Discriminative Feature Analysis in XAI for Multi-Class Classification of Oral Lesions

Paolo Fantozzi[1], Paolo Junior Fantozzi[2,3,†], Najwa Yousef[4,5,†], Mathilde Casagrande[2,3], Gianluca Tenore[2,3], Antonella Polimeni[2,3], James J. Sciubba[6], Tiffany Tavares[7], Umberto Romeo[2,3], Ahmed Sultan[4,5,‡] and Maurizio Naldi[1,*,‡]

[1]*Department of Law, Economics, Politics, and Modern Languages, LUMSA University, Rome, Italy*
[2]*Department of Oral and Maxillofacial Sciences, Sapienza University of Rome, Rome, Italy*
[3]*Department of Head and Neck, Umberto I University Hospital, Rome, Italy*
[4]*Department of Oncology and Diagnostic Sciences, University of Baltimore School of Dentistry, Baltimore, MD, USA*
[5]*Division of Artificial Intelligence Research, University of Maryland School of Dentistry, Baltimore, MD, USA*
[6]*Department of Otolaryngology, Head & Neck Surgery, The Johns Hopkins University, Baltimore, MD, USA*
[7]*Department of Comprehensive Dentistry, UT Health San Antonio School of Dentistry, San Antonio, TX, USA*

Abstract

The interpretability of machine learning (ML) models is critical in medical applications, particularly in diagnosing and classifying oral lesions. Traditional saliency maps highlight relevant features for classification, but typically focus on a single predicted class. However, this approach can lead to inconsistencies, misinterpretations, and an overwhelming number of comparisons in multi-class classification tasks. In this paper, we introduce a novel multi-class saliency map that integrates feature importance across all possible classifications, accounting for class assignment probabilities to enhance explainability. As an early demonstration of its performance, we report the results obtained with a dataset of 224 oral lesion images labeled by medical experts. Using deep learning models based on Vision Transformers (BEiT) and interpretability techniques such as LIME, we construct a unified saliency representation that highlights discriminative features across all classes and effectively eliminates misleading areas while emphasizing truly relevant regions for classification. These early results demonstrate improved clarity in feature attribution, supporting more reliable and interpretable AI-driven diagnostics in oral healthcare.

Keywords

Machine Learning, Explainable AI, Healthcare, Oral lesions, Multi-class classification, Saliency maps, Feature importance, Medical imaging

1. Introduction

The interpretability of machine-learning (ML) models is a must-have property nowadays, especially in application contexts where the results of such models have to be communicated to the general audience. Interpretability refers to the ability to understand and explain how a model makes decisions, which is often achieved by highlighting the features that have contributed most to that decision. This has become particularly critical with the advent of complex (hence, less transparent) computational structures like transformers [1]. Also, interpretability is going to become a distinctive feature that may have a commercial value [2]. A well-developed tool for interpretability is represented by saliency maps, which assign importance scores to pixels (in images) or features (in structured data) based on their contribution to the final output [3]. However, saliency maps are traditionally provided for a single class, typically the most likely class, i.e., the one that is assigned by the classifier to the instance at hand. For

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey
*Corresponding author.
†Joint contribution
‡Joint contribution
m.naldi@lumsa.it (M. Naldi)
0000-0002-5442-2239 (P. Fantozzi); 0000-0002-6807-3391 (P. J. Fantozzi); 0000-0001-9963-8052 (G. Tenore); 0000-0002-2679-7607 (A. Polimeni); 0000-0001-7515-5793 (J. J. Sciubba); 0000-0003-2439-2187 (U. Romeo); 0000-0001-5286-4562 (A. Sultan); 0000-0002-0903-398X (M. Naldi)

example, in a *dog* vs. *cat* classifier, if the model predicts *dog*, the saliency map will highlight the features that contributed to the dog class. Examples of this approach are [4] and [5]. Some methods generate a separate saliency map for each possible class to help understand which parts of the image contribute to different class scores. In the simple example just mentioned, such an approach would generate one map showing features contributing to *dog* and another map showing features contributing to *cat*. An example of such an approach is the paper by Shimoda [6].

However, such an approach has significant shortcomings. First, it does not provide a comprehensive view of the relevance of features. Some features may be relevant for one class but not for another, showing a lack of consistency. Simple logical operations cannot remedy this shortcoming. For example, taking the OR of the features deemed as relevant across the classes (i.e., simply listing the features that are relevant for at least one class) could end up obtaining the whole set of features. Hence, our analysis would not hold discriminative power. On the other hand, taking the AND of the features deemed relevant across the classes (i.e., considering just the features that are relevant for all the classes) could return an empty set. If we stick to the separate approach, we are then required to carry out multiple comparisons across class-specific maps, with the number of comparisons to be carried out growing as the square of the number of classes. This number would soon swamp our human capability to carry out comparisons. Also, individual class-wise saliency maps might falsely suggest that features not highlighted for a specific class are unimportant, making the whole procedure less robust to misinterpretation. We are then in strong need of an overall map to ensure that all critical regions are accounted for.

In this paper, we tackle that issue by proposing a multi-class saliency map that exploits an explainability-based approach to account for all the single-class saliency maps. This multi-class saliency map also accounts for the different relevance of output classes through their assignment probability. We test our proposal by applying it to a dataset made of real pictures of oral lesions, labelled by a team of medical experts. The dataset has been collected specifically for this purpose. We show that our multi-class approach allows us to correctly identify the most relevant areas in the picture while ruling out some patently non-relevant areas that appeared instead as relevant in some single-class maps.

Our paper is organized as follows. After briefly reviewing the literature in Section 2, we describe our multi-class saliency map in Section 3 and show the early results of its application in Section 4.

2. Related literature

Various machine-learning techniques have been applied to tasks related to oral cancer. Still, most of the works focus on Support Vector Machines, Artificial Neural Networks and Linear Regression, reaching 87.71% of published works until 2022[7]. The theme of explainability has been investigated in very few papers. In this section, we review the most relevant literature on the diagnosis of oral cancer through an ML approach, highlighting those papers where explainability has been sought after. A survey of the literature dealing with deep learning applications in the broader field of maxillofacial diseases is contained [8]. It highlights the lack of explainability as one of the major challenges. A more recent survey focusing on oral cancer is contained in [9], where again the explainability issue is highlighted.

We can now examine the small set of papers that use some form of ML explainability method instead. Wu et al. in [10] used Random Forest, Linear Discriminant Analysis (LDA), and Logistic Regression classifiers applied to various numerical and categorical features obtained by a public patient data registry. Shapley values were employed to extract the most relevant features. Rai et al. in [11] trained a CNN with input at the same time a picture and a microscope image. In parallel, image features are extracted from both the data sources and then passed to a classification layer for a cancerous vs non-cancerous classification. Duran-Sierra et al. in [12] use features extracted from maFLIM images, classifying each pixel by using either Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machines (SVM), and Logistic Regression (LOGREG). Then they apply a threshold over the probability map obtained by the output of the classifiers for the single pixels.

Some papers employ explainable techniques, mainly through the use of heatmaps. Cimino et al. presented a method to combine Convolutional Neural Networks with Case-Based Reasoning (CBR),

modifying the model architecture to get a measure of similarity to previous cases, and also exploiting saliency maps generated by GradCAM++ [13]. Again, a Grad-CAM technique was applied in [14] to explain oral cancer predictions with two machine learning models. Explainability is achieved by generating heatmaps that highlight image regions that are most influential in classification decisions. The same technique, enhanced through the use of the guided attention inference network (GAIN), is employed in [15] for oral cancer. An alternative approach is proposed in [16], where explainability is achieved by presenting the closest training instances where a close alignment is obtained between human-made decisions and the ML algorithm. This approach has been dubbed Informed Deep Learning (IDL). Rather than achieving explainability by design, a post hoc approach is taken in [17], where SHAP values are computed to explain the model's predictions in screening for oral cancers as a function of the patients' features. Similarly, the System Usability Scale (SUS) and System Causability Scale (SCS) are proposed in [18] to evaluate the explainability of ML-based diagnosis of oral tongue cancer.

3. Method

We want to address the problem of saliency maps for multi-class image classification. Most methods return a saliency map for each class, representing the importance scores for the features relative to the class (usually, the only one shown is the saliency map relative to the class returned by the classifier for the input). Instead, we would like to show a saliency map over the features that incorporates the contributions of each feature to all the classes. Let's consider a set of images, as is typically the case in most healthcare diagnostic tasks. We assume we will employ the raw intensity of pixels as the features.

We consider the classifier as a method returning a probability distribution over classes, namely the probability p_c that the instance at hand will be assigned class c, with the obvious constraint

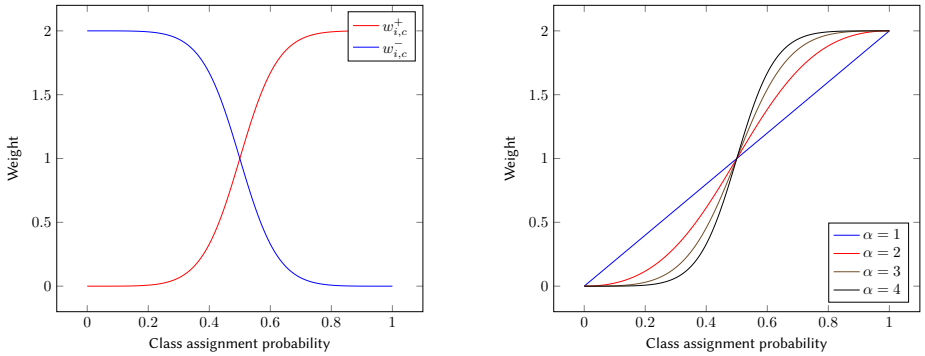
$$\sum_{c \in C} p_c = 1$$

We indicate the saliency map obtained by the method m for the instance (input image) i with respect to the class c by $S_{i,m,c}$. This is a single-class saliency map and is output by the explainability method of choice. Our aim is to arrive at a saliency map that incorporates the contribution of each feature across all the classes, i.e., the multi-class saliency map $\Delta_{i,m}$. Each saliency map is actually an N-by-M matrix of scores, where N and M are, respectively, the number of rows and the number of columns in the image, with the combination of row and column identifying a specific pixel.

The passage from the single-class saliency maps $S_{i,m,c}$ to the multi-class saliency map $\Gamma_{i,m}$ is accomplished through a three-stage process:

1. Single-class saliency map probability-based weighting;
2. Difference-Sum computation of the multi-class saliency map;
3. Score normalization.

As already hinted, we envisage starting with a set of single-class saliency maps $S_{i,m,c}$ ($c \in C$) provided by the explainability method of choice (e.g., LIME). For each class, the matrix $S_{i,m,c}$ is made of the scores $s_{i,m,c}^{(j,k)}$. For each class, we also assume that the score $s_{i,m,c}^{(j,k)}$ will be positive if the feature, i.e., the pixel (j,k), increases the probability that the instance i is assigned to the class c. It will be negative in the reverse case when that pixel (j,k) decreases that probability. However, we would also like scores to reflect the contribution of the feature to the actual classification. In a multi-class context, the class assigned to the instance is the one exhibiting the highest probability. Hence, we wish the score to reflect not just a probability increase of a given class but also the actual contribution of the feature towards shifting the classification decision towards that class. Since classification decisions are based on the assignment probability, we wish the feature score to be higher when the assignment probability is higher. On the other hand, if the assignment probability is low, that means that the feature (with a negative score) has really contributed to moving the classification decision away from that class. We can incorporate the influence of the assignment probability by properly weighting the saliency map



(a) Weighting functions with $\alpha = 4$ (b) Curve changing in terms of α

Figure 1: Class assignment probability curves

scores. We just need two weights $w^+_{i,c}$ and $w^-_{i,c}$ for negative and positive scores, respectively. As to the functional relationship between the weights and the class assignment probability, we have opted for the following sigmoidal shape, where α is a calibration parameter:

$$\begin{aligned} w^+_{i,c} &= \frac{2}{1 + (\frac{1}{p_c} - 1)^\alpha} \\ w^-_{i,c} &= 2 - \frac{2}{1 + (\frac{1}{p_c} - 1)^\alpha} \end{aligned} \tag{1}$$

We show an example of those functions in Figure 1a when $\alpha = 4$. The impact of the coefficient α can be seen in Figure 1b, where we see that $\alpha = 1$ gives rise to a linear weighting function.

For a given instance and class, the weights do not depend on the pixel, i.e., they are the same for all the pixels in the image. Hence, they serve to alter the ranking relationship between scores pertaining to different classes. However, the weight choice (i.e., $w^+_{i,c}$ or $w^-_{i,c}$) depends on the score sign. We form then an intermediate single-class saliency map $S^i_{i,m,c}$ whose elements are

$$s'^{j,k}_{i,m,c} = \begin{cases} s^{j,k}_{i,m,c} \times w^+_{i,c} & \text{if } s^{j,k}_{i,m,c} > 0 \\ s^{j,k}_{i,m,c} \times w^-_{i,c} & \text{if } s^{j,k}_{i,m,c} < 0 \end{cases} \tag{2}$$

We can now move to derive a multi-class saliency map based on several single-class saliency maps. We rely on the assumption that the most discriminative features are those that exhibit the largest differences in scores across classes. If a feature has the same score for all the classes, it is of no help in deciding the final class assignment. For example, a feature whose scores are -1 for the class c_1 and 1 for the class c_2 should be considered much more discriminative than a feature with scores -0.1 and 0.1, respectively. Based on the pairwise differences in scores across classes, we can build the multi-class saliency map $\Delta_{i,m}$, whose elements $\delta^{j,k}_{i,m}$ are the sum of all the pairwise differences (hence the name *difference-sum* saliency map)

$$\delta^{j,k}_{i,m} = \sum_{c_1, c_2 \in C \mid c_1 \neq c_2} |s'^{j,k}_{i,m,c_1} - s'^{j,k}_{i,m,c_2}|$$

Since weighting and differencing may have altered the score range possibly present in the explainability method, the resulting scores after weighting and differences would lie in an undefined range. In order to work with a specified range for convenience, we can proceed to normalize the scores by constraining them to be within the $[-1, 1]$ range.

The model we use to test our method is based on the BEiT architecture presented by Bao et al. in [19], that consists in a Vision Transformer pre-trained by using a masked image modeling approach tracing the original task employed in BERT architecture. We use a model (named `beit-base-patch16-224`

(a) Normal mucosa (b) Reactive/Benign (c) Premalignant (d) Malignant/Cancer

Figure 2: Examples of pictures for the four classes of oral lesions.

and released by Microsoft) pre-trained on ImageNet-21k (14 million images, 21,841 classes) at resolution 224x224, and fine-tuned on ImageNet 2012 (1 million images, 1,000 classes) at resolution 224x224. Furthermore, we fine-tuned the model on the dataset illustrated in dataset composed of 224 pictures representing different angles and framings of the inside of the mouth and including different kinds of oral lesions, partitioned in 4 classes: Malignant-Cancer, Normal Mucosa, Premalignant, Reactive-Benign. The images have been annotated by a team of doctors specialized in this type of lesions.

We rely on heatmaps produced by well-known explainability method (e.g. LIME), merging them in order to represent an aggregation revealing a different information with respect to the single heatmaps. Indeed the differences between two different heatmaps, associated with two different classes, would represent the most discriminative areas, that should be independent from any class. In the experiments we focused on LIME, presented by Ribeiro in [20].

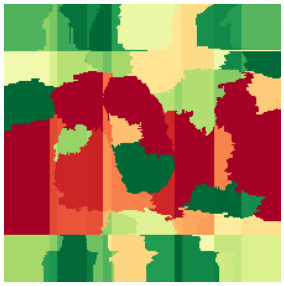
4. Results

For the purpose of this work, we have collected a dataset composed of 224 pictures representing the inside of the mouth as taken different angles and framings. The pictures show different kinds of oral lesions, partitioned into four classes: Malignant-Cancer, Normal Mucosa, Premalignant, Reactive-Benign. The images have been annotated by a team of doctors specialized in this type of lesions.

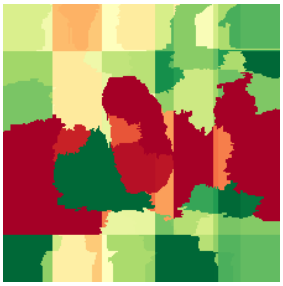
In Figure 2, we can see four examples pertaining to those classes. Those pictures differ significantly in terms of viewing angle, distance, and lighting conditions, making the identification effort more challenging. Figure 3a shows a Reactive/Benign lesion that is correctly classified by the model with nearly five-nine accuracy (see Table 1a, which reports the class assignment probability for the four classes). The most important part of the image is the excrescence on the patient's tongue, framed at the center of the picture. Green and red areas are the most important and the least important ones, respectively, driving the decision in LIME towards that class or another class. In our method, the green areas are the most discriminative ones, independent of the class, while the red areas are the least discriminative and do not influence the decision. We set $\alpha = 4$. The pictures in Figure 3c through Figure 3f show the single-class saliency maps, while Figure 3b shows the result obtained after applying our multi-class approach. We see that the single-class maps for the wrong decisions differ greatly from the single-class map for the right decision (Figure 3f). Our method allows us to recognize the areas that matter most, while ruling out some areas in the top and bottom parts of the picture that are clearly not relevant, as they mainly refer to the two black bands located there. Those areas appeared misleadingly light-green in some single-saliency maps. Also, the single-class saliency maps failed to recognize the excrescence as the most important element in the picture, as that area is totally coloured in red in Figure 3d and Figure 3e, and partly red in Figure 3c. The latter observation highlights another relevant shortcoming of that single-class saliency map, namely its lack of consistency across one image element: the excrescence (which is a relatively homogeneous element) is assigned quite different colours, located at the opposite end of the spectrum employed in the heatmap. Finally, the single-class saliency maps

(a) Original Image

(b) Our multi-class saliency map

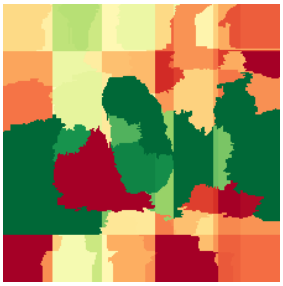


(c) LIME heatmap for Malignant-Cancer



(d) LIME heatmap for Normal Mucosa

(e) LIME heatmap for Premalignant



(f) LIME heatmap for Reactive-Benign

Figure 3: Heatmaps for a Reactive/Benign lesion testing case.

Class	Assignment Probability
Malignant/Cancer	$5.36 \cdot 10^{-6}$
Normal Mucosa	$6.14 \cdot 10^{-6}$
Premalignant	$8.1 \cdot 10^{-7}$
Reactive-Benign	0.999987

(a) Class Assignment Probability for the Reactive-Benign case of Figure 3

Class	Assignment Probability
Malignant/Cancer	$9.309673 \cdot 10^{-5}$
Normal Mucosa	0.29499677
Premalignant	$4.163922 \cdot 10^{-4}$
Reactive-Benign	0.70449376

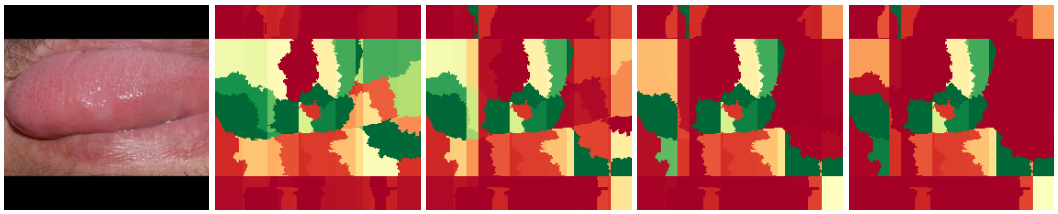
(b) Class Assignment Probability for the case of Figure 4

Table 1
Class Assignment Probabilities for the shown cases

share most areas, which are deemed relevant for all the classes. But this is quite contradictory, as an area that is relevant to all the classes bears no discriminative power and is then relevant to no class.

We can now consider a second case, where the ground truth is again the Reactive/Benign case, but some classification uncertainty is present that may influence the final output in the multi-class saliency map, due to the weights depending on the class assignment probability (see Figure 1a). The original picture is shown in Figure 4, where we can spot the big excrescence covering most of the central area of the picture, with the usual black belts located in the top and bottom areas. The class assignment probability values in Table 1b. While the correct class is identified with probability slightly larger than 70%, warranting correct classification under the majority rule, the Normal Mucosa decision takes nearly all the remaining 30%. This heavy presence of a wrong decision may drive the multi-class saliency map towards the single-class saliency map pertaining to that wrong classification.

We also consider the impact of the α values, showing the results for α in the [1,10] range. We see that all maps recognize the excrescence as the most relevant area and also highlight the left part where the excrescence is bigger. Finally, all maps correctly rule out the two black belts at the top and bottom of the picture. As to the impact of α, we expect that impact to be less relevant as the assignment gets more tilted towards one class (i.e., as the probability of a class approaches 1). Here, we are in an intermediate situation, where the majority assignment probability is not close to 1, so that we expect a significant



(a) Original image (b) Multi-class heatmap with $\alpha = 1$ (c) Multi-class heatmap with $\alpha = 4$ (d) Multi-class heatmap with $\alpha = 7$ (e) Multi-class heatmap with $\alpha = 10$

Figure 4: Impact α on multi-class saliency maps.

impact of α. Actually, we see a change as α grows, with the saliency maps being shifted towards extreme values, i.e., either dark green or dark red. In some cases, the shift turns some very-light-green areas into orange ones so as to capsize a weak result. The movement towards extreme values is, however, quite gradual, going through different shades of colour. The right choice about α depends on how much we are interested in a nuanced view of the discriminative areas with respect to how much we need to focus only on the most discriminative factors in the image. There is not a general rule, and each case should be analysed with respect to the planned use of the tool.

5. Conclusions

This study addresses a critical gap in explainable AI (XAI) for medical imaging, specifically in the classification of oral lesions. By introducing a multi-class saliency map, we overcome the limitations of traditional single-class interpretability methods, providing a more comprehensive and discriminative feature representation. Our approach effectively integrates saliency across multiple classes, offering improved clarity, reliability, and robustness in AI-driven diagnostics. The results demonstrate that this method successfully eliminates misleading feature attributions and highlights clinically relevant regions, making AI-assisted diagnosis more transparent and trustworthy for medical practitioners.

We can envisage some future areas of investigation to expand our work on the subject. In particular we plan to expand the dataset to include a broader range of oral lesions and imaging conditions (e.g., different ethnicities, age groups, and lighting variations). Also, we wish to evaluate the performance of our multi-class saliency approach against alternative interpretability methods. Finally, we would like to explore the application of our multi-class saliency in real-world medical AI workflows.

Acknowledgments

This work has been partially funded by the European Union–Next Generation EU within the framework of the PRIN 2022 Project “MEDICINE+AI, Law and Ethics for an Augmented and Human-Centered Medicine” (2022YB89EH) – CUP E53D23007020006

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] P. Fantozzi, M. Naldi, The explainability of transformers: Current status and directions, Computers 13 (2024) 92.

[2] P. Fantozzi, L. Laura, M. Naldi, Machine learning explainability as a service: Service description and economics, in: International Conference on the Economics of Grids, Clouds, Systems, and Services, Springer, 2024, pp. 244–253.
[3] E. Niebur, Saliency map, Scholarpedia 2 (2007) 2675.
[4] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).
[5] T. N. Mundhenk, B. Y. Chen, G. Friedland, Efficient saliency maps for explainable ai, arXiv preprint arXiv:1911.11293 (2019).
[6] W. Shimoda, K. Yanai, Distinct class-specific saliency maps for weakly supervised semantic segmentation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 218–234.
[7] X. A. López-Cortés, F. Matamala, B. Venegas, C. Rivera, Machine-learning applications in oral cancer: A systematic review, Appl. Sci. (Basel) 12 (2022) 5715.
[8] K. F. Hung, Q. Y. H. Ai, L. M. Wong, A. W. K. Yeung, D. T. S. Li, Y. Y. Leung, Current applications of deep learning and radiomics on ct and cbct for maxillofacial diseases, Diagnostics 13 (2022) 110.
[9] T. Thakuria, T. Rahman, D. R. Mahanta, S. K. Khataniar, R. D. Goswami, T. Rahman, L. B. Mahanta, Deep learning for early diagnosis of oral cancer via smartphone and dslr image analysis: a systematic review, Expert Review of Medical Devices 21 (2024) 1189–1204.
[10] M. P. Wu, G. Hsu, M. A. Varvares, M. G. Crowson, Predicting progression of oral lesions to malignancy using machine learning, Laryngoscope 133 (2023) 1156–1162.
[11] V. Rai, A. Chakrabarty, S. Bose, D. Pal, D. Bhattacharjee, F. Ahmed, S. R. Chowdhury, M. Maity, AI-driven smartphone screening for early detection of oral potentially malignant disorders, in: 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), IEEE, 2024.
[12] E. Duran-Sierra, S. Cheng, R. Cuenca, B. Ahmed, J. Ji, V. V. Yakovlev, M. Martinez, M. Al-Khalil, H. Al-Enazi, Y.-S. L. Cheng, J. Wright, C. Busso, J. A. Jo, Machine-learning assisted discrimination of precancerous and cancerous from healthy oral tissue based on multispectral autofluorescence lifetime imaging endoscopy, Cancers (Basel) 13 (2021) 4751.
[13] M. G. C. A. Cimino, G. Campisi, F. A. Galatolo, P. Neri, P. Tozzo, M. Parola, G. La Mantia, O. Di Fede, Explainable screening of oral cancer via deep learning and case-based reasoning, Smart Health 35 (2025) 100538.
[14] A. V. B. da Silva, C. Saldivia-Siracusa, E. S. C. de Souza, A. L. D. Araújo, M. A. Lopes, P. A. Vargas, L. P. Kowalski, A. R. Santos-Silva, A. C. de Carvalho, M. G. Quiles, Enhancing explainability in oral cancer detection with grad-cam visualizations, in: International Conference on Computational Science and Its Applications, Springer, 2024, pp. 151–164.
[15] K. C. Figueroa, B. Song, S. Sunny, S. Li, K. Gurushanth, P. Mendonca, N. Mukhia, S. Patrick, S. Gurudath, S. Raghavan, et al., Interpretable deep learning approach for oral cancer classification using guided attention inference network, Journal of biomedical optics 27 (2022) 015001–015001.
[16] M. Parola, F. A. Galatolo, G. La Mantia, M. G. Cimino, G. Campisi, O. Di Fede, Towards explainable oral cancer recognition: Screening on imperfect images via informed deep learning and case-based reasoning, Computerized Medical Imaging and Graphics 117 (2024) 102433.
[17] J. Adeoye, L.-W. Zheng, P. Thomson, S.-W. Choi, Y.-X. Su, Explainable ensemble learning model improves identification of candidates for oral cancer screening, Oral Oncology 136 (2023) 106278.
[18] R. O. Alabi, A. Almangush, M. Elmusrati, I. Leivo, A. Mäkitie, Measuring the usability and quality of explanations of a machine learning web-based tool for oral tongue cancer prognostication, International Journal of Environmental Research and Public Health 19 (2022) 8366.
[19] H. Bao, L. Dong, S. Piao, F. Wei, BEiT: BERT pre-training of image transformers, arXiv [cs.CV] (2021).
[20] M. T. Ribeiro, S. Singh, C. Guestrin, "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, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 1135–1144.