Automated classification of oral potentially malignant disorders and oral squamous cell carcinoma using a convolutional neural network framework: a cross-sectional study

Cristina Saldivia-Siracusa,[a] Eduardo Santos Carlos de Souza,[b] Arnaldo Vitor Barros da Silva,[c] Anna Luíza Damaceno Araújo,[d,e] Caíque Mariano Pedroso,[a] Tarcília Aparecida da Silva,[f] Maria Sissa Pereira Sant'Ana,[f] Felipe Paiva Fonseca,[f] Hélder Antônio Rebelo Pontes,[g] Marcos G. Quiles,[c] Marcio Ajudarte Lopes,[a] Pablo Agustin Vargas,[a] Syed Ali Khurram,[h] Alexander T. Pearson,[i,j] Mark W. Lingen,[k] Luiz Paulo Kowalski,[d,l] Keith D. Hunter,[m] André Carlos Ponce de Leon Ferreira de Carvalho,[b] and Alan Roger Santos-Silva[a,*]

[a]Departamento de Diagnóstico Oral, Faculdade de Odontologia de Piracicaba, Universidade Estadual de Campinas, Piracicaba, São Paulo, Brazil
[b]Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, São Paulo, Brazil
[c]Institute of Science and Technology, Federal University of São Paulo, São José dos Campos, São Paulo, Brazil
[d]Head and Neck Surgery Department, University of São Paulo Medical School, São Paulo, Brazil
[e]Hospital Israelita Albert Einstein, São Paulo, Brazil
[f]Department of Oral Surgery and Pathology, School of Dentistry, Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil
[g]Service of Oral Pathology, João de Barros Barreto University Hospital, Federal University of Pará, Belém, Brazil
[h]Unit of Oral and Maxillofacial Pathology, School of Clinical Dentistry, University of Sheffield, Sheffield, UK
[i]Section of Haematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA
[j]University of Chicago Comprehensive Cancer Center, Chicago, IL, USA
[k]Department of Pathology, The University of Chicago Medicine, Chicago, IL, USA
[l]Department of Head and Neck Surgery and Otorhinolaryngology, A.C. Camargo Cancer Center, São Paulo, State of São Paulo, Brazil
[m]Liverpool Head and Neck Center, ISMIB, University of Liverpool, Liverpool, UK

Summary

Background Artificial Intelligence (AI) models hold promise as useful tools in healthcare practice. We aimed to develop and assess AI models for automatic classification of oral potentially malignant disorders (OPMD) and oral squamous cell carcinoma (OSCC) clinical images through a Deep Learning (DL) approach, and to explore explainability using Gradient-weighted Class Activation Mapping (Grad-CAM).

Methods This study assessed a dataset of 778 clinical images of OPMD and OSCC, divided into training, model optimization, and internal testing subsets with an 8:1:1 proportion. Transfer learning strategies were applied to pre-train 8 convolutional neural networks (CNN). Performance was evaluated by mean accuracy, precision, recall, specificity, F1-score and area under the receiver operating characteristic (AUROC) values. Grad-CAM qualitative appraisal was performed to assess explainability.

Findings ConvNeXt and MobileNet CNNs showed the best performance. Transfer learning strategies enhanced performance for both algorithms, and the greatest model achieved mean accuracy, precision, recall, F1-score and AUROC of 0.799, 0.837, 0.756, 0.794 and 0.863 during internal testing, respectively. MobileNet displayed the lowest computational cost. Grad-CAM analysis demonstrated discrepancies between the best-performing model and the highest explainability model.

Interpretation ConvNeXt and MobileNet DL models accurately distinguished OSCC from OPMD in clinical photographs taken with different types of image-capture devices. Grad-CAM proved to be an outstanding tool to improve performance interpretation. Obtained results suggest that the adoption of DL models in healthcare could aid in diagnostic assistance and decision-making during clinical practice.

The Lancet Regional Health - Americas 2025;47: 101138
Published Online xxx
https://doi.org/10.1016/j.lana.2025.101138

*Corresponding author. Oral Diagnosis Department, Piracicaba Dental School, State University of Campinas (UNICAMP), Av. Limeira, no 901, Areão, Piracicaba, São Paulo, 13414-903, Brazil.
E-mail address: alan@unicamp.br (A.R. Santos-Silva).

Disclaimer: This summary is available in Portuguese in the Supplementary Material.

Funding This work was supported by FAPESP (2022/13069-8, 2022/07276-0, 2021/14585-7 and 2024/20694-1), CAPES, CNPq (307604/2023-3) and FAPEMIG.

Keywords: Artificial intelligence; Deep learning; Artificial neural network; Oral cancer; Head and neck cancer; Precancerous conditions

Research in context

Evidence before this study

Throughout the last few decades, Computer-Aided Diagnosis (CAD) has emerged as a valuable tool in medical diagnosis. Recognizing the inherent challenges of clinical and histopathological assessment of oral squamous cell carcinoma (OSCC) and oral potentially malignant disorders (OPMD), the integration of computer vision within Artificial Intelligence (AI) technology shows great promise in assisting oral healthcare providers with the screening and diagnosis of these oral lesions. Research based on AI approaches for clinical diagnosis assistance of OPMD and OSCC is currently limited. We searched English, Portuguese and Spanish language publications on this topic in the PubMed database without date restriction and tracked relevant bibliographic references within the found papers up to February 05, 2025, to identify any prior research utilizing AI models for the classification of OPMD and OSCC based on imaging data. We used the following key terms ('artificial intelligence' OR 'machine learning' OR 'deep learning' OR 'convolutional neural network' OR 'artificial neural network') AND (oral) AND ('cancer' OR 'carcinoma' OR 'potentially malignant disorders'). Results from this search revealed that CAD studies evaluating photograph-based data for OPMD and OSCC classification predominantly rely on the assessment of white-light intraoral photographs and fluorescence images. Other studies have focused on non-clinical imaging data inputs, such as histopathological or cytological preparations, cytometry images, or optical coherence tomography, among others. Additionally, non-imaging inputs such as medical records and written information about clinical and histopathological features have also been described. Overall, most reported approaches compare both OPMD and OSCC cases with normal mucosa or nonsuspicious/benign lesions and obtained accuracy values range from 73% to 100%. Notably, only a small proportion of these studies have described the use of explainable artificial intelligence methods as visual resources for result interpretation and explainability—such as Class Activation Mapping (CAM), Attention Rollout or Gradient-weighted Class Activation Mapping (Grad-CAM)—, but no comprehensive appraisal regarding the association between these visual explanations and performance metrics was identified.

Added value of this study

This study offers a thorough exploration of Deep Learning (DL) models in the context of OPMD and OSCC clinical diagnosis. It presents four DL models that accurately differentiate OPMD from OSCC using clinical photographs, while also providing a comprehensive assessment explainability and reasonability of these results through the Grad-CAM technique. This approach effectively addresses the "black box" drawback commonly associated with DL technologies. Furthermore, we explore potential variations in performance between professional and cellphone-captured images by comparing results from different devices, obtaining no statistically significant influence on the classification task. Both appraisals, to our knowledge, have not been previously conducted in the literature. Lastly, we discuss both the capabilities and limitations of DL models in CAD, offering valuable insights for future model development. This new evidence significantly advances the current understanding of AI applications in oral cancer research.

Implications of all the available evidence

Together with previously reported results in related research, our findings support the potential of AI as an innovative, non-invasive tool to assist clinical decision-making related to diagnosis of oral potentially malignant and malignant lesions. Additionally, we uniquely demonstrate the importance of results interpretability for DL models. While subsequent effort is needed to confirm the potential of these tools in uncontrolled, real-world clinical settings, we anticipate that the implementation of these models will have significant beneficial implications in oral healthcare clinical practice.

Introduction

Oral squamous cell carcinoma (OSCC) is a malignant neoplasm arising from the squamous epithelium of the oral mucosa (ICD-10 C00–C06), and it represents the most prevalent form of malignancy in the head and neck region, contributing to high mortality rates.[1] Conversely, oral potentially malignant disorders (OPMD) encompass a group of mucosal conditions (ICD-10 K13) characterized by a variable increased risk of developing OSCC.[2,3]

OSCC and OPMD can share overlapping clinical features. Consequently, oral clinicians face an important challenge when diagnosing these patients, as these two

entities have significantly different outcomes in terms of treatment and prognosis. Both OSCC and OPMD rely on microscopic evaluation as the diagnostic gold standard, a practice linked to intra- and inter-observer variability due to subjectivity in recognizing cytoarchitectural features.[3] This highlights the importance of new approaches for more consistent, efficient and accurate diagnosis of OPMD and OSCC to aid clinical decision-making by prioritizing OSCC referrals, implementing appropriate clinical follow-up protocols, and potentially alleviate the oral cancer burden, improving survival outcomes related to early diagnosis.[4,5]

Recently, interest has grown in leveraging artificial intelligence (AI) to support healthcare providers in diagnosing and understanding diseases.[6,7] Computer-Aided Diagnosis (CAD) is an approach in which AI systems are implemented to assist medical data interpretation, and has proven to be an effective tool for medical diagnosis.[8] Contemporary CAD techniques have relied heavily on Deep Learning (DL), and specifically convolutional neural networks (CNNs), as they have been proved to be particularly effective in recognizing patterns for classification and image-based diagnosis.[9] Nevertheless, challenges in interpretability exist, as CNNs are often seen as "black boxes", hindering understanding of the model's decision-making process. Hence, methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) have been proposed. Grad-CAM is a technique applicable to various CNNs that provides visual representation by highlighting regions of interest in input images during classification, which can be then interpreted by humans to obtain insights into the model's outputs.[10,11]

This study aims to develop and evaluate supervised DL models for automatic classification of OPMD and OSCC patients using clinical photographs. Additionally, we explore associations between performance and image-capturing devices, computational cost and prediction thresholds for reliability. Furthermore, we use Grad-CAM to assess explainability and enhance the interpretability of CNNs' predictions.

In view of this, the hypothesis we intend to test is whether AI models can effectively recognize high-level features and discern fine-grained visual patterns of OPMD and OSCC in clinical images.

Methods

This is a retrospective, cross-sectional study executed with the goal of creating a DL diagnostic assistance model for OPMD and OSCC binary classification.

Ethics committee approval

This study was performed in accordance with the Declaration of Helsinki and approved by the Piracicaba Dental Ethical Committee (registration number: 42235421.9.0000.5418), which also comprised Material Transfer Agreements between co-participant institutions to share clinical and demographic data, digital slides and clinical photographs. Informed consent was obtained from all participants. CLAIM (Checklist for Artificial Intelligence in Medical Imaging (CLAIM) recommendations were followed to report these results.[12]

Dataset

The dataset comprised 851 clinical photographs (taken using either a professional camera or cellphone) from 807 patients diagnosed with OPMD or OSCC between 2005 and 2022 at three Brazilian Institutions' oral medicine services: 758 patients from the Piracicaba Dental School (FOP) (Piracicaba, São Paulo, Brazil), 11 patients from the Federal University of Minas Gerais (UFMG) (Belo Horizonte, Minas Gerais, Brazil), and 38 patients from the Federal University of Pará (UFPA) (Belém, Pará, Brazil). In these oral medicine services, clinical practices to ensure adequate representative sampling of biopsy are strictly followed. Criteria for inclusion comprised of patients with intraoral lesions diagnosed as OPMD or OSCC (according to World Health Organization [WHO], 5ed).[13] Histopathological diagnosis was confirmed by evaluation of correspondent glass slides by the researchers, who are certified oral pathologists (C.S.S, A.L.D.A, A.R.S.S, H.A.R.P, F.P.F and P.A.V). Disagreements regarding diagnosis were resolved by consensus, and if consensus could not be reached, the case would be excluded. As it is the current gold standard for diagnosis, histopathological diagnosis according to the WHO classification for oral epithelial dysplasia and OSCC[13] was used to obtain the reference standard for the desired task. Patients who underwent multiple biopsies at the same anatomic location were each included as independent labels if clinical changes were observed, and new biopsies were conducted at intervals of at least three months. Clinical photographs were excluded if considered non-representative and poor quality (including significant blur, distortions caused by flash, motion artifacts, or cases where the lesion was not clearly visible) (n = 73). To deidentify data and protect health information of the patients, an alphanumeric code, non-related to personal information, was used to name the used images.

Because of the nature of both diseases, study size was determined by the number of available cases. In total, 404 images (52%) were labeled as OPMDs (no dysplasia, mild, moderate and severe oral epithelial dysplasia) and 374 (48%) were labeled as OSCC (microinvasive, frankly invasive/conventional, and verrucous).[13]

Training, validation (model optimization) and testing for image classification

The dataset was divided into training, validation (also known as tuning or model optimization) and internal

Subset	Class	Images (n)			Total
		FOP	UFMG	UFPA	
Training/Validation (90%)	OPMD	341	8	17	366
	OSCC	325	0	8	333
Test (10%)	OPMD	34	1	3	38
	OSCC	41	0	0	41

OPMD, Oral potentially malignant disorders; OSCC, Oral squamous cell carcinoma; FOP, Piracicaba Dental School; UFMG, Federal University of Minas Gerais; UFPA, Federal University of Pará.

Table 1: **Subset image distribution for training/validation and testing.**

testing subsets with 8:1:1 proportion (Table 1). To avoid data leakage and to maintain independence between training and testing data, a non-random, patient-level split was followed, so patients with more than one photograph of the same lesion were maintained in the training subset, and as such, multiple photos from a single patient would only be present in 1 of the 3 sets.

Data preparation and assessment

No image segmentation was conducted, manual or automatic. All the original images were resized to 224 × 224 pixels, to standardize the CNNs input. The images were kept at the RGB color space. Random data augmentation techniques were applied during training, consisting of random translations, rotations, mirroring, and shearing. During training, class weights were used to compensate for the slightly unbalanced class distribution. The training was carried out using hyperparameters detailed in Supplementary Material 1. We utilized the ModelCheckpoint and EarlyStopping callbacks, training for a maximum of 250 epochs. The Adam Optimizer with an initial learning rate of 0.001 was used, optimizing for the cross-entropy loss.

Eight architectures were explored: ConvNeXt,[14] EfficientNet,[15] Inception,[16] MobileNet,[17] ResNet[18] ResNetrs,[19] VGG[20] and Xception.[21] The models and algorithms were implemented using Python 3.10 and several open-source libraries specific to machine learning and image processing (TensorFlow + Keras, Scikit-Learn, and OpenCV). Training, tuning and internal testing using the dataset of interest were performed with each model for the classification task. A binary classification model was adopted, aiming for the models to discern between two classes: OMPD (class 0) and OSCC (class 1). Only the best performing models were reported.

Transfer learning strategies were applied to each CNN to overcome the limitation of using a small dataset. We transferred all convolutional layers from the pretrained models and added new subsequent classifier layers for our task. Initially, only the new, non-transferred layers were fitted while the transferred layers were kept frozen. After which, the entire CNN was fine-tuned at one tenth the learning rate. We performed this strategy using CNNs trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[22] dataset, and the classification dataset for The International Skin Imaging Collaboration 2019 challenge (ISIC 2019).[23–25] Moreover, we also attempted a two-step transfer learning process, where a CNN pretrained on the ILSVRC dataset was also further pretrained for the ISIC 2019 dataset, and later fine-tuned to our dataset. Nonetheless, assessment of CNN's performance without any transfer learning strategy was also done for performance comparative appraisal.

To evaluate performance, we used the mean accuracy, precision, recall, specificity, F1-score metrics and area under the receiver operating characteristics (AUROC) values, together with confusion matrices. Bar graphics were generated to assess the models' confidence in prediction generation for each sensor (professional and cellphone cameras). The computational cost of each CNN was assessed measuring the number of parameters, inference cost and inference time on a standardized hardware setup (CPU only inference, AMD Ryzen 7 3800x–16 cores @ 3.900 GHz, 31997MiB RAM @ 3200 MHz) and a single-board computer (CPU only inference, Raspberry Pi 4 Model B Rev 1.4–4 cores @ 1.800 GHz, 7631MiB RAM @ 3200 MHz).

Threshold assessment for valid predictions

Graphs were generated to evaluate the relationship between model accuracy and pre-defined 0.1 prediction thresholds ranging from 0.1 to 0.9. Results were reported as frequency and percentages.

Type of sensor

A quantitative analysis of the internal testing subset was conducted to explore associations between classification results and the image-capturing device type. Results were reported as frequency and percentages, with Fisher's exact test or Chi-square test applied where appropriate. Evaluation of variables with missing data was performed following a listwise deletion approach. A p-value of ≤ 0.05 was considered significant. All analyses were performed using SPSS version 25 (SPSS Inc., Chicago, USA).

Grad-CAM analysis

To assess interpretability, visual representations were generated for the best-performing CNN models using Grad-CAM, a class-discriminative localization technique which analyzes the gradients of the classification score flowing into the last convolutional layer of the CNN models.[10] This Explainable Artificial Intelligence (XAI) technique allows visualization of the regions that most influenced the predictions (focus points) generated during classification process for all input images in the testing subset, displaying them in a heatmap form.

Two researchers (C.S.S and A.L.D.A) reviewed the Grad-CAM output to compare the models' classification

reliability, splitting them into "trustworthy" if their Grad-CAM highlighted the targeted lesion accordingly, or "untrustworthy" if not. They also categorized them regarding their focus point location, distinguishing between "on-target" (heatmap highlighted approximately ≥80% of the lesion), "off-center" (approximately between 20% and 80%) or "off-target" (approximately ≤20% or less). Lastly, the models' reasonability was judged with an adapted scale from Selvaraju et al., ranging from clearly more/less reasonable (± 2), slightly more/less reasonable (± 1), and equally reasonable (0).[10] Disagreements were solved first by discussion and then by consulting a third author (A.R.S.S).

Funding sources did not have a role regarding study design, collection, analysis, interpretation, writing or decision to submit this paper for publication.

Results

A total of 778 clinical images from 681 patients were included in this study (Table 1). Of these, 378 (55.5%) were men and 303 (44.5%) were women, with a mean and median age of 61.19 and 61 years, respectively (Standard Deviation — 12.95) (range 18–94).

Models' performances

From the eight architectures explored, two CNN had the best results: ConvNeXt and MobileNet. Table 2 displays performance metrics for training/validation and internal test subsets.

During training, the worst results of both networks were generated when pre-training through transfer learning strategies were not executed (models 4 and 8). When transfer learning was conducted, overall performance improved greatly (from 0.661 to 1 AUROC for ConvNeXt, and from 0.838 to 1 for MobileNet), but the most favorable pre-training database differed between models: ConvNeXt obtained its best results when both ILSVRC and ISIC 2019 pre-training was performed (model 2), and MobileNet's highest performance resulted from the implementation of ILSVRC weights (model 5), reaching 100% mean accuracy, precision, recall, specificity, F1-score and AUROC (Table 2).

All models presented decreasing metrics in the independent internal testing subset, as expected. Despite this, the overall test results were solid. ConvNeXt combined with the use of ILSVRC transfer learning strategy performed as the best CNN (model 1), attaining superior performance with a mean accuracy of 0.799 and an AUROC of 0.863. MobileNet + ISIC 2019 (model 7) also had comparable performance, obtaining 0.790 of mean accuracy and 0.825 AUROC. Both CNNs also showed their lowest performance metrics during testing when no transfer learning strategies were performed (model 4 and 8) (Table 2).

Confusion matrices of the performance of the best four models (two of each CNN—models 1, 2, 5 and 7) during internal testing are shown in Fig. 1, revealing an overall efficient rate of correctly identifying both true positives (OSCC) and true negatives (OPMD), while highlighting a false negative (OSCC misclassified as OPMD) of 0.244 as the main limitation of the seemingly best performing model (model 1). ConvNeXt (models 1 and 2) showed a better accomplishment of OSCC classification while MobileNet achieved a slightly superior accomplishment of OPMD classification. Still, confusion matrices show that all models were more accurate in classifying OPMD, indicating that the malignant class is more prone to misclassification. However, upon evaluating the low number of false positives (OPMD misclassified as OSCC), we can infer the high quality of OSCC predictions combined with high precision values (Table 1).

Supplementary Material 2 serves as a general illustration of the models' performance and explores confidence range during the classification task for each sensor. For correct predictions, model 2 seemed to be the most confident for both cellphone and professional cameras, with 100% and 93% classified with a 0.9 confidence rate, respectively. For incorrect predictions, model 5 showed to be the model with the greatest number of cases within a

Model	CNN + Pre-trained weights	Set	Mean accuracy	Precision	Recall	Specificity	F1-score	AUROC
1	ConvNeXt + ILSVRC	Train	0.998	1.0	0.996	1.0	0.998	1.0
		Val	0.885	0.906	0.852	0.918	0.878	0.937
		Test	**0.799**	**0.837**	**0.756**	**0.842**	**0.794**	**0.863**
2	ConvNeXt + ILSVRC and ISIC 2019	Train	1.0	1.0	1.0	1.0	1.0	1.0
		Val	0.860	0.833	0.882	0.837	0.857	0.914
		Test	**0.750**	**0.818**	**0.658**	**0.842**	**0.729**	**0.847**
3	ConvNeXt + ISIC 2019	Train	0.870	0.842	0.892	0.848	0.866	0.937
		Val	0.788	0.771	0.794	0.783	0.782	0.821
		Test	0.758	0.761	0.780	0.736	0.771	0.787
4	ConvNeXt + None	Train	0.632	0.607	0.642	0.623	0.624	0.661
		Val	0.647	0.628	0.647	0.648	0.637	0.587
		Test	0.546	0.571	0.487	0.605	0.526	0.529
5	MobileNet + ILSVRC	Train	1.0	1.0	1.0	1.0	1.0	1.0
		Val	0.829	0.843	0.794	0.864	0.818	0.864
		Test	**0.777**	**0.896**	**0.634**	**0.921**	**0.742**	**0.853**
6	MobileNet + ILSVRC and ISIC 2019	Train	0.858	0.815	0.903	0.814	0.857	0.948
		Val	0.831	0.805	0.852	0.810	0.828	0.871
		Test	0.759	0.775	0.756	0.763	0.765	0.777
7	MobileNet + ISIC 2019	Train	0.885	0.928	0.829	0.942	0.876	0.959
		Val	0.815	0.818	0.794	0.837	0.805	0.881
		Test	**0.790**	**0.928**	**0.634**	**0.947**	**0.753**	**0.825**
8	MobileNet + None	Train	0.738	0.789	0.628	0.848	0.700	0.838
		Val	0.755	0.814	0.647	0.864	0.721	0.810
		Test	0.714	0.821	0.560	0.868	0.666	0.779

CNN, Convolutional neural network; AUROC, area under the receiver operating characteristics; Val, validation; ILSVRC, ImageNet Large Scale Visual Recognition Challenge dataset; ISIC, The International Skin Imaging Collaboration dataset; **In bold**, best performance according to AUROC.

Table 2: **Performance metrics.**

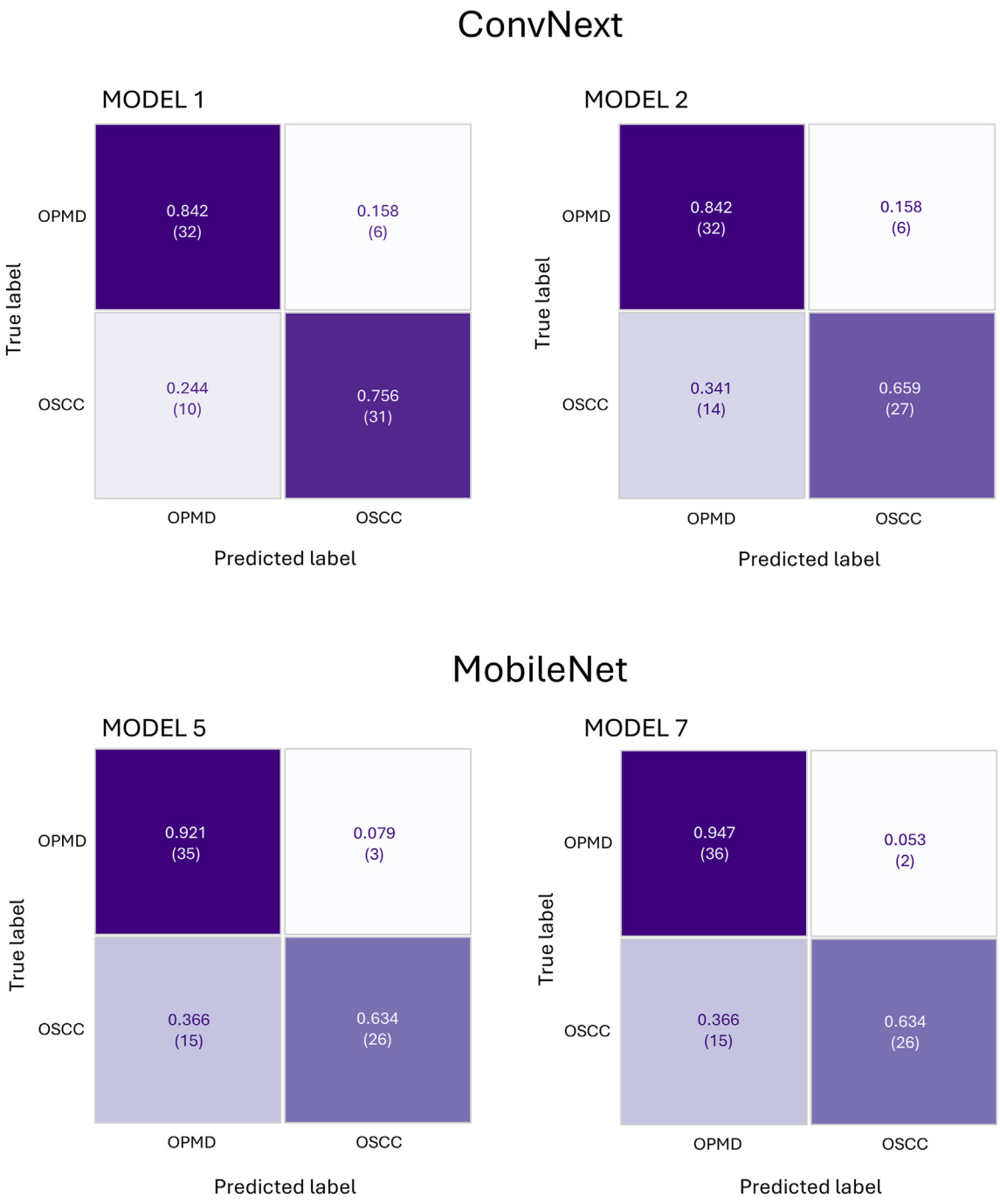

Fig. 1: Internal testing confusion matrices: ConvNeXt and MobileNet. MD: Oral potentially malignant disorders; OSCC, Oral squamous cell carcinoma.

wider range of confidence for cellphone images (0.5–0.9), while model 7 was highlighted having a wide confidence range not only for incorrect professional camera predictions, but also for all correct classification results, suggesting that, for many cases, the model was uncertain about its predictions.

The architecture's computational cost was also measured (Table 3). While MobileNet exhibited higher instantaneous power consumption on the desktop computer (76 W) and comparable power consumption on the single-board computer (4.8 W) to ConvNeXt, its significantly lower mean inference time (0.0072 s on the

Arquitecture	Number of parameters	Desktop computer			Single-board computer		
CNN		Instant power (w)	Mean inference time (s)	Energy per inference	Instant power (w)	Mean inference time (s)	Energy per inference
ConvNeXt	49,456,226	66 W	0.5104 s	9.3573 mW h	4.8 W	3.6356 s	4.8474 mW h
MobileNet	2,998,274	76 W	0.0072 s	0.1520 mW h	4.8 W	0.1121 s	0.1494 mW h

CNN: Convolutional neural network.

Table 3: **Computational cost.**

desktop and 0.1121 s on the single-board) confirms its overall superior efficiency. However, both are inferred in under a second, which is generally acceptable to most applications (Table 3).

Overall, according to performance metrics, the greatest model was ConvNeXt + ILSVRC (model 1), which reached almost perfect training metrics and the best performance in the test subset. MobileNet + ISIC 2019 (model 7) had a slightly lower but comparable performance during testing, while generating lower computational cost.

Threshold assessment for valid predictions

Results from Supplementary Material 3 showed that three of the four best-performing models achieved maximum accuracy based on over 81% of valid predictions when considering a threshold of 0.9 (model 1, 2 and 5), confirming prediction reliability. The model with the highest number of valid predictions at this threshold was Model 2, with 72 (91.1%) valid predictions, followed by Model 1, with 70 (88.6%). Conversely, the model that generated the fewest valid predictions at the highest threshold (0.9) was Model 7, with only 39.2% valid predictions. Overall, ConvNeXt demonstrated superior performance at higher thresholds, whereas MobileNet achieved the highest number of valid predictions when using a threshold of 0.6.

Type of sensor

Within the internal testing subset, a total of 20 (25.31%) images were taken with a cellphone camera and 59 (74.78%) with a professional camera. Distribution between correct and incorrect predictions according to the type of device used is shown in Table 4. Altogether, the models had a range of correct predictions between 80 and 90% when images were taken with cellphone camera, compared to 72.88–76.27% when a professional camera was used. Regarding incorrect predictions, ranges varied from 10 to 20% for cellphone cameras and from 23.73 to 27.12% for professional cameras. The inclusion of cellphone cameras did not seem to negatively influence the prediction rate, as the correct prediction rate was higher for this group, and confidence range was more consistent than for professional cameras when classifying correctly (Supplementary Material 2). Statistical analysis by Fisher's exact test and Pearson's chi-square test revealed no significant association between the type of sensor and the generated predictions (Table 4).

Grad-CAM analysis

To infer the decision-making process of CNN models, the top-4 performing models were assessed using Grad-CAM: ConvNeXt models 1 and 2, and MobileNet models 5 and 7. These images served as graphical representation depicted by color to make the models' decision understandable, indicating the focus points during classification, with blue denoting highest intensity (Fig. 2). When the identification was effective, Grad-CAM representations consistently highlighted relevant areas (Fig. 2a). Yet, deviations were also identified emphasizing broader or more imprecise areas that do not fit the targeted lesion, such as teeth, gloves, oral retractors or other oral and extraoral areas (Fig. 2b). These discrepancies between the performance metrics reached during the classification task and the heatmap representations indicate some unexpected findings in which the models' predictions were not supported by the Grad-CAM illustrations.

Interesting findings were revealed during Grad-CAM reliability evaluation. Despite not having the highest

Type of sensor	Cellphone camera	Professional câmera	p-value
Internal testing subset (n = 79)	**20 (100%)**	**59 (100%)**	
Models			
ConvNeXt + ILSVRC (model 1)			0.333[a]
Correct prediction	18 (90)	45 (76.27)	
Incorrect prediction	2 (10)	14 (23.73)	
ConvNeXt + ILSVRC + ISIC 2019 (model 2)			0.527[b]
Correct prediction	16 (80)	43 (72.88)	
Incorrect prediction	4 (20)	16 (27.12)	
MobileNet + ILSVRC (model 5)			1.000[a]
Correct prediction	16 (80)	45 (76.27)	
Incorrect prediction	4 (20)	14 (23.73)	
MobileNet + ISIC 2019 (model 7)			0.537[a]
Correct prediction	17 (85)	45 (76.27)	
Incorrect prediction	3 (15)	14 (23.73)	

ILSVRC, ImageNet Large Scale Visual Recognition Challenge dataset; ISIC, The International Skin Imaging Collaboration dataset. [a]Fisher's exact test double-sided p-value. [b]Pearson's chi-square test p-value.

Table 4: **Correlation between classification and type of sensor.**

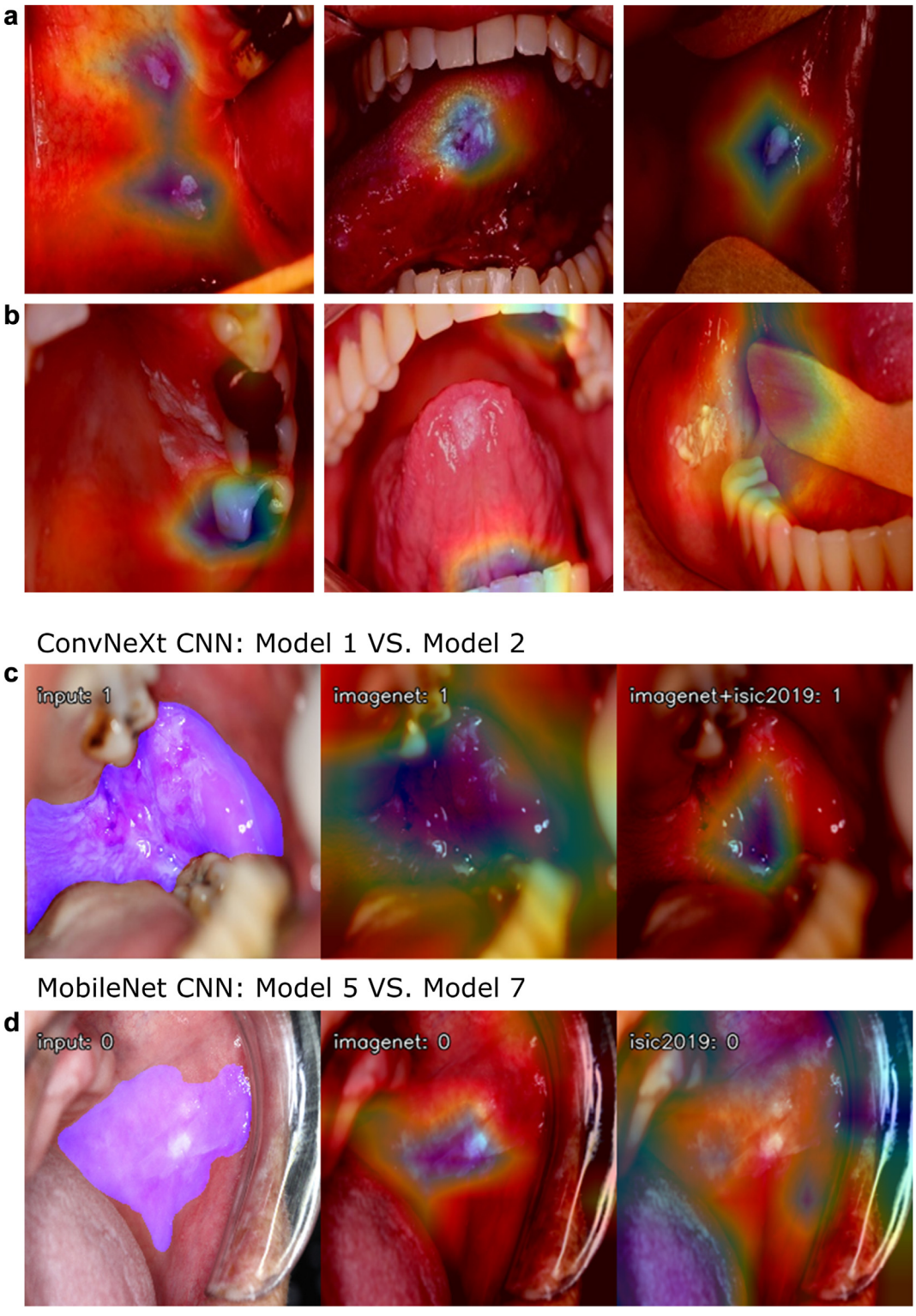

Fig. 2: Gradient-weighted Class Activation Mapping (Grad-CAM) visual representation: **a)** On-target focus points represented by blue highlighting of the lesion of interest; **b)** Off-target focus points, represented by blue highlighted areas outside the lesion of interest, targeting other irrelevant parts of the mouth or other distractors, such as teeth or buccal retractors; **c)** Comparison between ConvNext models 1 (ILSVRC pre-training) and 2 (ILSVRC + ISIC 2019 pre-training), showing correct explainable classification of OSCC (input 1), in which the model 1 achieved a

performance metrics, Model 5 (MobileNet + ILSVRC) showed the highest reliability: 46 cases were considered "trustworthy" in its ability to accurately identify the targeted lesion, opposed to model 1 with 41, model 2 with 36, and model 7 with 33 trustworthy Grad-CAM representations (Table 5). Also, 20 images displayed an "on-target" focus point's location within the heatmap, compared with 17 from model 1, 12 from model 2 and 18 from model 7. The Grad-CAM representations were, in cases, vastly different between models when compared side-by-side, highlighting that the CNN "learns" distinctly despite similar performance. Additionally, the most trustworthy models were based on different architectures but were both pre-trained with ILSVRC, and the application of a two-step transfer learning strategy did not seem to increase the model's capacity to recognize the affected area. As previously mentioned, we confirmed that the visual representation was, sometimes, dissonant with the obtained classification, as we identified that, in models 1 and 2, 29 and 28 cases were classified correctly by the model, respectively, but exhibited an untrustworthy Grad-CAM visual representation. In the case of models 5 and 7, this situation was found in 21 and 36 cases, respectively. Likewise, we also identified a lower number of misclassified cases (range 5–7 cases per model) in which Grad-CAM visualization was accurate regarding lesion localization.

Regarding ConvNeXt, model 1's predictions were categorized as clearly or slightly more reasonable than model 2 in a total of 32 cases against 25. Model 2 often exhibited less precision in generating heatmaps for lesion identification (Fig. 2c). When contrasting focus point location, model 1 images revealed 17 "on-target" and 26 "off-centered" heatmaps (Table 5), a difference that further highlighted the disparity between the two but was consistent with the modest variance in performance metrics.

A different pattern was observed between MobileNet's models 5 and 7: we noted that the model with the best metrics, in this case, was not the most reliable (Fig. 2d). While performance metrics of model 7 were better, suggesting usefulness of ISIC 2019 pre-training, Grad-CAM assessment revealed that the model's focus points were trustworthy in only 33 images, opposed to 46 trustworthy Grad-CAM representations of model 5, which involved the use of ILSVRC weights. According to reasonability assessment, outstanding differences were also noted: 33 Grad-CAM images of the model 5 (ILSVRC pre-trained model) were considered clearly and slightly more reasonable compared to 18 cases from model 7 (ISIC 2029 pre-training). ISIC 2019 pre-training seemed to be useful to increase performance metrics in classification, but this was not equivalent to good explainability and reliability (Table 5).

Altogether, we confirmed through Grad-CAM analysis that the greatest model regarding explainability was MobileNet (model 5), which benefited from a single-step transfer learning strategy with ILSVRC and demonstrated the highest reliability despite lower (but still competitive) performance metrics compared to other models.

Discussion

In the last years, deep convolutional networks have outperformed the state-of-the-art in many visual recognition tasks.[9] AI-based diagnosis research in the context of oral cavity is scarce, especially those using a DL approach through clinical photograph evaluation.[8,26] Through our approach, we explore CNN use for recognition of high-level features in clinical photographs and report four high-performing models for OPMD and OSCC classification comparing performance and computational cost, whilst also displaying and assessing Grad-CAM visualizations, which has superior relevance in this AI era to build not only accurate but "transparent" models to aid decision tasks.[10,11] We anticipate that these models could be impactful in future clinical practice for healthcare professionals' diagnostic assistance, particularly in regions around the world where access to infrastructure, resources and specialized oral medicine services are limited.

In our study, we aimed to achieve OPMD and OSCC classification. Two ConvNeXt-based models showed high and comparable efficacy, albeit employing different fine-tuning strategies during their development. Our results demonstrated the expected significance of pre-training strategies in enhancing learning capacity, as previously reported.[27] In this sense, it is important to note that different transfer learning strategies yielded contrasting results across models, which is an interesting area for further study.

We identified a decrease in performance metrics from the training to testing stages, which was anticipated. The model is expected to reach higher values during training since DL models optimizes itself by "learning from the error", meaning that the CNN updates their weights and bias to optimize the activation function. Since the validation subset is "held out" to avoid data leakage, it is also expected to see lower values in this subset when compared to training, which means

slightly more reasonable prediction; **d)** Comparison between MobileNet models 5 (ILSVRC pre-training) and 7 (ISIC 2019 pre-training), showing a clearly more reasonable on-centered heatmap for OPMD classification, while model 7 attained a correct classification but an off-target heatmap of the lesion of interest. Input 0: OPMD, input 1: OSCC. CNN, Convolutional neural network; OPMD, Oral potentially malignant disorders; OSCC, Oral squamous cell carcinoma; ILSVRC, ImageNet Large Scale Visual Recognition Challenge dataset; ISIC, The International Skin Imaging Collaboration dataset.

	Model 1 ConvNeXt (pre-trained with ILSVRC)	Model 2 ConvNeXt (pre-trained with ILSVRC + ISIC2019)	Model 5 MobileNet (pre-trained with ILSVRC)	Model 7 MobileNet (pre-trained with ISIC 2019)
Internal testing subset (n = 79)	**Number of cases (n)**			
Correctly classified cases	63	59	61	62
Trustworthiness of classification according to grad-CAM representation				
Trustworthy	41	36	46	33
Untrustworthy	38	43	33	46
Focus point's location				
On-target	17	12	20	18
Off-center	26	27	26	17
Off-target	36	40	33	44
Reasonability	Comparison between model 1 and 2		Comparison between model 5 and 7	
Prediction seems clearly more reasonable	24	17	19	7
Prediction seems slightly more reasonable	8	8	14	11
Both models' predictions seem equally reasonable	22		28	

Grad-CAM, Gradient-weighted Class Activation Mapping; ILSVRC, ImageNet Large Scale Visual Recognition Challenge dataset; ISIC, The International Skin Imaging Collaboration dataset.

Table 5: **Gradient-weighted Class Activation Mapping (Grad-CAM) explainability assessment.**

that the CNN can perform well on unseen data. Additionally, the test subset ultimately proves the models' generalization ability.

Our performance results initially indicated that the most promising architecture is ConvNeXt, which is also the newest CNN-based model. Using a classic convolutional network as a base, ConvNeXt incorporates characteristics of both transformers and CNN architecture.[14] However, the relevance of MobileNet must be noted, as it generated comparable results with our best performing network using a much lower computational cost while also attaining greater explainability and reliability. MobileNet CNNs are based on a streamlined architecture that uses depthwise separable convolutions to build light weight deep neural networks.[17] MobileNet was approximately 70 times faster than ConvNeXt on a desktop computer and about 32 times faster on a single-board computer. This difference is due to MobileNet's shallower architecture, which better utilizes the multiple cores of the desktop CPU by enabling greater parallelism. In contrast, ConvNeXt's deeper structure limits parallel processing, reducing its efficiency. On the single-board computer—simulating low-resource equipment such as embedded or mobile devices, the limited processing capacity restricted the potential for MobileNet's parallelism. As a result, both models fully utilized available resources, leading to similar instantaneous power consumption but reduced relative speed advantage for MobileNet. CPU utilization analysis confirmed this: on the desktop, ConvNeXt used 20–60% of all cores, while MobileNet maintained ~70%; on the single-board computer, both models exceeded 90% utilization. Depending on the purpose for which the network is developed (i.e., a screening mobile app), it may be determined that the use of a network with a lower computational cost is more useful, while for a computer program there is no problem using one at a higher cost. Based on this, along with the reported explainability outcomes, we consider that MobileNet, in conjunction with ILSVRC pre-training can be a model to be explored to develop assistance tools for clinical diagnosis.

Previous studies have trained CNN models for OSCC classification tasks.[28–43] A comparative table of related literature is presented in Supplementary Material 4. Higher accuracy for OSCC classification has been previously reported: Warin et al. achieved an AUROC of 0.98–1.00 using DenseNet121 to classify 980 images into "non-pathological", OSCC and OPMD.[43] Similarly, Tanriver et al. classified 684 images into benign, OPMD, and carcinoma categories, reporting an F1-score of 0.858,[39] although, in both studies, no explainability approach was considered. Some previous experiments have been conducted exclusively on tongue lesions,[30,34,38] while others have been based on collected images from online sources,[28,29,33,35,36,38,40] classification of "non-cancer" vs. "cancer" or "suspicious" vs. "not suspicious", and often grouping both OPMD and OSCC in the same category.[30,31,35,40] Various studies included normal mucosa as one of the classes.[32,41,42] Although some of previous work has incorporated XAI techniques for visual representation,[31,32,34,37,42,43] no formal explainability analysis has been reported. In this work, we aimed for early detection of potentially malignant and malignant lesions, so our sample included OPMD and OSCC. This approach closely resembles the substantial challenge of overlapping clinical features faced by oral clinicians during visual examination, as most dentists can

discriminate oral mucosa from pathological alterations, whereas discerning between potentially malignant and malignant lesions can be troublesome. Yet, we recognize that this decision restricted our sample size and understand that including *in-situ* (severe oral epithelial dysplasia) and microinvasive OSCC cases can carry a risk, as these borderline entities and overall dysplasia classification involves greater complexity because of subjective histopathological evaluation. Also, factors such as larger sample size,[17] sociodemographic data analysis, and implementation of other processing tools, can positively influence the results for multimodal DL approach in the future. Nevertheless, regardless of sampling and methodological limitations, we developed a pure classification model that generated competitive outcomes and provided a deep dive in various models' performance and interpretability.

For this study, most images were captured in clinical settings where standardized protocols for intraoral image acquisition are well established. However, a portion of the sample was obtained using cellphone cameras under non-standardized conditions. Utilizing images from different sensors in CNNs has been shown to enhance the generalization of the network's applicability, as the predictions become less dependent on image quality.[44] Implementing a pre-processing stage and employing resizing methods can ensure a certain degree of standardization, facilitating proper task execution.[44] Nevertheless, it is relevant to note that image capture involves many variables beyond quality attributes like size and resolution, including compositional features such as angle, framing and distance. Pictures taken with cellphone cameras or under non-standardized settings could introduce potential intrinsic bias associated with these variables. Since we opted to include photos taken with different devices, we analyzed the association between image-capture devices and classification results to see if it would have a discernible impact on the classification task. In our study, the use of lower-resolution cellphone images does not seem to adversely affect the models' performance. Whilst our results demonstrated no statistically significant influence, the models seemed to perform competently for this group, which can highlight the value of normalization and represent an advantage for future general use. Under subjective appraisal, we could not recognize any pattern within this group of pictures that could justify the higher confidence and prediction rate. Still, given to the limited number of cellphone images in our sample, the hypothesis that the aforementioned attributes could impact the results in a larger dataset should be considered and explored in prospective research.

In CNNs, particularly those used for image classification, understanding which parts of the input image are influential for the network's decision can be challenging. Grad-CAM exploits the spatial information preserved through convolutional layers and uses the feature maps produced by the last layer to understand which parts of an input image were important for a classification decision.[10,45] We conducted a comprehensive assessment of the models' performance and reliability, which has not been previously reported in medical literature using XAI methods. This was useful to critically understand that metrics did not fully represent overall capacity, revealing discrepancies in lesion recognition and classification. Some intriguing correlations can be made through this approach: for example, we identified that the number of false negatives in our best models' confusion matrices varied from 0.244 to 0.366 (Fig. 1), demonstrating low risk of missing an OSCC diagnosis, which would represent the worst-case scenario. In contrast, false positives, which would imply the risk of overtreatment, were notably lower (0.053–0.158). This situation was observed with some exuberant OPMD malignant-passing lesions that ended up being misclassified as OSCC by the CNN, resembling human interpretation. Grad-CAM cannot provide interpretability, as it is a visual representation and does not reveal the reasoning process influencing the results given by the model. Thus, clinicians' assessment is imperative to determine whether the explainability given by the model is suitable or not to the assigned task, and therefore, contribute to relevant insights to enhance AI performance.[45] We also acknowledge the potential for future research focused on XAI methods to deepen comprehension about interpretability.

In the context of oral diseases, the prevalence of OSCC and OPMD is low. In 2022, GLOBOCAN 5-year prevalence of lip and oral cavity cancer was estimated to be 0.0139% (13.9 per 100,000 habitants) (https://gco.iarc.fr/), while a recent meta-analysis reported a worldwide prevalence of 4.47% for OPMD.[2] Conversely, there are a wide range of oral lesions that may share similar clinical features such as color, surface and location, and are highly prevalent. Different from regular clinical practice, our models were trained in a restricted, controlled setting using a carefully selected sample of OPMD and OSCC cases. Therefore, to endorse widespread use of this technology, we acknowledge that the efficacy of these AI models in aiding diagnosis when presented with a broader repertoire of lesions has yet to be proven, which will be investigated in future studies.

Some limitations have been mentioned and must be noted while interpreting our results, such as lack of demographic data analysis including race/ethnicity, limited sample size and the intrinsic nature of subjective histopathological evaluation for the diagnosis of these lesions, especially when assessing borderline lesions such as *in-situ* and microinvasive OSCC. Moreover, due to factors such as intraoperative variables, visual subjectivity and expertise, incisional biopsies carry a risk of providing limited data. We aimed to

minimize these influencing factors by collecting a multicentric sample from reference oral medicine services, ensuring standardized biopsy protocols. Additionally, microscopic assessments were conducted by experienced oral pathologists to achieve histopathological consensus. To further strengthen diagnostic validity, we carefully selected cases and excluded any ambiguous ones. In addition, some influencing factors can affect Grad-CAM representations, such as network size or depth,[27] and we also only assessed the best-performing models, so reinterpretation of some explainability results could be possible when analyzing the remaining models, or by getting deeper into the model's layers. Finally, to help mitigate the lack of external validation, our dataset includes a wide range of imaging conditions, including images from different institutions and devices, which provides an interesting degree of variability and robustness to our results. Nevertheless, we recognize that this does not fully replace independent testing, and we plan to overcome this limitation in future studies.

This comprehensive evaluation offers a more understanding of the DL methods' capabilities and limitations for oral cancer diagnosis. Further studies using segmentation to delimit the lesion by experienced pathologists and stomatologists may be useful to provide more information to the model using other approaches to classification.

Conclusions

DL models proved to accurately distinguish OSCC from OPMD in clinical photographs. The ConvNeXt architecture combined with ILSVRC pre-training strategy, achieved metrics of 0.799 mean accuracy, 0.837 precision, 0.756 recall, 0.842 specificity, 0.794 F1-score and 0.863 AUROC, and obtained good explainability. MobileNet was demonstrated to be a low computational cost alternative with comparable performance results and superior explainability. Grad-CAM technique was proved to be an outstanding tool for performance interpretation, changing the optics on what model performed best, bringing the MobileNet architecture combined with ILSVRC pre-training strategy to the spotlight of the best model according to the heatmaps trustworthiness, focus and reasonability. The use of images captured with different devices did not statistically influence the classification task. This study provides a highly detailed exploration of the capabilities and limitations of DL models in the context of OPMD and OSCC clinical diagnosis.

Contributors

All authors had substantial contributions to this work: Conceptualization (C.S.S, A.L.D.A, E.S.C.S, A.V.B.S, A.C.P.L.F.C, A.R.S.S), methodology and design (C.S.S, A.L.D.A, E.S.C.S, A.V.B.S, A.T.P, M.W.L), data acquisition and curation (C.S.S, A.L.D.A, M.S.P.S, C.M.P, T.A.D, H.A.R.P, F.P.F, P.A.V), formal analysis and interpretation (C.S.S, A.L.D.A, E.S.C.S, A.V.B.S, F.P.F, S.A.K, M.G.Q, A.C.P.L.F.C, M.W.L, A.R.S.S, P.A.V, A.T.P, M.A.L, L.P.K, K.D.H), writing—original draft (C.S.S, A.L.D.A, E.S.C.S, A.V.B.S), and writing—review & editing (H.A.R.P, M.G.Q, M.A.L, P.A.V, S.A.K, M.G.Q, A.T.P, M.W.L, F.P.F, L.P.K, K.D.H, A.C.P.L.F.C, A.R.S.S). A.R.S.S is responsible for the decision to submit this manuscript. All authors gave their final approval and agreed to be accountable for all aspects of the work.

Data sharing statement

The study protocol and individual deidentified data underlining the results reported in this article will be available immediately following publication and ending 5 years following publication for researchers who provide a methodologically sound proposal for any type of analysis. Proposals should be directed to alan@unicamp.br, and data requestors will need to sign a data access agreement. All authors agree to be accountable for any aspects of the work and ensure that questions related to the accuracy or integrity of any part of the work will be appropriately investigated and resolved.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, DeepL Write and ChatGPT tools were used to improve readability. The authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Declaration of interests

S.A.K. declared receiving research support from Cancer Research UK, University of Sheffield; holding an unpaid role in the Swallows Head & Neck Cancer Charity Trustee, UK; and having stock options in Histofy. A.T.P. declared receiving consulting fees, payment or honoraria, and/or support for attending meetings/travel from ThermoFisher, Break-Through Cancer, Prelude, Abbvie, Elevar, and Caris.

Acknowledgements

This study was financed, partly, by the São Paulo Research Foundation—Brazil (FAPESP), process numbers 2022/13069-8 (C.S.S), 2021/14585-7 (A.L.D.A), 2024/20694-1 (A.R.S.S) and 2022/07276-0 (C.M.P); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES), Finance Code 001; Conselho Nacional de Desenvolvimento Científico e Tecnológico—Brazil (CNPq), process 307604/2023-3 (A.R.S.S); and Minas Gerais State Research Foundation—Brazil (FAPEMIG) (F.P.F.). Funding sources were not involved in the writing of the manuscript or the decision to submit it for publication. The corresponding author states that authors were not precluded from accessing data in the study.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.lana.2025.101138.

References

1 Lingen MW, Abt E, Agrawal N, et al. Evidence-based clinical practice guideline for the evaluation of potentially malignant disorders in the oral cavity: a report of the American dental association. *J Am Dent Assoc*. 2017;148:712–727.e10.
2 Mello FW, Melo G, Guerra ENS, Warnakulasuriya S, Garnis C, Rivero ERC. Oral potentially malignant disorders: a scoping review of prognostic biomarkers. *Crit Rev Oncol Hematol*. 2020;153. https://doi.org/10.1016/j.critrevonc.2020.102986.
3 Iocca O, Sollecito TP, Alawi F, et al. Potentially malignant disorders of the oral cavity and oral dysplasia: a systematic review and meta-analysis of malignant transformation rate by subtype. *Head Neck*. 2020;42:539–555.
4 Coelho KR. Challenges of the oral cancer burden in India. *J Cancer Epidemiol*. 2012;2012:701932. https://doi.org/10.1155/2012/701932.
5 García-Pola M, Pons-Fuster E, Suárez-Fernández C, Seoane-Romero J, Romero-Méndez A, López-Jornet P. Role of artificial

intelligence in the early diagnosis of oral cancer. A scoping review. *Cancers (Basel)*. 2021;13. https://doi.org/10.3390/cancers13184600.
6 Moxley-Wyles B, Colling R, Verrill C. Artificial intelligence in pathology: an overview. *Diagn Histopathol*. 2020;26:513–520.
7 Niazi KK, Gurcan MN, Khalid M, Niazi K, Parwani AV, Gurcan MN. Digital oncology 1 digital pathology and artificial intelligence. www.thelancet.com/oncology; 2019.
8 Mahmood H, Shaban M, Indave BI, Santos-Silva AR, Rajpoot N, Khurram SA. Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: a systematic review. *Oral Oncol*. 2020;110. https://doi.org/10.1016/j.oraloncology.2020.104885.
9 Zhang A, Lipton ZC, Li MU, Smola AJ. *Dive into deep learning*. 2022 [cited 2025 May 23]. https://doi.org/10.48550/arXiv.2106.11342 [Internet]. Version 0.17.1.
10 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29;* Venice, Italy. IEEE; 2017. p. 618–626. https://doi.org/10.1109/ICCV.2017.74.
11 Zhang Y, Hong D, McClement D, Oladosu O, Pridham G, Slaney G. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *J Neurosci Methods*. 2021;353. https://doi.org/10.1016/j.jneumeth.2021.109098.
12 Tejani A, Klontzas M, Gatti A, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6:e240300.
13 El-Naggar AK, Chan JK, Grandis RJ, Takata T, Slootweg P. *WHO classification of head and neck tumours*. 4th ed. Lyon: World Health Organization; 2017.
14 Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. http://arxiv.org/abs/2201.03545; 2022.
15 Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. http://arxiv.org/abs/1905.11946; 2019.
16 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. http://arxiv.org/abs/1409.4842; 2014.
17 Howard AG, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. http://arxiv.org/abs/1704.04861; 2017.
18 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. http://arxiv.org/abs/1512.03385; 2015.
19 Bello I, Fedus W, Du X, et al. Revisiting ResNets: improved training and scaling strategies. http://arxiv.org/abs/2103.07579; 2021.
20 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. http://arxiv.org/abs/1409.1556; 2014.
21 Chollet F. Xception: deep learning with depthwise separable convolutions.
22 Chollet F. Xception: Deep learning with depthwise separable convolutions. In: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26*. Honolulu, HI, USA: IEEE; 2017. p. 1800–1807. https://doi.org/10.1109/CVPR.2017.195.
23 Codella NCF, Gutman D, Celebi ME, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). http://arxiv.org/abs/1710.05006; 2017.
24 Combalia M, Codella NCF, Rotemberg V, et al. BCN20000: dermoscopic lesions in the wild. http://arxiv.org/abs/1908.02288; 2019.
25 Tschandl P, Rosendahl C, Kittler H. Data descriptor: the HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*. 2018;5. https://doi.org/10.1038/sdata.2018.161.
26 Mahmood H, Shaban M, Rajpoot N, Khurram SA. Artificial intelligence-based methods in head and neck cancer diagnosis: an overview. *Br J Cancer*. 2021;124:1934–1940.
27 Qiu Z, Rivaz H, Xiao Y. Is visual explanation with Grad-CAM more reliable for deeper neural networks? a case study with automatic pneumothorax diagnosis. http://arxiv.org/abs/2308.15172; 2023.
28 Alabdan R, Alruban A, Hilal AM, Motwakel A. Artificial-intelligence-based decision making for oral potentially malignant disorder diagnosis in internet of medical things environment. *Healthcare*. 2023;11. https://doi.org/10.3390/healthcare11010113.
29 Alanazi AA, Khayyat MM, Khayyat MM, Elamin Elnaim BM, Abdel-Khalek S. Intelligent deep learning enabled oral squamous cell carcinoma detection and classification using biomedical images. *Comput Intell Neurosci*. 2022;2022. https://doi.org/10.1155/2022/7643967.
30 Jubair F, Al-karadsheh O, Malamos D, Al Mahdi S, Saad Y, Hassona Y. A novel lightweight deep convolutional neural network for early detection of oral cancer. *Oral Dis*. 2022;28:1123–1130.
31 Talwar V, Singh P, Mukhia N, et al. AI-assisted screening of oral potentially malignant disorders using smartphone-based photographic images. *Cancers (Basel)*. 2023;15. https://doi.org/10.3390/cancers15164120.
32 Camalan S, Mahmood H, Binol H, et al. *Convolutional neural network-based clinical predictors of oral dysplasia : class activation map analysis of deep learning results*. 2021.
33 Anantharaman R, Anantharaman V, Lee Y. Oro vision: deep learning for classifying orofacial diseases. In: *Proceedings - 2017 IEEE International Conference on healthcare Informatics, ICHI 2017*. Institute of Electrical and Electronics Engineers Inc.; 2017:39–45.
34 Lee SJ, Oh HJ, Son YD, et al. Enhancing deep learning classification performance of tongue lesions in imbalanced data: mosaic-based soft labeling with curriculum learning. *BMC Oral Health*. 2024;24. https://doi.org/10.1186/s12903-024-03898-3.
35 Fu Q, Chen Y, Li Z, et al. A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: a retrospective study. *eClinicalMedicine*. 2020;27. https://doi.org/10.1016/j.eclinm.2020.100558.
36 Chen R, Wang Q, Huang X. Intelligent deep learning supports biomedical image detection and classification of oral cancer. *Technol Health Care*. 2024;32:S465–S475.
37 Rabinovici-Cohen S, Fridman N, Weinbaum M, et al. From pixels to diagnosis: algorithmic analysis of clinical oral photos for early detection of oral squamous cell carcinoma. *Cancers (Basel)*. 2024;16. https://doi.org/10.3390/cancers16051019.
38 Shamim MZM, Syed S, Shiblee M, et al. Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer. *Comput J*. 2022;65:91–104.
39 Tanriver G, Soluk Tekkesin M, Ergen O. Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders. *Cancers (Basel)*. 2021;13. https://doi.org/10.3390/cancers13112766.
40 Welikala RA, Remagnino P, Lim JH, et al. Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. *IEEE Access*. 2020;8:132677–132693.
41 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*. 2022;51:699–704.
42 Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P. Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. *J Oral Pathol Med*. 2021;50:911–918.
43 Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P, Vicharueang S. AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer. *PLoS One*. 2022;17. https://doi.org/10.1371/journal.pone.0273508.
44 Yousif MJ. Enhancing the accuracy of image classification using deep learning and preprocessing methods. *Artif Intell Robot Dev J*. 2024. https://doi.org/10.52098/airdj.2023348.
45 Zhang H, Ogasawara K. Grad-CAM-based explainable artificial intelligence related to medical text processing. *Bioengineering*. 2023;10. https://doi.org/10.3390/bioengineering10091070.