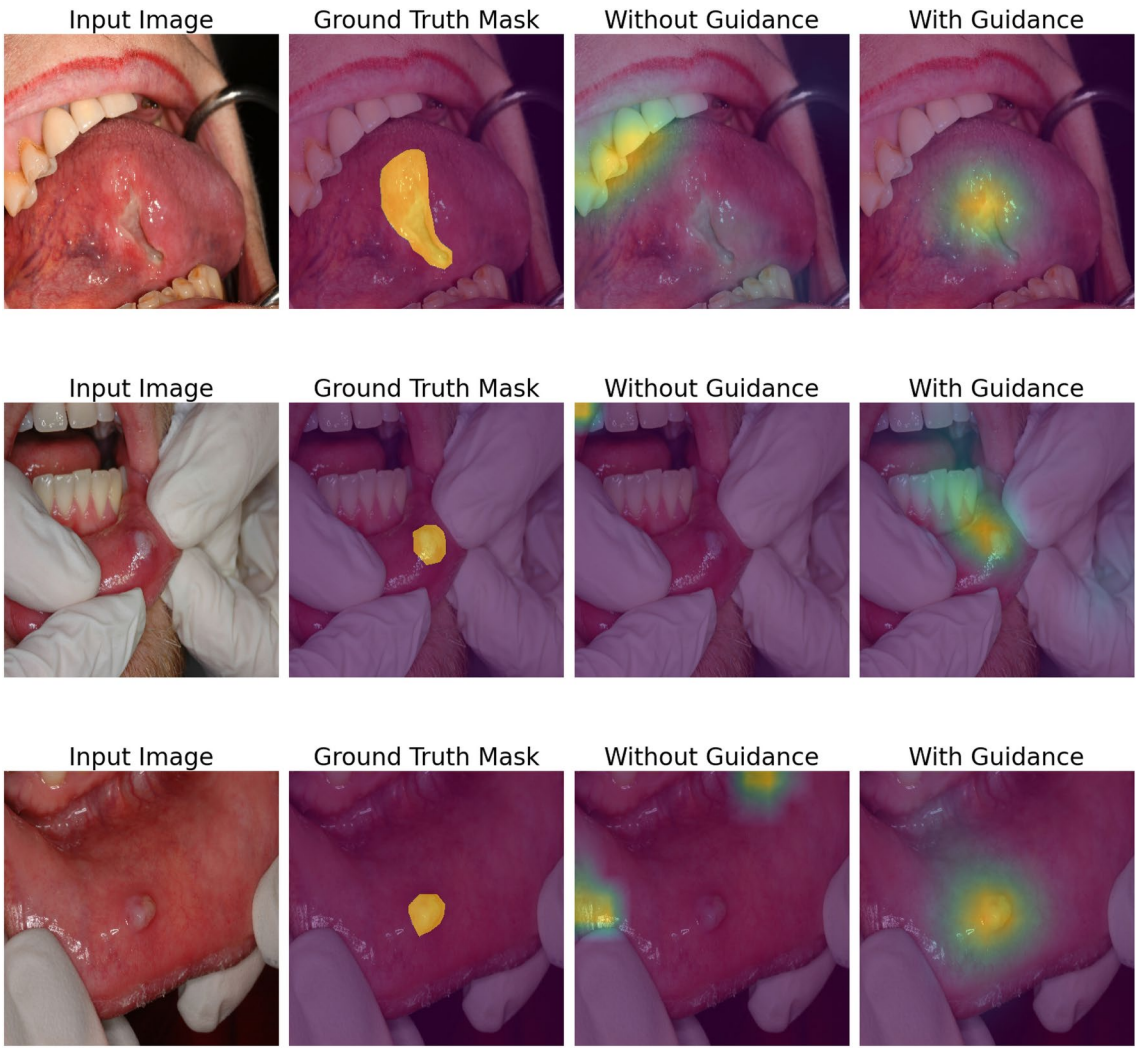

Fig. 1. Comparison of attention maps with and without guidance. We observe that the attention maps are significantly improved with our method compared to the baseline.

classification. This interpretability aspect is crucial for building trust and confidence in the model's predictions, thereby making it more suitable for integration into clinical practice.

Mitigating dataset bias: prediction of implausible combinations (true zeros)

We quantified the number of implausible combinations of oral lesion class and anatomical site (based on ground truth labels) predicted by each model, as shown in Table 4. By definition, all implausible combination predictions are incorrect classifications. The GAIN model reduced the number of implausible predictions to 62, a significant improvement from the 109 seen in the baseline model, reflecting a 43% relative improvement. Whereas, the GAIN+ASP only reduced the number of implausible predictions marginally compared to the baseline. Albeit surprising, this increase in number of implausible predictions with the addition of anatomical site classification

Baseline	GAIN		GAIN+ASP	
	Plaussible (↑)	Implaussible (↓)	Plaussible (↑)	Implaussible (↓)
Plaussible (↑) 579 (84.2%)	567	**12**	543	36
Implaussible (↓) 109 (15.8%)	59	**50**	41	68
688 (100%)	626 (91.0%)	**62 (9.0%)**	584 (84.9%)	**104 (15.1%)**

Table 4. Number of implausible lesion predictions for different methods. Our methods with GAIN model effectively reduces the implausible predictions compared to baseline. Whereas, the GAIN+ASP model shows marginal improvement over baseline model. Significant values are in bold.

	Baseline	GAIN	GAIN+ASP
Correct preds (↑)	3/15	8/15	**9/15**

Table 5. Number of correct predictions of Mucocele lesion type on Outer-lip anatomical site with different methods. Our methods with GAIN and ASP effectively minimizes the influence of dataset bias. Significant values are in bold.

task to the GAIN model could be attributed to the shared parameters of the CNN feature extractor backbone and the approximately similar contributions of the loss functions L_cl and L_{as} as identified by the optimal hyper-parameter values of α and γ in the combined loss function, L_{total} (see Appendix Tables A2 and A3).

Mitigating dataset bias: induced data zeros

To evaluate our models for robustness against dataset bias, we removed all images with ground truth diagnosis of Mucocele occurring at the outer lip site from the training set, effectively introducing *Data Zero* combination as described in methods section. We then evaluated the models' ability to mitigate this bias by examining the number of correctly predicted Mucocele lesions at the outer lip site in the test dataset (Table 5). Our GAIN+ASP model showed significant improvement, correctly predicting 9 out of 15 images, compared to the baseline model's 3 out of 15 correct predictions. Similarly, the GAIN model, without anatomical site prediction, also showed similar ability to mitigate dataset bias by correctly predicting 8 out of 15 images.

In summary, our experimental results validate the effectiveness of our proposed methodology for oral lesion classification. Our approach outperforms the baseline model in terms of accuracy, discrimination ability, and calibration. Additionally, the generated attention maps exhibit better alignment with the ground truth annotations, enhancing interpretability. These findings highlight the benefits of integrating additional guidance loss and anatomical site prediction in our methodology, resulting in improved performance, interpretability and robustness against dataset-bias compared to the baseline model.

Related work

Rokshad et al. recently reviewed the literature on previous attempts on oral lesion classification from intra-oral images using deep learning approaches[27]. Majority of primary studies identified in that review focused on binary classification task (e.g., suspicious vs non-suspicious[46]) or broad multi-class classification (e.g, ulcer vs papule vs macule[23], normal vs benign vs premalignant vs malignant[47]). Although, these approaches might have some clinical value, fine-grained multi-class classification as clinical diagnosis level has higher clinical utility. In our work we have focused on predicting 16 histopatholgically confirmed diagnosis classes. Moreover, a key limitation of all most all of the previous work in this area is the lack of investigation of explainability methods.

Comparable to the work presented here, Figueroa et al. implemented a guided attention network, with minimal adaptation to the original GAIN approach, to classify intra-oral images into suspicious and non-suspicious lesions[48]. In addition to the difference in fundamental task of binary classification and use of human labelled ground truth, in this work, we introduced anatomical site classification task in addition to the diagnostic classification task. Our adaptation was particularly motivated to mitigate dataset bias, which the previous work did not explore.

Welikala et al. on the other had used a soft attention approach with the attention weights estimated from specific layers of the architecture rather than a shared parameter approach of GAIN and did not investigate the impact of their approach on dataset bias[46]. Moreover, both of these previous approaches reported only marginal improvement in classification accuracy, whereas our approach shows significant improvement in classification accuracy and explainability with the added advantage of mitigation of dataset bias.

Conclusion

In this paper, we propose a deep learning-based approach for oral lesion classification that emphasizes interpretability and tackles dataset bias. Our methodology, utilizing the Guided Attention Inference Network (GAIN) framework, generates attention maps to highlight important regions in oral cavity images and promote trust between the model and healthcare professionals. By incorporating guidance loss and pixel-level annotations, our approach mitigates dataset bias and enhances generalization. Experimental results demonstrate

the effectiveness of our approach, outperforming baselines in accuracy and interpretability, with robustness to dataset bias.

Our research has significant implications for clinical practice, aiding clinicians in making informed decisions, fostering trust in automated systems, and improving patient outcomes. Future research can explore advanced interpretability techniques and incorporate multi-modal information to further improve accuracy and clinical utility. Overall, our work contributes to accurate and interpretable oral lesion classification, advancing patient care in oral healthcare.

Methods

Our method integrates three key components: (i) classification stream, (ii) guidance stream, and (iii) anatomical site prediction stream.

Classification stream

The classification stream is analogous to a typical multiclass classification using CNNs. In the classification stream, the network is trained using the oral cavity images, where each image is labeled with one of the 16 types of different lesions. During training, the network learns to extract relevant features from the input images and optimize its parameters to minimize the classification loss; typically computed using a suitable loss function, such as cross-entropy loss.

Guidance stream

The guidance stream is introduced to provide additional guidance to the network about the classification task. This guidance stream, denoted as S_g, utilizes external guidance, such as pixel-level segmentation masks, to guide the network's attention more precisely. By leveraging pixel-level labels, we are able to provide detailed guidance to the network regarding the specific regions of interest in the oral cavity images. This stream operates in parallel with the *Classification Stream* and shares the same network parameters (Fig. 2). The first step in this stream is to generate the trainable attention maps to influence the network's learning. The attention maps reflect the areas in the oral cavity images which contributes to the network's final prediction. These attention maps can be utilized further during training in order to promote the network to focus it's attention on region of interest in the image. By training the network to focus on the relevant regions, we aim to improve the overall classification performance and interpretability of the network's decision-making process as well as mitigate the dataset bias for improved generalization.

To generate the attention maps, we leverage the intermediate feature maps obtained from the *Classification Stream*. These attention maps are generated using the fundamental framework of Grad-CAM (Gradient-weighted Class Activation Mapping)[49]. Given an input image, we obtain the activations of the units in a specific convolutional layer, denoted as $f_{l,k}$, where l represents the layer index and k represents the unit index. For each class c corresponding to the ground-truth label, we calculate the gradient of the class score $\hat{y}^c$ with respect to the activations $f_{l,k}$. These gradients are then subjected to global average pooling to derive the neuron importance weights $w^c_{l,k}$.

$$w_c^{l,k} = \mathrm{GAP}\left(\frac{\partial \hat{y}^c}{\partial f_{l,k}}\right)$$

where GAP$(\cdot)$ represents the global average pooling operation. These importance weights reflect the contribution of each activation map to the prediction of class c. We then apply a 2D convolution operation on the activations f_l using w_c as kernel weights, which effectively generates map with a combination of activations. Finally, the ReLU operation is applied to obtain the attention map A^c as shown in the equation.

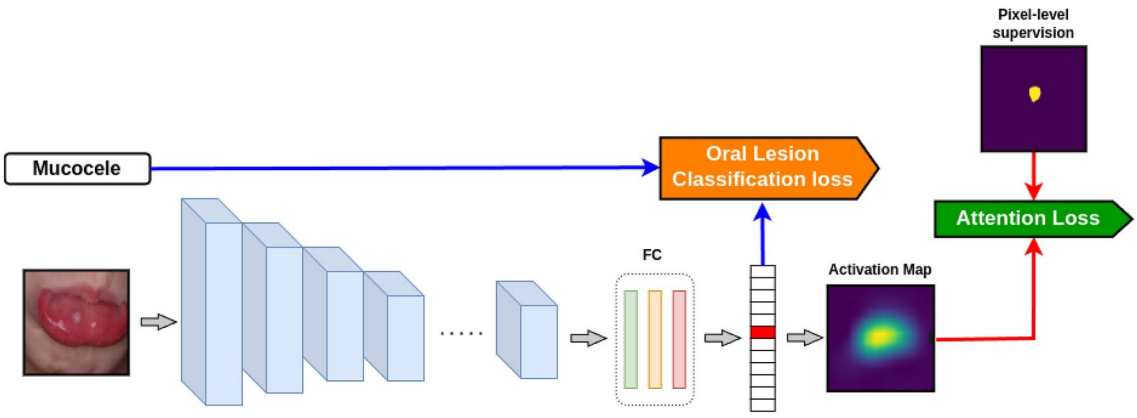

Fig. 2. Visual representation of our approach with both the classification and guidance streams, denoted as GAIN.

$$A^c = \mathrm{ReLU}\left(\mathrm{conv}(f_l, w_c)\right)$$

The resulting attention map A^c represents the regions in the image that contribute most to the network's prediction of class c. Our hypothesis is that the end-to-end trainable attention map A^c influences the network's learning by effectively focusing on relevant regions, thereby making predictions more plausible within the area of interest. These attention maps, combined with pixel-level annotations, provide supplementary information that guides the network in accurately identifying and understanding the characteristics of oral lesions. The discrepancy between the attention maps and the provided pixel-level annotations is calculated by the guidance loss, denoted as L_g. This loss encourages the attention maps to align with the precise regions of interest defined by the external supervision (e.g. pixel-level segmentation masks), leading to more accurate and detailed attention maps. We define the guidance loss L_g as:

$$L_g = \frac{1}{n}\sum (A^c - H^c)^2$$

where A^c is the generated attention map and the H^c represents the pixel-level mask highlighting the region of interest.

Anatomical site prediction stream

Expanding upon our approach, we introduce an auxiliary anatomical site prediction task in the pipeline. Illustrated in Fig. 3, this extension involves the integration of the anatomical site classification task alongside the existing lesion classification stream and the guidance stream with attention loss. Notably, all three streams utilize a common set of CNN features. However, a new MLP head is introduced specifically for the anatomical site prediction task. The neural network is tasked with predicting the precise anatomical location of the lesion, selecting from a set of eight potential locations. To achieve this, we employ a cross-entropy loss function. It is important to highlight that this multifaceted approach aims to maximize the benefits derived from shared features, thus enhancing the model's performance that comprehensively addresses both lesion classification and anatomical site prediction tasks.

Our hypothesis posits a fundamental correlation between anatomical sites and lesion types (see the *Dataset bias* section below), asserting that the ability to learn these correlation structures may enhance the performance of the model in the lesion prediction task. This premise builds on the notion that recognizing the likelihood of certain lesions appearing in specific anatomical areas can improve the model's ability to identify those lesions.

The model, by recognizing the anatomical context, gains a richer set of features that are inherently valuable for distinguishing between different types of lesions. This is not a unidirectional benefit; the insight gained from predicting the anatomical site feeds back into the lesion prediction task. Consequently, understanding where

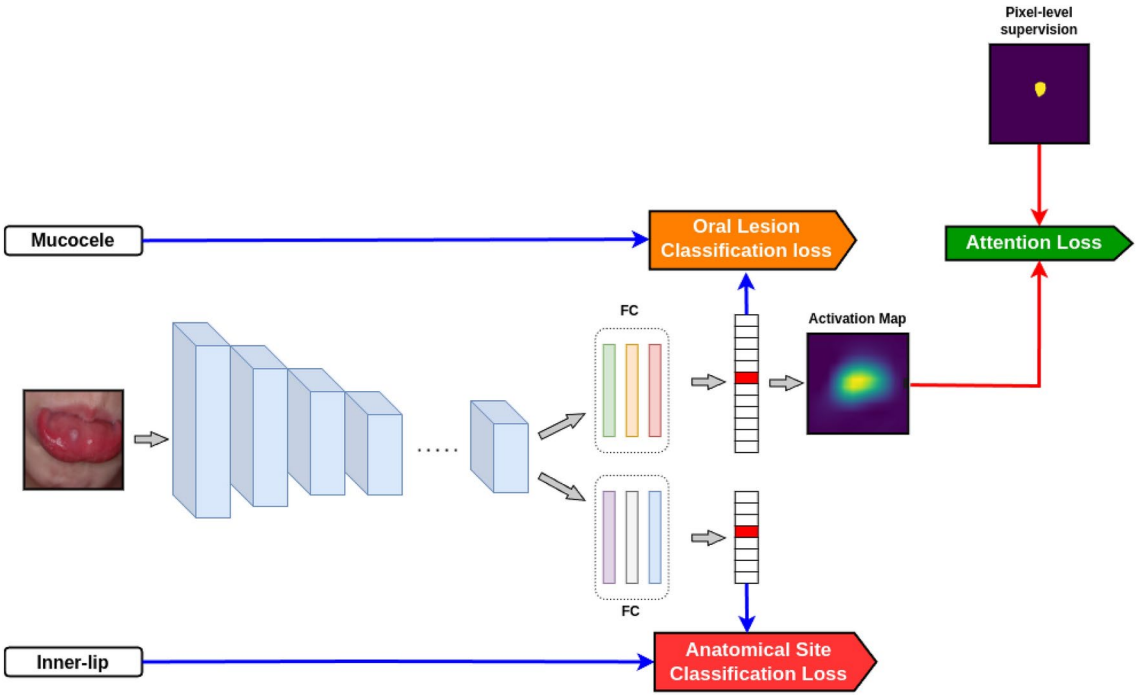

Fig. 3. Visual Representation of our approach that integrates the classification, guidance, and anatomical site prediction streams, denoted as GAIN+ASP.

a lesion is likely to occur can provide additional learning signal to the network during training. This dual-task learning framework not only improves the model's predictive accuracy but also its interpretability. When the model predicts a lesion, it also provides the anatomical context, offering a more comprehensive understanding of its prediction.

Combining the streams

We bring together the three streams by optimizing a finely-tuned loss function, denoted as L_{total}. This function harmonizes the insights drawn from each stream by combining the classification loss (L_{cl}) with the guidance loss (L_g), and the anatomical site classification loss (L_{as}). This combined loss function is given as:

$$L_{total} = \alpha L_{cl} + \beta L_g + \gamma L_{as}$$

Here, α, β, and γ serve as weighting parameters that adjust the contribution of each loss component to the total loss. By optimizing the combined loss function, the network learns to enhance its focus on input images, incorporating both the discriminative features captured by the classification stream and the refined attention maps from the guidance stream. This joint optimization process not only sharpens the model's focus on relevant image regions but also equips it with an understanding of the intricacies involved in oral lesion identification, thereby reducing dataset bias and enhancing overall predictive performance. The detailed algorithm of our method is presented in the appendix and implementation in `pytorch` is provided https://gist.github.com/MadathilSA/48ca84378e393b7737311d18d1c0b5aa.

Dataset and pre-processing

Data source

We established a retrospective cohort of patients who consulted with the oral pathologist in our team (PC) from 1999 to 2021. During this period, 2,765 images across 16 oral lesion classes were compiled from 1079 patients referred to the oral pathologist. Dental experts in the team, meticulously cleaned the dataset for quality by removing images with invisible lesions or multiple lesions in the same image, resulting in 1,888 typical images of 16 oral lesion types (Fig. 4a).

The images were collected by oral pathologists as part of the routine clinical examination of patients with oral lesions and were included in their electronic health records. The images were captured using digital cameras of varying quality under dental clinic lighting conditions. Ground truth labels for the oral lesion types were obtained from histopathology reports (the gold-standard diagnosis) for lesions requiring a biopsy (e.g., pre-malignant lesions) and from expert clinician diagnoses for lesions not requiring a biopsy. Two dentists annotated the oral lesions, generating the ground truth labels for the segmentation masks.

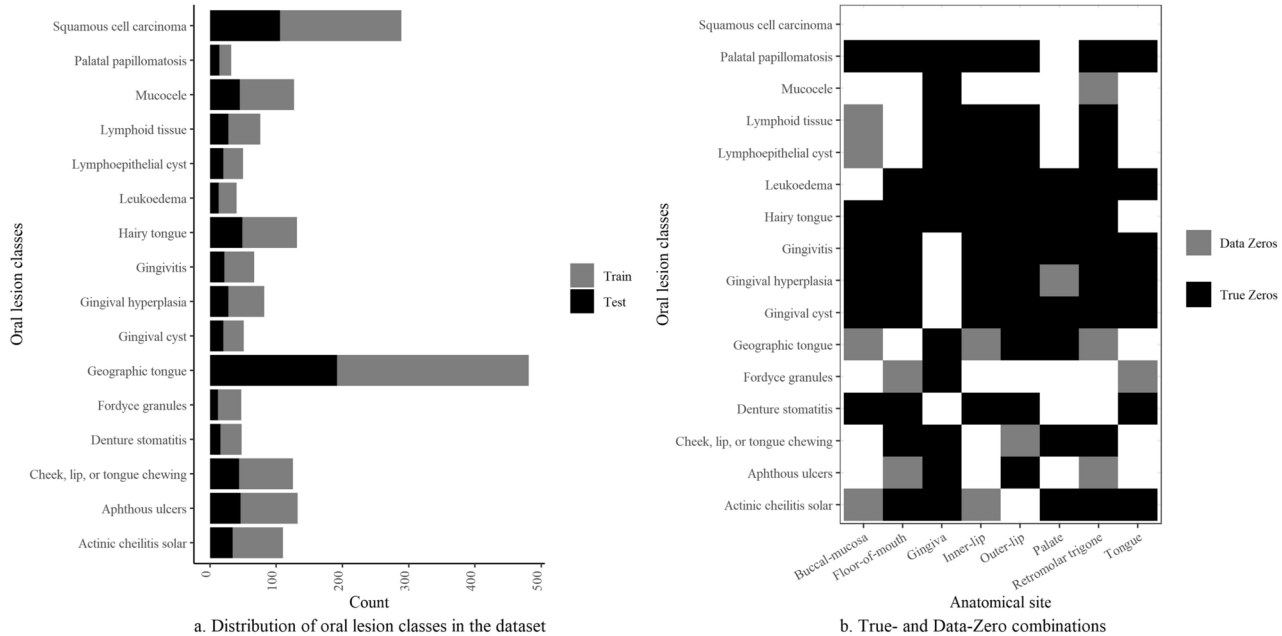

Fig. 4. Dataset characteristics. (**a**) Illustrates the distribution of oral lesion classes, with geographic tongue being the most prevalent. (**b**) Presents the data zero and true zero combinations. The grey-shaded cells (data zeros) indicate plausible combinations for which no data is available in our dataset, the black-shaded cells (true zeros) represent implausible combinations, and the white cells represents plausible combinations where data is available.

All steps of this research were conducted in accordance with the Declaration of Helsinki. No identifiable information or images are presented in this manuscripts. An informed consent was obtained from all participants during the data collection process. The project received approval from the Research Ethics Review Board at McGill University [Protocol no: A07-M40-21B].

Class imbalance

Class imbalance is a prevalent challenge in oral lesion classification tasks, where certain lesion types occur infrequently[50]. Figure 4a illustrates the distribution of oral lesion classes in our dataset. To address the impact of class imbalance, we employed class weighting during the training process[51]. Class weights were calculated as the inverse of the class frequencies in the training dataset. Lesion types with lower occurrences were assigned higher weights, while lesion types with higher occurrences received lower weights. This approach aimed to balance the influence of each lesion type during the training process, giving more emphasis to the underrepresented classes. By assigning appropriate class weights, we aimed to alleviate the adverse effects of class imbalance and improve the overall performance of our oral lesion classification model.

Dataset bias

The classification of oral lesions from intra-oral images presents unique challenges, notably due to the variability of lesion incidence based on anatomical sites within the oral cavity. Certain lesions are exclusive to specific sites, while others may never occur in those areas. For instance, Hairy Tongue is condition that only appears on the tongue's surface and is implausible on other sites such as the buccal mucosa. We define such implausible combinations as *True Zero* combinations. The oral pathology expert in the team was consulted in creating the *True Zero* combination based on the expert evidence synthesis. On the other hand, oral squamous cell carcinoma (OSCC), a rare lesion, can theoretically occur at any anatomical site within the oral cavity. However, despite being plausible, there may be no training images available for this lesion type at certain locations. We refer to these plausible but absent combinations in the training data as *Data Zero combinations*. Figure 4b illustrates the *True Zero* and *Data Zero* combinations of lesions and anatomical sites pertinent to this study.

Additionally, the limited accessibility to certain anatomical sites (e.g., the palate) means that images of lesions in these areas can only be captured in specific ways. These factors contribute to a high correlation between anatomical sites and lesion types, leading to potential bias in CNN models. While datasets that are heterogeneous, exhaustive, and contain a sufficiently large number of images for each unique combination of anatomical site and oral lesion type may help mitigate this bias, real-world datasets often make it challenging for models to distinguish between *Data Zero* and *True Zero* combinations. Our group has previously developed a method to incorporate external knowledge of implausible combinations by augmenting the loss function of CNN models[45]. In contrast, this study explores an alternative approach, aiding the model in learning the correlation structure through guided attention to the lesion's location and supervised learning of the lesion's anatomical site.

Data leakage

A common challenge in intra-oral image classification tasks is data leakage, which occurs when the algorithm learns using information from the data that should not be utilized for the task at hand, such as the shape of teeth[52]. This issue is further exacerbated by the use of multiple angles of the same lesion, leading to an illusion of high performance. To avoid data leakage, we first created clusters of images representing the same oral lesion using the EXIF data, patient IDs, lesion types, and anatomical sites. These clusters were then randomly assigned to either training or test sets, resulting in a training set of 1,200 images and a test set of 688 images (Fig. 4a). This approach helps ensure that the model does not gain an unfair advantage by seeing the same lesion in both training and testing phases.

Data augmentation

To improve the diversity of our oral lesions dataset, we employed a range of data augmentation techniques. These included random rotations within a certain range, horizontal and vertical flips, random translations, and adjustments in brightness and contrast, all while maintaining the aspect ratio of the images. These transformations were meticulously designed to preserve the intrinsic characteristics of oral cavity images, ensuring that the shape and size of the lesions remained unaltered. This careful approach allowed us to create a more robust dataset, suitable for training our models.

Experimental and training setup

We conducted a comprehensive series of experiments to evaluate the effectiveness of our approach in three key areas: (i) accurately classifying oral lesions, (ii) improving interpretability through generated attention maps, and (iii) mitigating dataset bias.

Models

We employed a transfer learning approach with EfficientNet-B5 (EB5) model as the backbone architecture[53]. Initially, a baseline model was established by fine-tuning the EB5 model pre-trained on ImageNet-1K[54], incorporating only the *Classification stream*. Building upon this, we introduced the guidance loss through the *Guidance stream* resulting in `(GAIN)` model. Subsequently, we added the anatomical site classification loss via the *Anatomical site prediction stream*, yielding the `(GAIN+ASP)` model. Except for the hyper-parameter tuning procedure detailed below, all other aspects of fine-tuning were kept consistent across the three model variants. By comparing the performance of our approach with the baseline model, we aim to demonstrate the additional benefits and improvements achieved by the integration of guidance stream, which trains the network's attention

to enhance interpretability and mitigate dataset bias. The inclusion of anatomical site prediction further enhances interpretability and reduces dataset bias.

Hyper-parameter optimization

Our method requires three additional hyper-parameters to be optimized for the combined loss function. We utilized Bayesian Optimization, implemented via BoTorch[55] on the Adaptive Experimentation Platform[56], to tune these parameters. Specifically, we kept the hyper-parameters for EfficientNet-B5 and AdamW Optimizer[57] at their recommended settings and focused the optimization process on our method's hyper-parameters. Details of the optimization procedure and the optimal hyper-parameters identified are provided in Appendix. We fine-tuned all model variants on our dataset, adjusting weights of all the layers to optimize performance using appropriate loss functions and the optimal parameters identified.

Metrics

Our choice of evaluation metrics was guided by the recent 'Metrics Reloaded' framework from[58]. This framework recommends a problem driven approach that enhances the user-centric application of the model. Based on to this framework, we have a combination of metrics for image-level classification (classifying the image into one of 16 classes) and object localization (network's attention on the lesion area). We chose the following metrics for image-level classification: (i) per-class balanced accuracy (BA), (ii) per-class brier score (BS), and (iii) ROC-One-vs-Rest AUC (AUC). To assess the interpretability of the generated attention maps, we utilized recommended segmentation metrics: (i) Intersection over Union (IoU), and (ii) Dice-Sørensen coefficient (DSC). These metrics collectively ensure a comprehensive evaluation of our model's performance and interpretability.

Mitigating dataset bias

To evaluate our method's ability to mitigate dataset bias, we conducted two ablation studies focusing on *True Zero* and *Data Zero* combinations. First, we quantified the implausible lesion predictions made by each model. A prediction is considered implausible if the predicted lesion type, given the ground truth anatomical site, contradicts established anatomical norms (Fig. 4b).

Second, we introduced a category of *Data Zero* combination by removing a specific combination of anatomical site and oral lesion type from the training dataset (specifically, Mucocele lesions occurring on outer-lip). We then retrained all model variants from scratch on this modified dataset. The subsequent evaluation focused exclusively on the models' performance on this particular lesion type. The rationale behind this approach is that if a model is influenced by dataset bias, its performance would deteriorate when encountering certain lesion types in anatomical locations it hasn't been trained on. By conducting these studies, we aimed to assess the robustness of our models and their ability to generalize beyond the biases inherent in the training data.

Data availability

Due to the limited access to the oral cavity, images of specific anatomical locations (e.g. palate) could only be taken in a limited number of ways. Further, some images may contain identifiable patient information (e.g., parts of eyes, skin color, dentition), preventing us from providing dataset publicly. However, dataset can be made available for reproducing results or collaborative research upon request to the corresponding author.

Received: 10 September 2024; Accepted: 28 November 2024
Published online: 30 December 2024

References

1. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024).
2. Marcu, L. & Yeoh, E. A review of risk factors and genetic alterations in head and neck carcinogenesis and implications for current and future approaches to treatment. *J. Cancer Res. Clin. Oncol.* **135**, 1303–14 (2009).
3. Yu, T., Wood, R. & Tenenbaum, H. Delays in diagnosis of head and neck cancers. *J. Can. Dent. Assoc.* **74** (2008).
4. Brinkerhoff, B. T. & Choong, N. W. Diagnosis to treatment interval and outcome in patients with locally-advanced squamous cell carcinoma of the head and neck in a veterans affairs medical center. *J. Cancer Sci. Ther.*[SPACE]https://doi.org/10.4172/1948-5956.1000122 (2012).
5. Hammerlid, E., Silander, E., Harnestam, L. & Sullivan, M. Health-related quality of life three years after diagnosis of head and neck cancer: A longitudinal study. *Head Neck* **23**, 113–25 (2001).
6. Jacobson, J. J. et al. The cost burden of oral, oral pharyngeal, and salivary gland cancers in three groups: Commercial insurance, medicare, and medicaid. *Head Neck Oncol.*[SPACE]https://doi.org/10.1186/1758-3284-4-15 (2012).
7. Sankaranarayanan, R. et al. Effect of screening on oral cancer mortality in Kerala, India: A cluster-randomised controlled trial. *The Lancet* **365**, 1927–33 (2005).
8. Subramanian, S. et al. Cost-effectiveness of oral cancer screening: Results from a cluster randomized controlled trial in india. *Bull. World Health Organ.* **87**, 200–6 (2009).
9. Sankaranarayanan, R. et al. Long term effect of visual screening on oral cancer incidence and mortality in a randomized trial in Kerala, India. *Oral Oncol.* **49**, 314–21 (2013).
10. Olson, C., Burda, B., Beil, T. & Whitlock, E. Screening for oral cancer: A targeted evidence update for the us preventive services task force. *U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews* (2013).
11. Brocklehurst, P., Baker, S. & SR, P. M. A qualitative study examining the experience of primary care dentists in the detection and management of potentially malignant lesions. 1. factors influencing detection and the decision to refer.. *Br. Dent. J.* **208**, 72–3 (2010).
12. Choudhary, A. et al. Dentist's knowledge regarding oral mucosal lesions: Revealing the diagnostic dilemma. *Int. J. Health Allied Sci.* **8**(1), 68 (2019).
13. Ali, M., Joseph, B. & Sundaram, D. Dental students' ability to detect and diagnose oral mucosal lesions. *J. Dent. Educ.* **79**, 140–5 (2015).

14. Lopez-Jornet, P., Camacho-Alonso, F. & Molina-Minano, F. Knowledge and attitudes about oral cancer among dentists in spain. *J. Eval. Clin. Pract.* **16**, 129–33 (2010).
15. Bacci, C., Donolato, L., Stellini, E., Berengo, M. & Valente, M. A comparison between histologic and clinical diagnoses of oral lesions. *Quintessence Int.* **45**, 789–94 (2014).
16. Morgan, R., Tsang, J., Harrington, N. & Fook, L. Survey of hospital doctors' attitudes and knowledge of oral conditions in older patients. *Postgrad. Med. J.* **77**, 392–4 (2001).
17. Bataineh, A. B., Hammad, H. M. & Darweesh, I. A. Attitude toward oral biopsy among general dental practitioners: Awareness and practice. *J. Orofac. Sci.* **7**(1), 19 (2015).
18. Williams, P., Poh, C., Hovan, A., Ng, S. & Rosin, M. Evaluation of a suspicious oral mucosal lesion. *J. Can. Dent. Assoc.* **74**, 275–80 (2008).
19. W, I. Potentially malignant disorders of the oral and oropharyngeal mucosa; present concepts of management.. *Oral Oncol.* **46**, 423–5 (2010).
20. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* **25** (2012).
21. Jubair, F. et al. A novel lightweight deep convolutional neural network for early detection of oral cancer. *Oral Dis.* **28**, 1123–1130 (2022).
22. Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S. & Jantana, P. Performance of deep convolutional neural network for classification and detection of oral potentially malignant disorders in photographic images. *Int. J. Oral Maxillofac. Surg.* **51**, 699–704 (2022).
23. Gomes, R. F. T. et al. Use of artificial intelligence in the classification of elementary oral lesions from clinical images. *Int. J. Environ. Res. Public Health* **20**, 3894 (2023).
24. Welikala, R. A. et al. Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. *IEEE Access* **8**, 132677–132693. https://doi.org/10.1109/ACCESS.2020.3010180 (2020).
25. Tanriver, G., Soluk Tekkesin, M. & Ergen, O. Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders. *Cancers* **13**, 2766 (2021).
26. Shamim, M. Z. M. et al. Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer. *Comput. J.* **65**, 91–104 (2022).
27. Rokhshad, R. et al. Artificial intelligence for classification and detection of oral mucosa lesions on photographs: A systematic review and meta-analysis. *Clin. Oral Invest.*[SPACE]https://doi.org/10.1007/s00784-023-05475-4 (2024).
28. Alabi, R. O. et al. Machine learning in oral squamous cell carcinoma: Current status, clinical concerns and prospects for future-a systematic review. *Artif. Intell. Med.* **115**, 102060 (2021).
29. Salahuddin, Z., Woodruff, H. C., Chatterjee, A. & Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput. Biol. Med.* **140**, 105111 (2022).
30. Alabi, R. O., Almangush, A., Elmusrati, M. & Mäkitie, A. A. Deep machine learning for oral cancer: From precise diagnosis to precision medicine. *Front. Oral Health* **2**, 794248 (2022).
31. Castelvecchi, D. Can we open the black box of AI?. *Nat. News* **538**, 20 (2016).
32. Zhang, Q.-S. & Zhu, S.-C. Visual interpretability for deep learning: A survey. *Front. Inform. Technol. Electron. Eng.* **19**, 27–39 (2018).
33. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* **32**, 18069–18083 (2020).
34. ElShawi, R., Sherif, Y., Al-Mallah, M. & Sakr, S. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Comput. Intell.* **37**, 1633–1650 (2021).
35. Corbella, S., Srinivas, S. & Cabitza, F. Applications of deep learning in dentistry. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **132**, 225–238 (2021).
36. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 1–9 (2019).
37. Wang, F., Casalino, L. P. & Khullar, D. Deep learning in medicine-promise, progress, and challenges. *JAMA Intern. Med.* **179**, 293–294 (2019).
38. Kleppe, A. et al. Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer* **21**, 199–211 (2021).
39. Gerber, M., Pillay, N. & Khammissa, R. A comparative study of supervised and unsupervised neural networks for oral lesion detection. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 01–07 (IEEE, 2021).
40. Li, K., Wu, Z., Peng, K.-C., Ernst, J. & Fu, Y. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9215–9223 (2018).
41. Kattenborn, T., Leitloff, J., Schiefer, F. & Hinz, S. Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS J. Photogr. Remote Sens.* **173**, 24–49 (2021).
42. Venkataramanan, S., Peng, K.-C., Singh, R. V. & Mahalanobis, A. Attention guided anomaly localization in images. In *European Conference on Computer Vision*, 485–503 (Springer, 2020).
43. Tommasi, T., Patricia, N., Caputo, B. & Tuytelaars, T. A deeper look at dataset bias. *Domain adaptation in computer vision applications* 37–55 (2017).
44. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **54**, 1–35 (2021).
45. Besombes, C., Patel, A. & Madathil, S. A. Incorporating expert prior knowledge for oral lesion recognition. In Maughan, K., Liu, R. & Burns, T. F. (eds.) *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023* (OpenReview.net, 2023).
46. Welikala, R. A. *et al.* Clinically guided trainable soft attention for early detection of oral cancer. In Tsapatsoulis, N. *et al.* (eds.) *Computer Analysis of Images and Patterns*, 226–236, (Springer International Publishing, Cham, 2021). https://doi.org/10.1007/978-3-030-89128-2_22
47. Song, B. et al. Mobile-based oral cancer classification for point-of-care screening. *JBO* **26**, 065003. https://doi.org/10.1117/1.JBO.26.6.065003 (2021).
48. Figueroa, K. C. et al. Interpretable deep learning approach for oral cancer classification using guided attention inference network. *J. Biomed. Opt.* **27**, 015001. https://doi.org/10.1117/1.JBO.27.1.015001 (2022).
49. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (2017).
50. Song, B. et al. Classification of imbalanced oral cancer image data from high-risk population. *J. Biomed. Opt.* **26**, 105001–105001 (2021).
51. Barandela, R., Sánchez, J. S., Garcıa, V. & Rangel, E. Strategies for learning in class imbalance problems. *Pattern Recogn.* **36**, 849–851 (2003).
52. Mohammad-Rahimi, H., Rokhshad, R., Bencharit, S., Krois, J. & Schwendicke, F. Deep learning: A primer for dentists and dental researchers. *J. Dent.* **130**, 104430 (2023).
53. Tan, M. & Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. & Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, vol. 97 of *Proceedings of Machine Learning Research*, 6105–6114 (PMLR, 2019).

54. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis. (IJCV)* **115**, 211–252. https://doi.org/10.1007/s11263-015-0816-y (2015).
55. Balandat, M. *et al.* BoTorch: Programmable Bayesian optimization in PyTorch. arxiv e-prints (2019).
56. Bakshy, E. *et al.* Ae: A domain-agnostic platform for adaptive experimentation. In *NeurIPS Systems for ML Workshop* (2018).
57. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019* (OpenReview.net, 2019).
58. Maier-Hein, L. et al. Metrics reloaded: Recommendations for image analysis validation. *Nat. Methods* **21**, 195–212. https://doi.org/10.1038/s41592-023-02151-z (2024).

Acknowledgements
We would like to acknowledge the contributions of Mehak Khanna, and Mohammed Al-Shehri for their contribution toward data cleaning and pre-processing. This project was funded by the Canadian Institue of Health Research Project Grant [PJT-438778]. S Madathil is a recipient of a Career Award from the Fonds de Recherche du Québec-Santé. This research was partly enabled by support provided by Calcul Quebec (calculquebec.ca) and the Digital Research Alliance of Cancer (alliancecan.ca).

Author contributions
Conceptualization: A.P., F.T., M.D., P.C., S.M.; Data curation: A.P., C.B., T.D., M.S., S.M.; Formal analysis: C.B., S.M.; Funding acquisition: F.T., P.C., S.M.; Investigation: A.P.; Methodology: A.P., C.B., S.M.; Project administration: A.P., S.M.; Resources: S.M.; Software: A.P., C.B., S.M.; Supervision: S.M.; Validation: A.P.; Visualization: A.P.; Writing—original draft: A.P., S.M.; Writing—review & editing: A.P., C.B., T.D., M.S., F.T., M.D., P.C., S.M.

Declarations

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/s41598-024-81724-0.

Correspondence and requests for materials should be addressed to S.M.