Enhancing Explainability in Oral Cancer Detection with Grad-CAM Visualizations

Arnaldo V. Barros da Silva[1(✉)], Cristina Saldivia-Siracusa[2], Eduardo Santos Carlos de Souza[3], Anna Luíza Damaceno Araújo[4], Marcio Ajudarte Lopes[2], Pablo Agustin Vargas[2], Luiz Paulo Kowalski[4], Alan Roger Santos-Silva[2], André C. P. L. F. de Carvalho[3], and Marcos G. Quiles[1]

[1] Institute of Science and Technology, Federal University of São Paulo (UNIFESP), São José Dos Campos, SP, Brazil
{arnaldo.barros,quiles}@unifesp.br

[2] Oral Diagnosis Department, Piracicaba Dental School, University of Campinas (FOP-UNICAMP), Piracicaba, SP, Brazil
c234721@dac.unicamp.br, malopes@fop.unicamp.br, {pavargas,alan}@unicamp.com.br

[3] Institute of Mathematical and Computer Sciences, University of São Paulo (ICMC-USP), São Carlos, SP, Brazil
{eduardo.santos.souza,andre}@usp.br

[4] Head and Neck Surgery Department, University of São Paulo Medical School, Sao Paulo, SP, Brazil
lp_kowalski@uol.com.br

Abstract. Late diagnosis of oral cancer significantly compromises patient outcomes. A promising approach to speed up the diagnostic process involves the use of Deep Learning (DL) models for medical image analysis. However, a notable challenge with these models is their lack of interpretability. To address this, techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) have been developed. Grad-CAM generates heatmaps that highlight image regions most influential for classification decisions. In our study, we evaluated the performance of two DL models renowned for their high accuracy in oral cancer classification. Our analysis extended beyond mere accuracy metrics; we employed Grad-CAM to provide visual explanations of the models' decisions. Furthermore, we investigated subclass accuracy rates and the distribution of prediction confidences to gain a deeper insight into the models' performance and robustness in oral cancer detection. This comprehensive evaluation approach offers a more nuanced understanding of the capabilities and limitations of DL methods in the context of oral cancer diagnosis.

Keywords: Oral cancer diagnosis · Deep Learning · Interpretability · Grad-CAM · Model evaluation

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024, LNCS 14813, pp. 151–164, 2024.
https://doi.org/10.1007/978-3-031-64605-8_11

1 Introduction

Oral cancer ranks among the most common types of cancer worldwide [21], characterized by late diagnoses and high mortality rates. The delay in diagnosis can be attributed to various factors, with scant public awareness and lack of knowledge among healthcare professionals cited as primary contributors to this scenario [28]. Developing effective strategies to promote early detection is crucial to improve treatment outcomes and survival prospects for patients affected by this condition.

Using Computer-Aided Diagnosis (CAD) [7] is a highly effective approach for reducing diagnosis time. By analyzing medical images from various sources such as Hyperspectral Imaging, Magnetic Resonance Imaging, Computed Tomography, among others [16], it can enhance lesion detection. A key example of the power of CAD is illustrated in the work conducted by [25]. They employed a Deep Learning model to classify oral cancer on a dataset of clinical oral photographs and achieved 99% precision and 100% recall. This provides clear evidence that CAD is an indispensable tool in the context of medical diagnosis.

Deep Learning methods have the capability to automatically learn complex features from images as they enhance classification accuracy without relying on direct intervention from human experts [11]. However, a significant limitation of these models is their "black-box" nature [3], where the internal decision-making process is not easily understood by humans. This lack of interpretability can be a critical obstacle, especially in sensitive applications such as the medical field. In response, methods like CAM [30], Grad-CAM, Grad-CAM++ [4,19], LIME [18] and SHAP [15] have been developed, enabling the visualization of specific regions in input images that most influenced predictions generated by Convolutional Neural Networks (CNNs) [1].

The study [29] investigated the effectiveness of CAM, Grad-CAM, and Grad-CAM++ techniques in interpreting CNNs used for classifying multiple sclerosis (MS) cases from brain magnetic resonance imaging scans. It found that integrating Grad-CAM with robust CNN models like VGG19 yielded superior results. The research underscored the importance of combining Grad-CAM with advanced CNN architectures and quantification methods to identify key brain regions affected by MS pathology. In study [2], researchers investigated the interpretability of models in detecting retinoblastoma in fundus images. They utilized InceptionV3 to train a neural network on a dataset consisting of 400 retinoblastoma images and 400 images without the condition. Subsequently, they employed LIME and SHAP to explain the model's predictions. The findings demonstrated that these techniques effectively identified critical features in the images for precise diagnosis. This study implies that such an approach holds promise for enhancing the detection and treatment of retinoblastoma.

The main objective of this study is to refine the evaluation pipeline for oral cancer classification models, focusing on two ConvNeXt-based models [14] trained on a proprietary dataset developed by the authors. This dataset, which remains undisclosed to the public, serves as a novel resource for testing the efficacy of these models. Through multiple experimentation, these models

demonstrated superior efficacy with comparable performance, albeit employing different fine-tuning strategies during their development.

This paper is structured as follows: Sect. 2 reviews related literature, Sect. 3 details the experimental methodology, Sect. 4 discusses the findings, and Sect. 5 concludes the study, highlighting its contributions to the field of medical diagnostics and early cancer detection.

2 Related Works

In this section, studies applying Explainable Artificial Intelligence (XAI) methods in medical images will be presented, both to assess model accuracy and to provide additional tools to healthcare professionals. Additionally, works justifying the selection of Grad-CAM as an XAI method compared to other approaches will be discussed.

2.1 XAI for Medical Imaging

The work by Hauser et al. [10] sheds light on the increasing interest in Explainable Artificial Intelligenc (XAI) to address the black-box nature of Deep Neural Networks (DNNs) in skin cancer detection. The authors analyzed 37 articles published between January 2017 and October 2021 and found that while 19 studies applied existing XAI methods to interpret classifier decisions, 4 proposed new approaches or refined established techniques. Besides, only 3 studies evaluated the performance and confidence of healthcare professionals when using DNN-based decision support systems with XAI. These findings underscore the need for a more systematic and rigorous evaluation of the utility of XAI in this context.

In a similar vein, Warin et al. [26] evaluated the performance of CNN algorithms in classifying and detecting potentially malignant oral diseases and Oral Squamous Cell Carcinoma (OSCC) and Oral Potentially Malignant Disorders (OPMD) in photographic images. The researchers developed several image classification models, highlighting DenseNet-196 as the most effective, achieving AUCs of 1.00 and 0.98 for OSCC and OPMDs, respectively. Furthermore, object detection models were employed, with Faster R-CNN leading, achieving AUCs of 0.88 and 0.64 for OSCC and OPMDs, respectively. The researchers used Grad-CAM to illustrate the model's assertiveness. Similarly, the study of the DiCoMLP-Mixer model [17] combined dilated convolutions and a multilayer perceptron to assess oral connective tissue, capturing dense contextual representations at multiple scales. The backbone of DiCoMLP-Mixer also incorporates the attention mechanism of a transformer model. Once again, Grad-CAM was used to reinforce the assertiveness of the method.

In contrast, the work proposed by Figueroa et al. [9] developed a training method that makes the network's predictions more understandable, accurately directing its attention to cancerous regions of the image. The adopted strategy involves using Grad-CAM with the guided attention inference network

[12]. Results indicate that this approach enhances CNN interpretability, enabling more reliable lesion classification.

Finally, the study by Dorrich et al. [8] investigated the ability of CNNs to classify head and neck cancer histopathology, addressing the lack of interpretability with XAI techniques. The researchers trained a CNN to classify tumor and non-tumor tissue and another CNN for semantically segmenting four classes. The classification model's performance reached an accuracy of 89.9% on previously unseen data, while the segmentation model achieved a mean Intersection over Union of 0.690 and 0.782 specifically for tumor tissue. Explainable Artificial Intelligence methods evidenced that both networks base their decisions on features consistent with pathology experts' opinions.

2.2 Comparisons Between XAI Methods

In the work [5], a CNN was trained to classify images of LEGO pieces, and the LIME and Grad-CAM methods were applied to generate visual explanations. The results indicated that, overall, Grad-CAM outperformed LIME. According to 80% of the respondents, Grad-CAM produced more reliable heatmaps, providing more robust visual interpretations of the CNN's decisions.

In the study [23], a variety of models such as Vision Transformers, MobileNet, and ResNet were trained on a dataset of endoscopic images. Among the XAI techniques evaluated, propagation-based methods such as GradCAM, GradCAM++, and LayersCAM, as well as feature-based methods like SHAP and LIME, stood out. It was observed that among the propagation methods, Grad-CAM exhibited superior performance, especially in challenging classes such as the Ileocecal Valve. In overall comparison, it also outperformed SHAP and LIME, although it was noted that if the model does not employ backpropagation mechanisms, such as decision trees or k-nearest neighbors, the path to pursue would be SHAP or LIME. Another limitation of Grad-CAM was also identified: when trying to locate objects with multiple occurrences of the same class, most of the time it generated a failure.

Finally, the investigation [27] explored models for classifying lesions on fruit leaves, employing the methods GradCAM, SmoothGrad, and LIME for explainability. The results revealed that GradCAM offers more intuitive and understandable interpretations. SmoothGrad, despite highlighting the importance of pixels in lesion areas, lacks clarity in explaining leaf appearance. Similarly, the results of LIME were considered unsatisfactory, limited to describing leaf appearance without providing a robust explanation.

3 Experiments

3.1 Dataset

The dataset that has been analyzed in this study includes a total of 837 clinical photographs that were taken by professional cameras or mobile devices. These



Fig. 1. Examples of images in the dataset. The first row depicts images of the OPMD class, while the second row displays images of the OSCC class.

photographs were obtained from three different institutions. The majority of the images, which corresponds to 758 patients, were collected from the *Faculdade de Odontologia de Piracicaba* (FOP) in Piracicaba, São Paulo, Brazil. In addition to this, the dataset also includes contributions from 11 patients from the *Universidade Federal de Minas Gerais* (UFMG) in Belo Horizonte, Minas Gerais, and 38 patients from the *Universidade Federal do Pará* (UFPA) in Belém, Pará, Brazil. To ensure that the patient labeling was accurate, a group of specialists conducted a histopathological analysis of the corresponding glass slides, aiming to review the diagnosis and eliminate any doubts.

The images were categorized into two main categories following the World Health Organization (WHO) guidelines from 2017 [20]. The first category was labeled as OPMDs, which included 351 cases ranging from no evidence to mild or moderate dysplasia. The second category as labeled as OSCC, which included 486 cases ranging from carcinoma in situ and micro-invasive to overtly invasive/conventional and verrucous forms. Some examples of images from each class can be seen in Fig. 1.

Our experimental evaluation was conducted using the test subset of the dataset, which represents 10% of the entire dataset. This subset comprises 33 examples of Oral Potentially Malignant Disorders (OPMDs) and 46 examples of Oral Squamous Cell Carcinoma (OSCC), totaling 79 unique cases. The distribution of these images, categorized by their collection location, is detailed in Table 1, providing a clear view of the dataset's composition.

It is important to note that the models utilized in this research were pre-trained and directly applied to the test data without undergoing any further fine-tuning or hyperparameter adjustments. This approach was deliberately chosen because the primary focus of our investigation was to assess the performance and applicability of these already-trained models in the context of our specific dataset. Consequently, there was no need to employ the training or validation sets typically used in the model development phase.

Table 1. Distribution of clinical photographs by collection location and class for the test subset: FOP (Faculdade de Odontologia de Piracicaba), UFMG (Universidade Federal de Minas Gerais), UFPA (Universidade Federal do Pará).

Class	Images			
	FOP	UFMG	UFPA	Total
OPMD	30	1	2	33
OSCC	45	0	1	46

3.2 Models

Several models with distinct architectures have been trained based on the constructed dataset, with results indicating that the most promising architecture is ConvNeXt. Built upon a classical convolutional network such as ResNet [13], ConvNeXt incrementally evolves by incorporating distinctive features from transformers [24]. Among these features are the introduction of a "patchify" layer, deep and wide convolutions, the implementation of non-local self-attention, and the utilization of larger kernels.

The two best-performing ConvNeXt models were obtained using two different strategies. **Model A** involved fine-tuning a pre-trained model on the ImageNet dataset [6], while **Model B** followed the same approach with an additional fine-tuning step using the ISIC 2019 dataset [22]. The accuracy achieved by Model A was 78.48%, whereas Model B achieved higher accuracy of 79.75%. The chosen ConvNeXt architecture has a total of 49,456,226 trainable parameters with an input shape of 224×244x3.

3.3 Grad-CAM

As previously mentioned, the selected XAI technique for the experiments was Grad-CAM. An advantage of this approach is its universal applicability to any CNN architecture, without requiring modifications to the network structure. This sets it apart from its predecessor, Class Activation Mapping (CAM), which necessitates the inclusion of specific additional layers like the Global Average Pooling (GAP) layer and softmax activation within the network architecture. By bypassing the need for these supplementary layers, Grad-CAM streamlines the interpretability process.

To explain its results, Grad-CAM calculates the gradient ($\frac{\partial y^c}{\partial A^k}$) of the score y^c for class c with respect to the activation of the feature map A^k from convolution k in the deepest convolutional layer. The gradients are then globally aggregated via max-pooling to determine the convolution's significance for the class, represented as α_c^k. This computation yields a localization map L^c, obtained through weighted summation followed by a Rectified Linear Unit (ReLU) activation function, which highlights only positive contributions.

4 Results

4.1 Accuracy by Class and Collection Location

In order to obtain a comprehensive understanding of the performance of the classification models, we have compiled two tables - Table 2 for model A and Table 3 for model B - which present the accuracy rates by class and collection location. These tables reinforce the notion that the selection of a model should not be solely based on overall accuracy values. For instance, Model A achieves a 10% higher accuracy rate in predicting the OPMD class for images collected at FOP, while Model B exhibits a 6.67% higher accuracy rate for the OSCC class. It is worth mentioning that only Model B presents an accurate prediction in the only OSCC example collected at UFPA.

Table 2. Model A accuracy separated by class and institution for each example in the test set.

Class	FOP	UFMG	UFPA
OPMD	**83,33%**	0%	50%
OSCC	77,77%	–	0%

Table 3. Model B accuracy separated by class and institution for each example in the test set.

Class	FOP	UFMG	UFPA
OPMD	73,33%	0%	50%
OSCC	**84.44%**	–	**100%**

4.2 Venn Diagram

The results obtained from the subclass and collection region accuracy indicate that even though the models have similar architecture and overall accuracy, they learned different image features to produce their predictions. To verify this, we can create a Venn diagram that represents shared correct predictions, distinct correct predictions, and common errors between the models. Figure 2 shows the diagram with the results for the two analyzed models. We noticed that there were images that one model classified correctly while the other did not. This suggests that the models may have learned some divergent features. Furthermore, the results emphasize that it is not as straightforward to claim that one model is better than another. For such a statement to hold, the model with higher accuracy must correctly classify all instances that the other model got right along with additional examples.

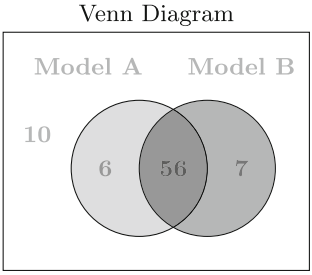


Fig. 2. A Venn Diagram illustrating the predictions of ConvNeXt with only ImageNet fine-tuning (Model A) and ConvNeXt utilizing both ImageNet and ISIC 2019 fine-tuning (Model B).

4.3 Confidence Distribution

In the domain of utilizing models in CAD systems, minimizing false positives is of vital importance, given that such erroneous predictions could gravely misdirect a patient towards incorrect treatment paths. To this end, analyzing classification confidences assumes significant importance as it facilitates the detection of misguided predictions, which are made with low confidence. Moreover, it may prove beneficial to establish a separate class for "undefined predictions," which could help in eliminating certain false positives.

In Figs. 3 and 4 we show the distributions of confidences for wrong and correct predictions for Model A and Model B respectively. Upon analyzing the figures, it is evident that creating a new class is not a simple task. Overall, the models, both in correct and incorrect predictions, tend to exhibit a high confidence value in their decisions.

4.4 Visual Explanations

In order to expand upon the analysis of model performance beyond just accuracy rates, we utilized Grad-CAM to generate visual explanations for all test images using both models. These visualizations were then presented side by side to better understand the differences between them.

The first observation is that in many cases, both models correctly predicted the OSCC class. However, when this happened, Model B generated a less scattered heatmap than Model A, as seen in Fig. 5. This characteristic suggests that Model B may have learned better to detect lesion regions.

Another observation is that when the models predicted the OPMD class, their heatmaps focused on regions that had little or no intersection with the lesions observed by specialists. This suggests that perhaps both models have learned to distinguish between "with lesion" and "without lesion" rather than OPMD and OSCC. Figure 6 shows this trend.

Finally, Fig. 7 shows examples where the models diverge from each other, indicating that even with similar architectures and similar training methods, both models learn different features.

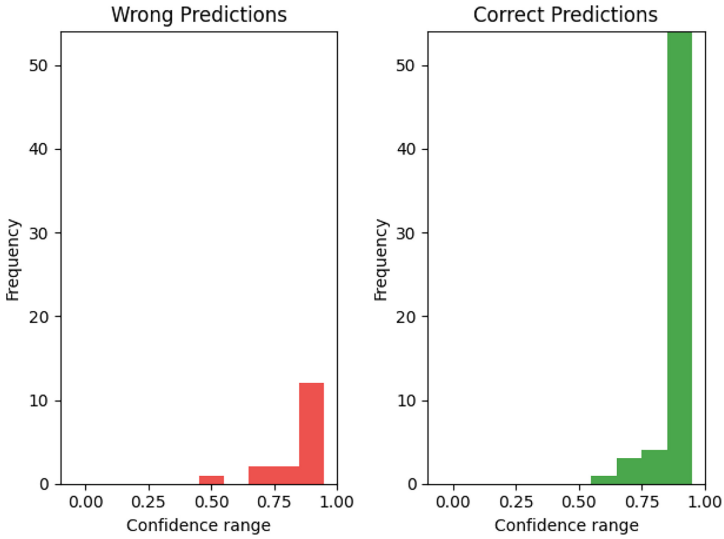


Fig. 3. Distribution of confidences for wrong predictions (red) and correct predictions (green) made by Model A in the test set. (Color figure online)

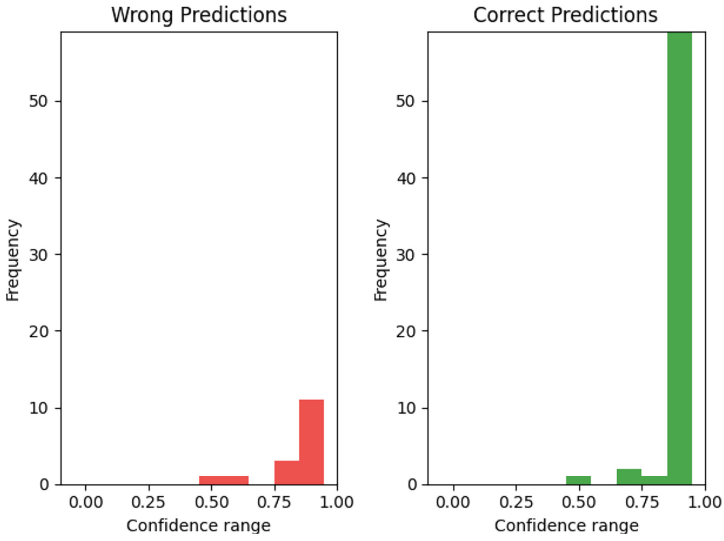


Fig. 4. Distribution of confidences for wrong predictions (red) and correct predictions (green) made by Model B in the test set. (Color figure online)

Delving deeper, according to medical experts, during a visual assessment, model B appeared more confident in its visual explanations. Approximately 56.96% of the maps generated for the test set by model B focused on the correct regions of the lesions, while only 40.51% of the maps from model A showed this level of accuracy.

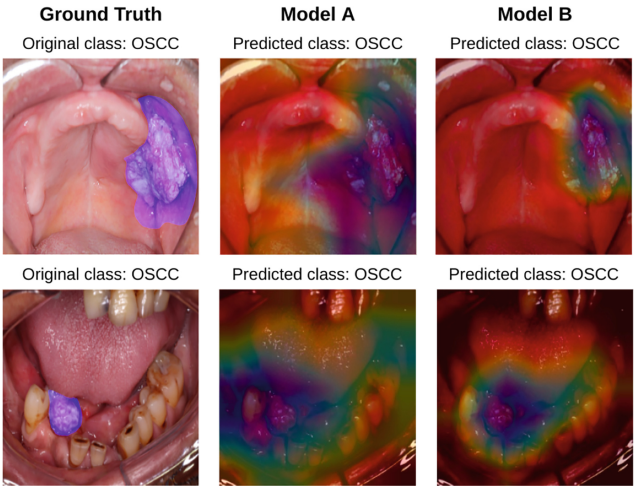


Fig. 5. Examples of visual explanations generated via Grad-CAM in images from the test set (first column) using Model A (second column) and Model B (third column) where Model B showed less spread out heatmaps.

4.5 Data Preprocessing

After noticing that the heatmap was focusing on areas of the image that were not related to the lesion, such as the teeth, we decided to crop the image to only include the lesion region based on segmentation annotations provided by specialists. We hoped that this would improve the accuracy of our classification without retraining the models. However, our results showed that this hypothesis was incorrect. The accuracy of the first model decreased from 78.48% to 75.95% after implementing the preprocessing step, and the accuracy of the second model decreased from 79.75% to 70.89%.

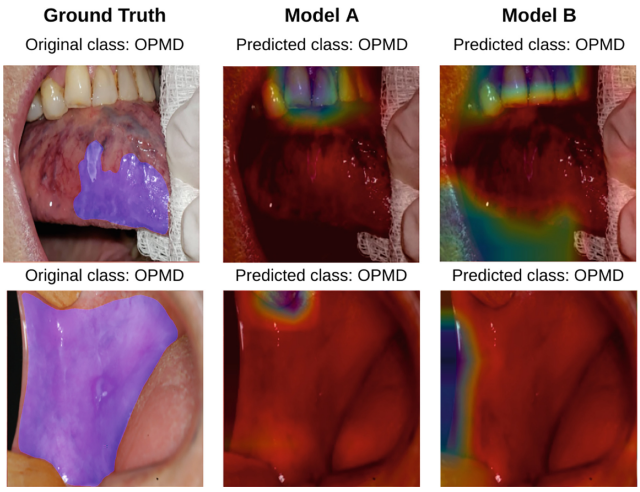


Fig. 6. Examples of visual explanations generated via Grad-CAM on images from the test set (first column) using Model A (second column) and Model B (third column) where both models showed heatmaps that did not intersect with lesions identified by experts.

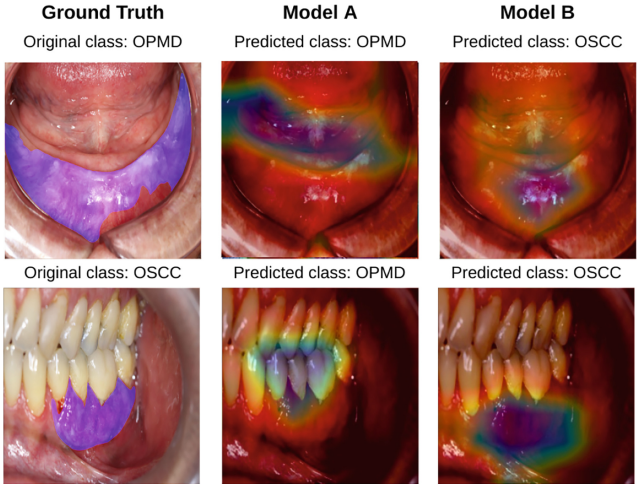


Fig. 7. Examples of visual explanations generated via Grad-CAM in images from the test set (first column) using Model A (second column) and Model B (third column) where the predictions generated by the models diverge.

5 Conclusion

This study aimed to provide a more robust pipeline for evaluating medical image classification models, given the cautious nature of tasks directly impacting people's health. To achieve this, two models based on the ConvNeXt architecture were analyzed, and trained on the same data but with different fine-tuning strategies.

Initially, when considering only overall accuracy, Model B seemed to be the optimal choice, showing a 0.9% improvement over Model A. However, when dividing the test set into subclasses, more significant differences emerged: Model A exhibited superiority in OPMD classification, while Model B excelled in OSCC classification.

Delving deeper into the investigation, a Venn Diagram was constructed to compare the predictions of each model. The results revealed that the models are not easily interchangeable, suggesting that they likely learned distinct features despite sharing similar architectures and being trained on the same data.

This distinction between models became even clearer when examining the regions each model focused on during predictions, using the Grad-CAM technique. Remarkably, while Model B excelled in identifying lesion regions in OSCC cases, both models did not demonstrate a focus on lesion regions in OPMD cases.

Using the explanations provided by Grad-CAM with the assistance of medical expertise, we found that model B was the most reliable overall. However, the overall confidence was still low, which likely hinders the use of this model in the real world despite its high accuracy.

Furthermore, other analyses were conducted, such as evaluating the distribution of model confidences and making data modifications to potentially improve results. In summary, this study provided a clear experimentation pipeline, indicating that the choice of the best model depends more on the specific context and objectives of the application than on the overall accuracy achieved.

Acknowledgements. The authors gratefully acknowledge support from FAPESP (São Paulo Research Foundation), projects No. 2020/09835 − 1, 2022/13069 − 8 and 2022/09285−7, and the CNPq (National Council for Scientific and Technological Development), project No. 313680/2021 − 3.

Disclosure of Interests. The authors declare no competing interests relevant to the content of this article.

References

1. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. IEEE (2017)
2. Aldughayfiq, B., Ashfaq, F., Jhanjhi, N., Humayun, M.: Explainable AI for retinoblastoma diagnosis: interpreting deep learning models with lime and shap. Diagnostics **13**(11), 1932 (2023)

3. Castelvecchi, D.: Can we open the black box of AI? Nat. News **538**(7623), 20 (2016)
4. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847. IEEE (2018)
5. Cian, D., van Gemert, J., Lengyel, A.: Evaluating the performance of the lime and grad-cam explanation methods on a LEGO multi-label image classification task. arXiv preprint arXiv:2008.01584 (2020)
6. Deng, J., et al.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
7. Doi, K.: Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput. Med. Imaging Graph. **31**(4–5), 198–211 (2007)
8. Dörrich, M., et al.: Explainable convolutional neural networks for assessing head and neck cancer histopathology. Diagn. Pathol. **18**(1), 121 (2023)
9. Figueroa, K.C., et al.: Interpretable deep learning approach for oral cancer classification using guided attention inference network. J. Biomed. Opt. **27**(1), 015001–015001 (2022)
10. Hauser, K., et al.: Explainable artificial intelligence in skin cancer recognition: a systematic review. Eur. J. Cancer **167**, 54–69 (2022)
11. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
12. 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, pp. 9215–9223 (2018)
13. Li, S., Jiao, J., Han, Y., Weissman, T.: Demystifying resnet. arXiv preprint arXiv:1611.01186 (2016)
14. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017). http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
16. McAuliffe, M.J., Lalonde, F.M., McGarry, D., Gandler, W., Csaky, K., Trus, B.L.: Medical image processing, analysis and visualization in clinical research. In: Proceedings 14th IEEE Symposium on Computer-Based Medical Systems, CBMS 2001, pp. 381–386. IEEE (2001)
17. Pratiher, S., Chattoraj, S., Nawn, D., Pal, M., Paul, R.R., Konik, H., Chatterjee, J.: A multi-scale context aggregation enriched mlp-mixer model for oral cancer screening from oral sub-epithelial connective tissues. In: 2022 30th European Signal Processing Conference (EUSIPCO), pp. 1323–1327 (2022). https://doi.org/10.23919/EUSIPCO55093.2022.9909942
18. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13-17 August 2016, pp. 1135–1144 (2016)

19. Selvaraju, R.R., 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 IEEE International Conference on Computer Vision, pp. 618–626 (2017)
20. Soluk-Tekkeşin, M., Wright, J.M.: The world health organization classification of odontogenic lesions: a summary of the changes of the 2017 (4th) edition. Turk Patoloji Derg **34**(1), 1–18 (2018)
21. Tranby, E.P., et al.: Oral cancer prevalence, mortality, and costs in medicaid and commercial insurance claims data. Cancer Epidemiol. Biomark. Reven. **31**(9), 1849–1857 (2022)
22. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1), 1–9 (2018)
23. Varam, D., et al.: Wireless capsule endoscopy image classification: an explainable AI approach. IEEE Access **11**, 105262–105280 (2023)
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. **30** (2017)
25. 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. **50**(9), 911–918 (2021)
26. 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 **17**(8), e0273508 (2022)
27. Wei, K., Chen, B., Zhang, J., Fan, S., Wu, K., Liu, G., Chen, D.: Explainable deep learning study for leaf disease classification. Agronomy **12**(5), 1035 (2022)
28. 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 (2020). https://doi.org/10.1109/ACCESS.2020.3010180
29. 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 **353**, 109098 (2021)
30. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)