Biomedical Signal Processing and Control 111 (2026) 108337

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Multiclass classification of oral mucosal lesions by deep learning from clinical images without performing any restrictions

Alejandro Redondo [a], Katerina Ivaylova [b], Margarita Bachiller [a,*], Mariano Rincón [a], José Manuel Cuadra [a], Faleh Tamimi [c], José Luis López-Cedrún [d], Márcio Diniz-Freitas [e], Lucía Lago-Méndez [f], Guillermo Rubín-Roger [h], Jesús Torres [b], Leticia Bagán [g], Gonzalo Hernández [b], Rosa María López-Pintor [b]

[a] Department of Artificial Intelligence, National University of Distance Education (UNED), Juan del Rosal, 16, 28040 Madrid, Spain
[b] ORALMED Research Group, Department of Dental Clinical Specialties, Complutense University of Madrid, Plaza Ramón y Cajal, s/n, 28040 Madrid, Spain
[c] College of Dental Medicine, QU Health, Qatar University, Doha, Qatar
[d] Complexo Hospitalario Universitario de A Coruña, A Coruña, Spain
[e] Special Care Unit, OMEQUI Research Group, School of Medicine and Dentistry, Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela University, Santiago de Compostela, Spain
[f] EOXI Lugo-Monforte-Cervo, Galician Health Service (SERGAS), Chantada, Spain
[g] Department of Stomatology, Valencia University, Valencia, Spain
[h] Hospital Universitario de Cabueñes, Gijón, Asturias, Spain

ARTICLE INFO

Keywords:
Images classification
Oral cancer
Oral potentially malignant disorders
Deep learning
Convolutional neural network
Skip connection networks
Visual transformers
ConvNeXt

ABSTRACT

Oral cancer is a frequently malignant tumor that can be detected during an oral examination. Unfortunately, it is often diagnosed in advanced stages, which leads to low survival rates of about 50% at five years. Due to the low survival rate, it is crucial to develop automated systems that allow the classification of oral lesions according to their severity, aiding in the early diagnosis of oral cancer.

This study aims to investigate the effectiveness of using clinical images and deep learning based models to perform a multiclass classification of oral mucosal lesions in color photographs taken without following any acquisition protocol. The classification differentiated four classes: malignant, potentially malignant, benign and healthy. The dataset included a total of 3246 images from 1013 patients, with 40 different categories of oral lesions, including healthy oral mucosa. The images showed different areas of the oral cavity and were captured from different perspectives by diverse dentists and maxillofacial surgeons in the practice.

For the classification, different deep learning architectures were applied and compared, from the best known convolutional neural networks (CNN) and skip connection networks (SCN), to more innovative architectures such as visual transformers and a recent hybrid architecture, ConvNeXt v2. The ConvNeXt v2 Tiny architecture, with 85.53% accuracy, 85.02% precision, 85.50% recall, 84.92% F1-score, and 97.40% ROC AUC for an input image size of 354 × 354 pixels, outperformed the other architectures on the same database. The present model improved on previous proposals by considering a greater number of oral lesions and output classes.

1. Introduction

Oral and lip cancer is ranked 16th by the Global Cancer Observatory. The most common type is oral squamous cell carcinoma (OSCC), with the latest data in 2022 showing an incidence of 389,846 new cases and 188,438 deaths. The incidence and mortality of this type of cancer are most prevalent in Asia and Europe [1]. Oral cancer (OC) is a global health challenge, as it has a low survival rate of approximately 50% within five years of diagnosis [2]. This low survival rate has not improved in recent years despite established preventive programs and improved treatments for this tumor [3–5]. This may be because

* Corresponding author.
E-mail addresses: aredondo275@alumno.uned.es (A. Redondo), kateriva@ucm.es (K. Ivaylova), marga@dia.uned.es (M. Bachiller), mrincon@dia.uned.es (M. Rincón), jmcuadra@dia.uned.es (J.M. Cuadra), fmarino@qu.edu.qa (F. Tamimi), lopezcedrun@centromaxilofacial.com (J.L. López-Cedrún), marcio.diniz@usc.es (M. Diniz-Freitas), Lucia.Lago.Mendez@sergas.es (L. Lago-Méndez), grubin@hotmail.es (G. Rubín-Roger), jesus.torres@ucm.es (J. Torres), leticia.bagan@uv.es (L. Bagán), ghervall@ucm.es (G. Hernández), rmlopezp@odon.ucm.es (R.M. López-Pintor).

https://doi.org/10.1016/j.bspc.2025.108337
Received 17 April 2024; Received in revised form 7 May 2025; Accepted 27 June 2025
Available online 19 July 2025

OC is diagnosed at very advanced stages, which is associated with worse survival, higher treatment costs, and complications such as facial dysfunction and deformity [4,6]. Therefore, it is essential to improve diagnostic methods that favor an early diagnosis of OC.

Oral potentially malignant disorders (OPMD) are a group of oral lesions and conditions that increase the risk of developing OC [7,8]. The prevalence of OPMD varies from 1%–5% with significant geographic variation. Scientific evidence has shown that a small proportion of these OPMD can lead to OC [7].

One of the most common clinical signs of OC is the appearance of oral mucosa lesions. However, there is a wide variety of oral lesions, including benign, potentially malignant, and malignant lesions. Thus, early diagnosis of OC can be achieved by correctly diagnosing an oral lesion, placing it on a continuum from normal oral mucosa to malignancy. Dentists and other clinicians are therefore advised to be alert for signs of OPMD and OC [8].

Some studies have shown that patients with OC preceded by OPMD who attended regular check-ups were diagnosed with OC at an earlier stage, improving their survival [9]. This is clearly important as the 5-year survival rate increases to 75%–90% when OC is diagnosed at earlier stages [10,11]. Therefore, oral lesions should be systematically evaluated, and those suspicious of malignancy should be referred to a specialist for biopsy and subsequent histopathological analysis, which is considered the gold standard. However, many clinicians do not adopt these indications because they have difficulty diagnosing oral lesions and require further consultations before referring the patient to the appropriate specialist [12–16].

It is worth noting that diagnosis is complicated because there is a great variety of oral lesions with similar clinical appearance, and many systemic conditions can give rise to similar oral manifestations. Conversely, oral lesions are uncommon, and many clinicians lack experience diagnosing them. All this can delay the diagnosis of oral lesions, including OC and OPMD [14–16].

Oral lesions are visible and usually accessible when examining the oral mucosa and can be photographed with a smartphone or reflex camera for further analysis. Based on these images, a computer vision system could help diagnose and classify these oral lesions, facilitating the correct derivation of cases to the corresponding specialists and favoring the early diagnosis of OC and OPMD. However, the development of this system is complex and challenging due to the variability of the images, which can capture different areas of the oral cavity from different perspectives or even contain appliances (orthodontics, dentures) and the specialists' instruments.

During the past few years, research has focused on developing automatic systems for image analysis based on artificial intelligence techniques [17,18]. Many of these systems use deep learning techniques (DL), which have had a substantial impact on this type of problems because they allow to obtain generalizable results, where previously it was not possible due to the high economic and time cost of first obtaining relevant features and then modeling the relationships of these features. When DL techniques are employed, image variability is addressed in the proper definition of the dataset for model training, which is composed of varied samples representing the characteristics of the target population. DL techniques have been used for pigmented skin lesion classification [19], lung cancer classification [20], brain tumor detection and classification [21] and particularly in the analysis of oral lesion images [22–24].

The main objective of this study is to investigate the effectiveness of using clinical images and DL based models to perform a multiclass classification of oral mucosal lesions in color photographs taken without restrictions. The model will classify any image of oral lesions into four classes according to their severity: healthy, benign, potentially malignant, and malignant. The input to the model will be the digital color photograph taken from any perspective and without following any acquisition protocol. Based on our own dataset, which is larger and more diverse than other existing ones, we will perform a comparative study between the most commonly used and successful architectures for classifying oral lesions in recent years and the alternative proposed in this research ConvNeXt v2. This comparison, realized under the same conditions, will allow the selection of the best architecture for the classification of oral lesions prioritizing the best model in terms of diagnostics.This classifier may be used in an automatic system to assist clinicians in the early diagnosis of OC and OPMD. Using just one photograph, the model will provide a preliminary diagnosis to assess the severity of the case and determine if a referral to a specialist is necessary.

The remainder of the paper will be structured as follows. Section 2 describes previous studies on DL techniques applied to oral lesions classification. Section 3 describes the materials and methods. Section 4 shows the experiments done. Section 5 reflects the analysis of the results. Finally, Sections 6 and 7 shows the discussion and conclusions of the study, respectively.

2. Related works

Recent studies have employed DL algorithms to diagnose OC due to their ability to identify patterns in complex visual data [24–26]. Publications have highlighted three groups of neural networks for classifying oral lesions: convolutional neural networks (CNNs) [27], skip connection networks (SCN) such as ResNet [28] or DenseNet [29], and the more recent vision transformers (ViTs) strategy [30]. The CNN architecture is composed of a convolutional module, whose objective is to extract the feature vector by passing through a series of filters, and a fully connected classification module that interprets the resulting feature vector to assign the image to a specific category. The SCN architecture differs from CNN in the convolutional module as there are now skip connections from one layer to another deeper layer when the network has multiple layers. ViT also differs from CNN in the first module of the architecture. It uses self-attention mechanisms to understand all pixels simultaneously instead of convolutions that only consider pixels belonging to the filter kernel.

CNNs [31] are characterized by their ability to capture local and spatial patterns in images through convolution layers that extract features at different levels of abstraction, and pooling layers, which reduce dimensionality while preserving the most relevant features. Pandit et al. [32] showed a brief description of CNNs applications in computer vision and natural language processing. Various architectures, including VGGNet [33], EfficientNet [34], Inception [35], MobileNet [36] and HRNet [37], have been applied to OC. Shamim et al. [38] used the VGGNet architecture to perform a binary (benign and precancerous) and a multiclass classification of five classes of tongue lesions. Similarly, Welikala et al. [39] employed the same architecture for a binary classification. Tanriver et al. [40] used the EfficientNet and Inception architectures to automatically classify oral lesions into three classes: benign lesions, OPMD and carcinomas. Gomes et al. [41] also employed Inception-v3 to perform a multiclass classification of six categories in different areas of the oral cavity. In a different study, Birur et al. [42] used the MobileNet architecture to detect malignant and OPMD lesions in resource-poor areas, specifically in India. They compared the performance of MobileNet, InceptionV3, and VGG19 and showed that MobileNet significantly reduced the number of parameters and model size with minimal difference in accuracy. On the other hand, Lin et al. [43] used the HRNet architecture to build a multi-class detection model for five categories: no lesion, aphthous ulcers, low-risk OPMD lesion, high-risk OPMD lesion and OC.

SCN were developed as an evolution of the CNN to improve training by using skip connections. This reduces the vanishing gradient and provides richer features. Within this neural network several architectures have been used such as ResNet and DenseNet. Shamin et al. [38] used the ResNet50 architecture to conduct multiclass classification involving five types of tongue lesions. Similarly, Warin et al. [44] employed the same architecture for the binary classification of OPMD

Table 1
Summary of metrics of the analyzed articles.

Architecture	Type	Dataset size	Balanced dataset	Number of classes	Number of lesions	Lesion zone	TL	Type of image	Unique perspective	Capture protocol	Size	Accuracy	Precision	Recall	F1-Score
VGGNet-19 [38]	CNN	200	Yes	2	8	Tongue	Yes	Whole Image	Yes	Yes	224	98	–	89	–
VGGNet-19 [39]	CNN	2155	No	2	–	Oral Mucosa	Yes	Whole Image	Yes	Yes	224	80.88	77.06	85.71	81.16
EfficientNet-B4 [40]	CNN	684	No	3	30	Multiple Zones	Yes	ROI	No	No	380	–	86.9	85.5	85.8
Inception-v4 [40]	CNN	684	No	3	30	Multiple Zones	Yes	ROI	No	No	299	–	87.7	85.5	85.8
Inception-v3 [41]	CNN	5069	No	6	6	Multiple Zones	Yes	ROI	No	No	299	95.09	86.32	85.25	85.45
MobileNet [42]	CNN	32 128	No	2	7	Multiple Zones	Yes	Whole Image	No	No	–	79	–	82	–
HRNet [43]	CNN	1448	No	5	–	Multiple Zones	Yes	ROI	No	No	512	–	84.3	83	83.6
ResNet-50 [38]	SCN	200	Yes	5	8	Tongue	Yes	Whole Image	Yes	Yes	224	97	–	–	–
ResNet-50 [44]	SCN	600	Yes	2	5	Multiple Zones	Yes	ROI	No	Yes	224	–	92	98.39	95
DenseNet-121 [45]	SCN	6176	No	2	26	Multiple Zones	Yes	Whole Image	No	No	224	84.1	–	89.6	–
DenseNet-169 [48]	SCN	980	No	3	4	Multiple Zones	Yes	ROI	No	No	224	–	95–98	95–99	95–98
DenseNet-121 [46]	SCN	700	Yes	2	–	Multiple Zones	Yes	ROI	No	No	224	–	100	98.75	99
DenseNet-201 [47]	SCN	2178	No	2	7	Multiple Zones	Yes	Whole Image	No	No	224	–	86	85	86
ResNet-101 [49]	SCN	2155	No	5	16	Multiple Zones	Yes	Whole Image	No	No	224	–	52.13	49.11	50.57
ViT [47]	ViT	2178	No	2	7	Multiple Zones	Yes	Whole Image	No	No	224	–	77	77	77
Swin [47]	ViT	2178	No	2	7	Multiple Zones	Yes	Whole Image	No	No	224	–	86	86	86

and images without oral lesions. Fu et al. [45], Warin et al. [46], and Talwar et al. [47] used different versions of the DenseNet architecture to classify oral lesions into two classes: OSCC or healthy oral mucosa. Meanwhile, Warin et al. [48] used the same architecture to consider three output classes, differentiating between OPMD, OSCC, and healthy.

On the other hand, among the systems focused on detecting oral lesions is Welikala et al. [49] that present as alternative the use of the Faster R-CNN architecture [50] together with ResNet-101 to jointly detect different high and low risk OPMDs and separating them into five classes. Another example is presented by Warin et al. [44] which uses the same Faster R-CNN architecture to locate OPMD disorders but now the classification is performed using the ResNet-50 architecture.

In recent years, Vision Transformers (ViTs) have emerged [51] inspired by the Transformers widely used in natural language processing (NLP) tasks with results that outperform those of recurrent neural networks. ViTs incorporate attention mechanisms that relate different image patches, although their drawback lies in their quadratic computational complexity. Swin Transformers [52] improve ViTs performance by incorporating hierarchical feature maps and shifted windows, which makes transformers a benchmark given their remarkable performance on a wide range of vision tasks. A study presented by Talwar et al. [47] used both models to distinguish between two categories of suspicious and non-suspicious oral lesions.

Table 1 summarizes the relevant information from the previous studies, including the architecture used, number of images in the dataset, number of classes and oral lesions considered, area of the oral cavity under study, use of transfer learning (TL) with pre-trained ImageNet weights, type of input to the model differentiating between whole image or region of interest (ROI) that includes the lesion, the use or not of a protocol for image acquisition, size of the input images to the model and metrics used for the evaluation. As can be observed, prior studies have used datasets with a limited number of images and most not exceeding 2500 images (7 of them do not exceed 1000), except [41,42,45].

Precisely, the major drawback of all these DL strategies is that the number of available images to train the models is usually small. Consequently, these systems exhibit limitations, addressing a diminished scope of oral lesions and offering fewer output classes. One way to achieve satisfactory results is to reduce the problem. Table 1 shows that the studies with better results simplify the problem by concentrating on a particular region of the oral cavity, such as the tongue [38], analyzing fewer lesion categories [38,41,42,44,47,48] or focusing on the region containing the lesion rather than the whole image [40,41,43,44,46,48]. This reduces the possibility of finding artifacts, such as gloves or medical instruments, and other oral cavity structures, such as teeth. When more lesion categories and the complete image are considered as input, the results worsen significantly [47,49].

Among the most recent DL techniques are hybrid architectures that combine elements from different model types to leverage their individual strengths. One such architecture is ConvNeXt [53], which integrates elements from ViTs, such as GELU activation, asymmetric stage design, and large convolutional kernels, into CNNs. These hybrid architectures increase their ability to model long-range relationships, limited in CNNs, and they achieve good performance on small datasets, improving to the ViTs. ConvNeXt retains the typical CNN structure with convolutional layers and introduces a Global Average Pooling (GAP) layer, which flattens the output of the convolutional layers into a single feature vector. It also incorporates “batch normalization” and “non-linear activation functions”, which enhance learning performance and robustness against overfitting. A variant of this architecture, ConvNeXt v2 [54], introduces enhancements such as Fully Convolutional Masked Autoencoders for self-supervised pretraining and a novel Global Response Normalization (GRN) layer to strengthen inter-channel feature competition. This second generation comes in multiple versions, including ConvNeXt V2 Tiny, optimized for deployment in resource-constrained environments, and ConvNeXt V2 Base, designed to maximize accuracy and overall performance. ConvNeXt v2 outperforms well-known architectures in various tasks, including ImageNet classification [53], COCO [55] object detection, and ADE20K [56] segmentation, as indicated by Woo et al. [54]. Additionally, these innovative architectures have recently been applied in medicine for mammogram classification [57] and blood vessel segmentation [58], making it a promising alternative for oral lesion classification.

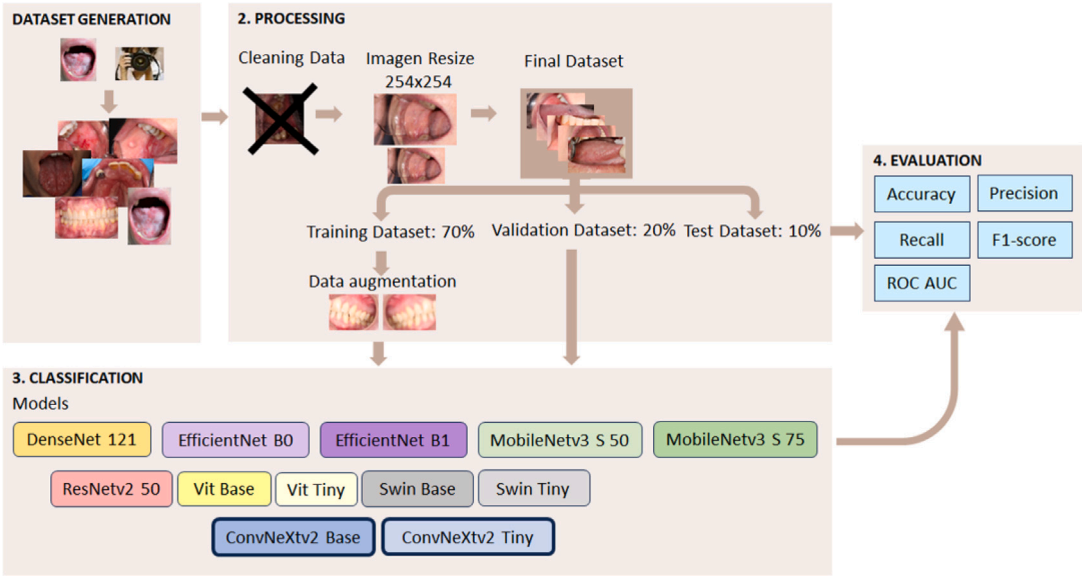

Fig. 1. Overview of the approach for oral lesion classification.

3. Materials and methods

In order to carry out this study, it was first necessary to create a representative and clean sample of the problem. Next, a total of 12 DL models were implemented and evaluated with 5 metrics. Fig. 1 provides an overview of the approach.

3.1. Dataset

For this study, a dataset of 3270 oral mucosal images was generated from 1013 white patients. The images of the lesions were taken by various oral medicine specialists, maxillofacial surgeons, and dentists in Spain. The study protocol received approval from the Ethics Committee of the Hospital San Carlos de Madrid (no. 21/187-E) and was conducted in accordance with the ethical principles of the Declaration of Helsinki.

The final dataset of this study included RGB images in JPEG format of the oral cavity taken with reflex cameras (with or without macro lens or flash) and smartphone cameras of any brand or model. No specific protocol or restrictions were followed for image capture. The photographs were acquired from different perspectives and distances, and included artifacts such as dental appliances, dental restorations, gloves used by the practitioner, gauzes, separators or dental mirrors. Images from all locations of the oral mucosa were included, from the external aspect of the lips to the oropharynx. The distribution of the number of images of each class according to location is shown in Table 2, in which six locations have been distinguished, covering images acquired from any perspective: gingiva and alveolar ridge, buccal mucosa, palate, floor of the mouth, tongue, lip and oropharynx. Note that the sum of the images in the table is greater than the number of images contained in the dataset since some of them present oral lesions visible in more than one oral location.

No images that could identify patients were collected. ROI was not marked. Only images with sufficient definition to correctly visualize the oral mucosa with or without oral lesions were selected. The images were of varying resolutions, with an average image size of 3456 × 5184 pixels and a standard deviation of 806 × 554 pixels. Fig. 2 presents several examples extracted from the database used in this study that show the variability of the images included.

The oral lesions were clinically diagnosed by oral medicine specialists and maxillofacial surgeons who classified them into four groups: healthy, benign, OPMD, and OC. Malignant and OPMD were also diagnosed histologically according to current criteria. Histological diagnoses were also made for certain benign lesions, such as exophytic, blistering, or granulomatous lesions.

The OC group primarily consisted of OSCC images but also included melanoma and verrucous carcinoma cases. The OPMD group comprised oral lesions as defined in the WHO Collaborating Centre for Oral Cancer Consensus Report 2020 [59]. The most common OPMDs were oral lichen planus, leukoplakia, and proliferative verrucous leukoplakia. The group of benign lesions includes benign tumors, fungal and viral infections, benign ulcerative lesions, white lesions, lingual changes, benign pigmented lesions, autoimmune blistering disorders, hypersensitivity

Table 2
Number of images of each class separated by location.

Localization	OC	OPMD	Benign	Healthy
Gingiva and Alveolar ridge	144	464	372	94
Buccal mucosa	48	473	136	107
Palate	26	61	70	71
Floor of the mouth	36	36	13	51
Tongue	107	183	234	226
Lip	10	22	185	54
Oropharynx	6	7	12	2

Fig. 2. Different perspectives of the oral cavity under different lighting conditions. Different appliances or instruments used by dentists can also be observed.

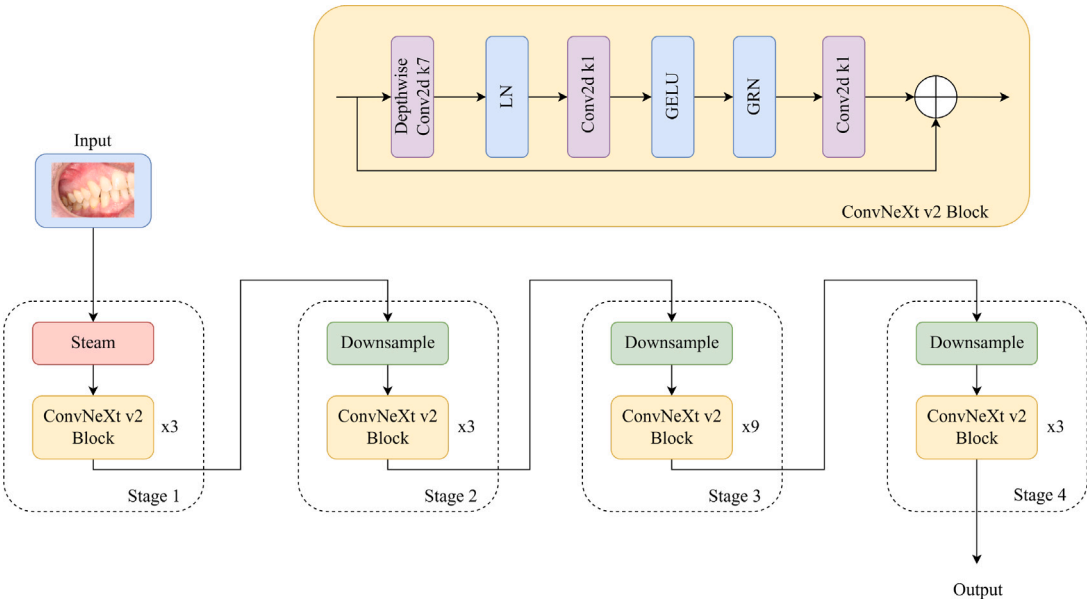

Fig. 3. ConvNeXt v2 Tiny architecture.

reactions, and lip lesions, among others. Additionally, up to 610 images of healthy oral mucosa were included.

3.2. Data processing

The first step was to clean the data, which involved an automatic part, where images that did not have three channels (not RGB) were eliminated. A manual review was also carried out, in which blurred or out-of-focus images were discarded. A total of 24 images were removed from the original dataset. Consequently, the final dataset comprised 3246 images.

Table 3 presents the final distribution of images, showing the subgroups of oral lesions considered in each output class/group, the specific types of lesions included in each subgroup and the number of images available for each of them.

To conduct the experiments, we used stratified sampling. The final dataset was divided into three subsets: 70% for training, 20% for validation, and 10% for testing. We verified that each of these subsets contained representative samples of every class (Table 4). Furthermore, the training data underwent online data augmentation using the Torchvision Transforms library. This was done to account for the various conditions that may be encountered during the daily use of the application, such as rotations, flips, changes in contrast, gaussian blurs, and perspective changes.

3.3. Architectures

In the classification phase, we apply the most widely used DL architectures in oral lesion classification. Particularly the used models were EfficientNet B0, EfficientNet B1, MobileNetv3 S 50, MobileNetv3 S 75, ResNetv2 50, DenseNet 121, Vit Base, Vit Tiny, Swin Base and Swin Tiny. In addition, as a novelty, this work proposes the use of the ConvNeXt v2 architecture, which is briefly presented below. A detailed description can be found at [54].

The EfficientNet B0, EfficientNet B1, MobileNetv3 S 50, Mobile Netv3 S 75 architectures are included in the CNN subgroup. EfficientNet [34] modifies the depth or number of layers and the width or number of channels per layer. The difference between EfficientNet B0 and EfficientNet B1 is the number of layers of the intermediate blocks. MobileNetv3 [60] reduces the complexity introducing depthwise separable convolutions, the linear bottleneck and inverted residual structure. ResNetv2 50 and DenseNet 121 architectures belong to the SCN subgroup. These architectures are characterized by using skip connections. ResNet [61] takes a previous output as an input for a future layer, and DenseNet [29] takes all previous outputs as inputs for a future layer. There are different versions of both architectures. The difference between them is in the number of layers. Finally, the ViT Base, ViT Tiny, Swin Base, Swin Tiny [51,52] belong to the ViT subgroup. Their architecture leverages the self-attention mechanism, enabling the model to interpret images holistically. However, ViT models have huge number of parameters. The Tiny versions, pretrained on large-scale datasets, was created to avoid this problem. On the other hand, Swin is an enhancement to ViT which employs a hierarchical structure, progressively merging patches into larger representations at different resolutions.

The architecture of ConvNeXt v2 [54,62], shown in Fig. 3, consists of four stages, each containing different blocks. These blocks can be of three types: Stem, Downsample, and ConvNeXt v2. The Stem block, present only in the first stage, is responsible for the initial image processing. The Downsample block, located at the beginning of the other stages, generates a multiscale hierarchical feature map. Finally, the ConvNeXt v2 block, which is the fundamental block of the model, present in all states. This last block, shown in Fig. 3, contains the novel non-linear activation gaussian error linear unit (GELU) [63] and the global response normalization (GRN) [54] module. GRN aims to improve the feature map by increasing contrast and channel selectivity.

This study uses two versions of the architecture, Tiny and Base [54]. Tiny version (Fig. 3) consists of 4 stages, each with a different number of ConvNeXt v2 Blocks (3, 3, 9, and 3, respectively). Base version also consists of 4 stages with a varying number of ConvNeXt v2 Blocks (3, 3, 27, and 3, respectively). Thus, the initial part of this architecture remains identical to that depicted in Fig. 3. However, in stage 3, the number of ConvNeXt v2 Blocks increases from 9 to 27. Consequently, the Tiny version, with approximately 28M parameters, is less dense than the Base version, which has almost 87M parameters.

3.4. Hyperparameter configuration and model training

The hyperparameter settings used in model training were the same for both models and included a maximum of 100 epochs, transfer learning (TL) from ImageNet [53], and batch size of 8 images. The appropriate batch size was determined through experimentation with different values (4, 8, 16, and 32), being 8 the optimum.

Table 3
Distribution of the 3246 images used in this work in class/group, subgroups-types of lesion for each output class.

Class/Group	Subgroup-Lesion type		Number of images	Total
OC	Oral squamous cell carcinoma		329	347
	Melanoma		16	
	Verrucous carcinoma		2	
OPMD	Oral lichen planus		886	1238
	Leukoplakia		228	
	Proliferative verrucous leukoplakia		113	
	Other potentially malignant disorders		11	
Bening	Benign tumors	Fibroma	101	1056
		Angioma	54	
		Torus and exostosis	46	
		Pyogenic granuloma	41	
		Papilloma	37	
		Mucocele	15	
		Gingival enlargement	15	
		Other benign tumors	20	
	Fungal-associated infections	Erythematous candidiasis	27	
		Angular cheilitis	13	
		Pseudomembranous candidiasis	12	
	Viral infections	Recurrent oral herpes simplex	43	
		Primary herpetic primoinfection	12	
	Benign ulcerative lesions	Recurrent aphthous stomatitis	38	
		Behçet disease	18	
		Traumatic ulcers	79	
		Other benign ulcers	14	
	Benign white lesions	Frictional hyperkeratosis	45	
		Linea alba	14	
		Other benign white oral lesions	52	
	Benign lingual alterations	Coated tongue	23	
		Geographic tongue	23	
		Other non-pathological alterations of the tongue	20	
	Pigmented lesions of the oral mucosa	Oral melanotic macule	30	
		Other non-pathological pigmentations of the oral mucosa	6	
	Autoimmune blistering disorders	Pemphigoid	142	
	Hypersensitivity reactions	Angioedema	43	
		Contact stomatitis	27	
		Other hypersensitivity reactions	33	
	Lip lesions	Desquamative lip lesions	7	
		Granulomatous cheilitis	3	
	Other lesions	Tonsillitis	3	
Healthy	Healthy		605	605

Table 4
Number of samples in each subset of data obtained with random seed 13.

Number of images per subset	Training images	Validation images	Test images	Total number of images
Malignant	236 (68.01%)	78 (22.48%)	33 (9.51%)	347 (100%)
Potentially malignant	880 (71.08%)	223 (18.02%)	135 (10.9%)	1238 (100%)
Benign	725 (68.65%)	227 (21.5%)	104 (9.85%)	1056 (100%)
Healthy	431 (71.24%)	121 (20%)	53 (8.76%)	605 (100%)

On the other hand, the models used advanced hyperparameters such as the Nadam (Nesterov-accelerated Adaptive Moment Estimation) optimizer [64], categorical cross entropy [65] for the loss function, ReduceLROnPlateau [66] for the learning rate, ImageNet [53] for TL, L2 regularization [67], and early stopping.

The algorithms required between five and ten hours to be fully trained in a high-performance environment with Nvidia V100 GPUs. This is due to the large number of mathematical operations, such as derivatives, needed to fit the nearly 28 million parameters present in ConvNeXt v2 Tiny. However, the inference time was reduced to 10–50 ms per image.

4. Model evaluation

4.1. Metrics

The purpose of model evaluation [68] is to assess its ability to make accurate predictions on unseen data (test set) after training. In this study, we conducted a quantitative analysis using various metrics to compare models trained on the same dataset. The evaluation metrics employed were accuracy, precision, recall, and F1-score.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = 2 * \frac{recall * precision}{recall + precision} \tag{4}$$

where, for a given class C_i, TP, TN, FP, FN represent the True Positives, True Negatives, False Positives and False Negatives, respectively.

A valuable tool for evaluating the performance of classification models is the Receiver Operating Characteristic (ROC) curve, along with its associated metric—the Area Under the Curve (AUC-ROC) [69]. The ROC curve is a graphical representation that shows the relationship between the true positive rate and the false positive rate across different decision thresholds. This visualization allows for analyzing the model's ability to discriminate between classes, which can assist physicians in selecting the most appropriate threshold to optimize the correct identification of a specific class over the others. The AUC-ROC metric provides a scalar value between 0 and 1 that summarizes the overall performance of the classifier. This value is particularly useful for comparing different classification models, especially in contexts with imbalanced classes, where other metrics such as accuracy may be misleading due to their bias toward the majority class.

4.2. Experiments to evaluate the study

This study presents two sets of experiments designed to evaluate algorithm performance using the generated dataset. The first experiment used the most common architectures in the classification, as well as the proposed architecture, ConvNeXt v2. The objective in all cases was to classify the severity of oral lesions into four classes: healthy, benign, potentially malignant, and malignant. The experiment allowed the selection of the three most suitable architectures based on metrics. In the second experiment, we examined the impact of the input image size on the three previously chosen models to determine the optimal solution.

4.2.1. Experiment 1. Choice of architectures

As previously stated, this experiment aimed to identify the top three performing configurations, that is, those that demonstrated superior performance across the most metrics. Two key factors were considered for each architecture: the use of TL and the density of the architecture. Firstly, for the same architecture, more and less dense models were chosen, for example, from the ConvNeXt v2 architecture, the Base and Tiny models were selected, with Tiny having fewer parameters than Base. For each selected model, two runs were performed: one using TL from ImageNet [53] to initialize the model parameters and the other using the gaussian initializer. To ensure result comparability, we standardized the input image size to 224 × 224 pixels and consistently split the dataset across all configurations (see Table 4). We opted for an image size of 224 × 224 pixels due to constraints in computing resources. In addition, we repeated random train-validation-test splits three times to provide a reliable estimate of model performance. The outcomes were consistent, with metric values differing by no more than 1.5 points.

4.2.2. Experiment 2. Choice of image size

The objective of this experiment was to determine the optimal input image size. The choice of image size represents a trade-off between image quality and processing time. The larger the image size, the more detail is preserved in textures, edges, and small regions, which can help improve classification accuracy, but it could also lead to the model being overfitted to irrelevant details when a large dataset is not available. On the other hand, processing time must also be considered, as it increases with larger image sizes.Taking this into account, three image resolutions have been evaluated: 224 × 224, 384 × 384, and 512 × 512 pixels. These image sizes, which are no larger than 512 pixels per dimension, facilitate training on GPUs with 16 GB of VRAM. Consequently, predictions entail minimal computational demands, facilitating the use of these models by users without access to high-performance environments.

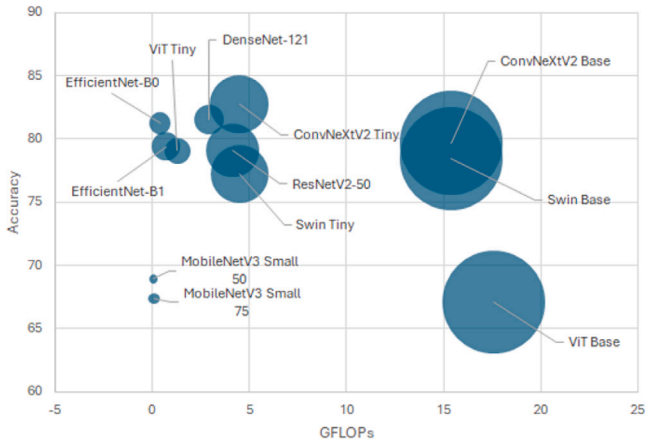

Fig. 4. Bubble chart in which each circle represents a model. The horizontal position indicates the GFLOPs, the vertical position the accuracy and the circle size the number of parameters for 224 × 224 pixel input images.

5. Results

Table 5 presents the results of the first experiment, including the model name, neural network type, whether TL was applied, and the evaluation metrics on the test set. Upon examining this table, it is evident that TL was essential in achieving a satisfactory performance. Additionally, the use of TL decreased the number of training epochs required (a good model was obtained within the first 20 epochs). Conversely, performances without TL exhibited slower convergence rate, were more unstable, and yielded inferior results. It was observed that less dense models performed better for the same architecture, likely due to the dataset size. Thus, the architectures that exhibited the best metrics, as shown in Table 5, were ConvNeXt v2 Tiny, DenseNet 121, and EfficientNet B0, which were trained using TL and an input image size of 224 × 224 pixels, which was the resolution mostly used in the studies presented in Table 1. These architectures achieved above 80% on all metrics.

On the other hand, a comparison of all these models in terms of computational and spatial complexity, when the input to the network is an image of size 224 × 224 pixels, is shown in Table 6, whose values have been extracted from the original articles [29,34,51,52,54,60,61]. Computational complexity is measured in GFLOPs (Giga Floating Point Operations per Second) which indicates how many billions of floating point operations are required to process an input through the network. Spatial complexity is the amount of memory a model needs during its execution, including the storage of the model parameters (weights and biases). It is observed that the ConvNeXt v2 Tiny model has a worse computational complexity than CNNs and SCNs (see Table 6), however performance metrics are better (see Table 5). This idea is clearly observed in the bubble chart of Fig. 4, in which the GFLOPs, accuracy and number of parameters of all models are represented. It can also be observed that the model with the lowest value of GFLOPs is MobileNetv3 S 50 however its metric scores are in the order of 70%.

Table 7 presents the results of the second experiment, including the model name, neural network type, size of the input images, and performance metrics obtained on the test set for the three best architectures (ConvNeXt v2 Tiny, DenseNet 121 and EfficientNet B0). This table indicates that the worst performance occurred with images

Table 5
Results of Experiment 1: Performance obtained with different architectures. The three best values for each of the metrics have been marked in bold and the three best models with the best overall values have been painted in green.

Architecture	Type	TL	Accuracy	Precision	Recall	F1-score	ROC AUC
ConvNeXt v2 Base	Hybrid	Yes	79.69	79.03	78.98	78.37	**95.33**
ConvNeXt v2 Base	Hybrid	No	46.76	54.57	37.06	38.84	70.58
ConvNeXt v2 Tiny	Hybrid	Yes	**82.76**	**81.96**	**83.15**	**82.41**	**95.46**
ConvNeXt v2 Tiny	Hybrid	No	44.0	45.71	39.26	40.32	73.96
DenseNet 121	SCN	Yes	**81.53**	**81.77**	**82.42**	**81.75**	95.31
DenseNet 121	SCN	No	55.38	56.97	58.49	55.63	80.90
EfficientNet B0	CNN	Yes	**81.23**	78.35	**82.49**	**79.86**	**95.59**
EfficientNet B0	CNN	No	48.30	45.70	50.40	46.95	75.02
EfficientNet B1	CNN	Yes	79.38	78.31	79.86	79.04	93.85
EfficientNet B1	CNN	No	41.23	30.93	31.07	30.23	59.06
MobileNetv3 S 50	CNN	Yes	68.92	69.73	72.11	70.38	89.96
MobileNetv3 S 50	CNN	No	51.38	51.96	47.64	48.05	77.63
MobileNetv3 S 75	CNN	Yes	67.38	69.19	71.29	68.65	89.57
MobileNetv3 S 75	CNN	No	50.15	47.79	49.86	48.35	76.84
ResNetv2 50	SCN	Yes	79.07	78.06	81.16	79.15	94.02
ResNetv2 50	SCN	No	52.0	51.44	57.44	52.33	79.16
Swin Base	ViT	Yes	78.46	77.18	80.12	78.37	93.29
Swin Base	ViT	No	43.69	30.23	35.42	31.28	69.61
Swin Tiny	ViT	Yes	77.23	77.91	76.12	76.44	93.90
Swin Tiny	ViT	No	46.46	45.11	50.33	46.56	74.47
ViT Base	ViT	Yes	67.07	66.58	68.07	67.02	88.62
ViT Base	ViT	No	46.15	44.29	38.71	39.65	68.34
ViT Tiny	ViT	Yes	79.07	**79.62**	77.72	78.56	94.86
ViT Tiny	ViT	No	43.38	49.50	40.98	42.46	72.67

Table 6
Computational and spatial complexity of DL models.

Architecture	GFLOPs	Parameters
ConvNeXt v2 Base	15.4	87 696 900
ConvNeXt v2 Tiny	4.47	27 869 572
DenseNet 121	2.9	6 957 956
EfficientNet B0	0.39	4 012 672
EfficientNet B1	0.7	6 518 308
MobileNetv3 S 50	0.06	572 324
MobileNetv3 S 75	0.09	1 020 972
ResNetv2 50	4.12	23 508 548
Swin Base	15.4	86 747 324
Swin Tiny	4.5	27 522 430
ViT Base	17.6	85 801 732
ViT Tiny	1.3	5 525 188

scaled to 224 × 224 pixels, whereas the best outcomes were consistently achieved with images sized at 384 × 384 pixels. Among them, the ConvNeXt v2 Tiny model stands out again, with values for 384 × 384 pixel images of 85.53% accuracy, 85.02% precision, 85.50% recall, 84.92% F1-score and 97.40% ROC AUC.

6. Discussion

This study analyzed the performance of twelve DL models for classifying the severity of oral lesions in images into four categories: healthy, benign, potentially malignant and malignant. The study included 40 specific types of oral lesions, including healthy oral mucosa, located anywhere in the oral cavity. In addition, no protocol was established to capture the images, and it was not required for the clinician to specify the location of the lesion, since the input to the model was the complete image. In previous studies, the number of lesions is significantly lower, and images are usually classified into just two categories (see Table 1). Studies that use the whole image as input to the model analyze 26 or fewer lesion types. This number increases to 30 when the model input is a ROI, in which performance improves slightly due to the reduced input information. However, these models require localization of the ROI by the clinician. Our proposal expands the number of oral lesions and output classes without the need for clinician intervention.

As a preliminary step, we created a new database with images from various clinicians. The images were annotated and reviewed by specialists in the field such as dentists specialized in oral medicine and maxillofacial surgeons. As the images came from different clinicians, it was necessary to clean the dataset by removing those that were too small, not in RGB format or blurred.

The experiments enabled the comparison of several well-known DL architectures on the same image dataset. Among the three best architectures, ConvNeXt v2 Tiny (this study's novel proposal) standed out slightly when the input size was 254 × 254 pixels, which was the usual size in the literature. In addition, in this study, we experimented with variations in the size of the input image, always considering the dataset size and its use in devices with limited computational resources, such as mobile phones. The ConvNeXt v2 Tiny architecture, using 384 × 384 input images, demonstrated the best performance for oral lesion classification, achieving an accuracy of 85.53%, a precision of 85.02%, a recall of 85.50%, an F1-score of 84.92%, and a ROC AUC of 97.40%. However, the values in the metrics of the EficientNet B0 model also showed good performance, close to the ConvNeXt v2 Tiny model for all image sizes considered.

If we observe the studies summarized in Table 1 that used the complete image as input and different areas of the oral cavity, DenseNet-201 [47] stands out with a F1-score of 86%. However, in this study, only seven different oral lesions and two output classes are considered, significantly reducing the learning difficulty. In the case of DenseNet-121 [45], with an accuracy of 84.1%, only 26 oral lesions and 2 output classes are considered. These values are similar to those achieved with ConvNeXt v2 Tiny but, in our case, we employed a higher number of output classes and oral lesion types. In Table 1 it is observed that when the number of classes is increased as in [49], the F1-score is reduced to 50.57%. Therefore, our proposal has considerable improvements over previous studies. For a future application to help diagnose oral lesions by assisting dentists and other medical professionals, it is important to classify them into at least four classes. It is necessary to distinguish benign lesions, which will not progress to OC, from OPMDs that may progress to OC and need to be biopsied and periodically reviewed by

Table 7
Results Experiment 2. Performance obtained with different architectures and different image sizes. The best value for each of the metrics has been marked in bold and the model with the best overall values has been painted green.

Architecture	Type	Size	Accuracy	Precision	Recall	F1-score	ROC AUC
ConvNeXt v2 Tiny	Hybrid	224	82.76	81.96	83.15	82.41	95.46
ConvNeXt v2 Tiny	Hybrid	384	**85.53**	**85.02**	85.50	**84.92**	**97.40**
ConvNeXt v2 Tiny	Hybrid	512	83.07	83.39	82.64	82.94	96.92
DenseNet 121	SCN	224	81.53	81.77	82.42	81.75	95.31
DenseNet 121	SCN	384	84.30	82.23	85.05	83.16	96.52
DenseNet 121	SCN	512	81.84	79.45	82.59	80.67	95.19
EfficientNet B0	CNN	224	81.23	78.35	82.49	79.86	95.59
EfficientNet B0	CNN	384	84.92	83.22	**86.90**	84.69	97.22
EfficientNet B0	CNN	512	82.15	80.05	84.83	81.94	96.77

specialists. Periodic review of OPMDs by specialists can aid in the early diagnosis of OC. Early detection of OC will reduce treatment costs and impact patients' quality of life. Moreover, OC lesions should be referred with the utmost urgency for treatment by the maxillofacial surgeon. We believe that since specialized medical care is expensive and insufficient in some countries, it is necessary to prioritize urgent cases to provide adequate service and contain costs. If only two classes (benign and malignant) were handled, a coarse classification is being made with the risk of not making an early diagnosis of OC. Therefore, the four output diagnoses proposed in this study give a functionality and an answer to the reality of the diagnostic problem of oral lesions surpassing other previous tools.

The study of prediction errors in classes that are particularly sensitive from the medical point of view can be achieved using the confusion matrix (see Fig. 5). The best model, ConvNeXt v2 Tiny 384 × 384 pixels, misclassified 9.09% of malignant cases as potentially malignant and 9.09% as benign. In the first case we believe that this is not a big mistake, as OPMD should be biopsied and therefore the biopsy would show that it is an OC, but the case of a misdiagnosis of OC as a benign lesion, it poses a significant error. However, whether these cases have been mistaken for a benign lesion that needs to be biopsied or removed should be further studied to assess the extent of the error. On the other hand, the model classified 5.1% of OPMD cases as benign. These cases should also be further reviewed to assess whether the confusion is with lesions requiring biopsy or excision. Finally, the model classified 5.1% of OPMD cases as healthy. These errors are important and could be due to the fact that few images of certain types of lesions are available. Therefore, it is necessary to evaluate each individual case and further research to assess, in the near future, whether a larger dataset with proportional representativeness of all types of lesions can solve these problems and increase accuracy.

The future implementation of an automatic system based on DL models such as the one proposed in this study could pose several difficulties. Among them is the knowledge and skill to use mobile applications, which could be aggravated in professionals without technological knowledge. In addition, the quality of the cameras is not the same in all cell phones, which could influence detection. On the other hand, although today most of the world's inhabitants have access to the Internet, there may be regions with low coverage. It must also be assumed that there may be a risk of obtaining a wrong diagnosis, so, confirmation by the specialist is necessary. This could be a problem in poor countries with difficulties of referral to specialized centers. Finally, it is important to improve the dataset to cover the entire world population and reduce racial bias, as only images of white Spanish patients were included in our study.

7. Conclusion and future work

In this study, the aim was to select the best DL architecture for the diagnosis of oral lesions in four classes according to the severity of the lesion from clinical images. Notably, it imposed no restrictions on the type of lesion or the area in which it was located. The input to model were small photographic images with a size of 384 × 384 pixels. The advantage of this image size is the possibility of using it in environments with scarce computer resources. According to the performance metrics, the new ConvNeXt v2 Tiny model for 384 × 384 pixel images presents the most satisfactory results, achieving an accuracy of 85.53%, precision of 85.02%, recall of 85.50%, F1-score of 84.92%, and ROC AUC of 97.40%.

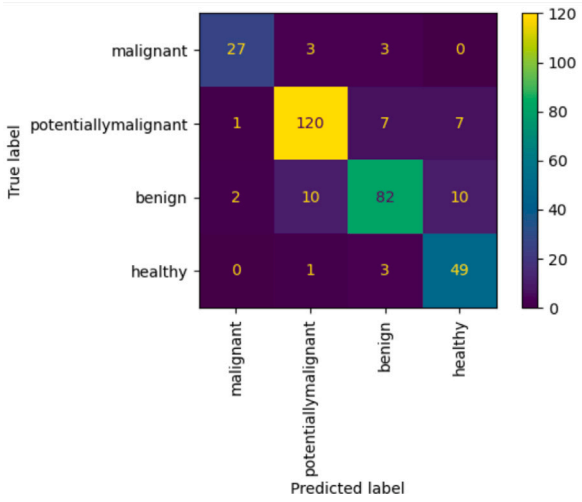

Fig. 5. Confusion Matrix of the ConvNext v2 Tiny model 384 × 384 pixels.

This study may help to develop a tool to assist healthcare professionals in making a first diagnosis of the severity of oral lesions and help in the early diagnosis of OC. However, for clinical application, it would be necessary to further improve the performance of the system by avoiding errors in the classification of OC or OPMD type lesions. Therefore, it is crucial to increase the dataset with more representative samples of all oral lesions and to balance the dataset to increase the accuracy and reliability of the developed model.

Finally, also as future work, inspired by Rabinovici-Cohen et al. [70], who advocate for leveraging image metadata such as lesion appearance and location within the classification layer of CNN models to improve OSCC classification, we suggest a model based on different types of data. This approach would integrate color oral images with structured patient data to enable more personalized predictions. This method would exploit crucial data from the patient's medical history including smoking habits, alcohol consumption, existing diseases, and medication use.

CRediT authorship contribution statement

Alejandro Redondo: Visualization, Methodology, Investigation, Resources, Validation, Writing – original draft, Data curation,

Writing – review & editing, Formal analysis, Software. **Katerina Ivaylova:** Resources, Data curation, Writing – review & editing. **Margarita Bachiller:** Methodology, Writing – review & editing, Writing – original draft, Data curation, Supervision, Visualization, Conceptualization. **Mariano Rincón:** Supervision, Writing – review & editing, Conceptualization, Funding acquisition, Data curation. **José Manuel Cuadra:** Writing – review & editing, Software, Resources. **Faleh Tamimi:** Conceptualization, Resources. **José Luis López-Cedrún:** Resources. **Márcio Diniz-Freitas:** Resources. **Lucía Lago-Méndez:** Resources. **Guillermo Rubín-Roger:** Resources. **Jesús Torres:** Resources. **Leticia Bagán:** Resources. **Gonzalo Hernández:** Resources. **Rosa María López-Pintor:** Resources, Writing – original draft, Conceptualization, Funding acquisition, Writing – review & editing, Data curation.

Funding

This research has been funded by the Instituto de Salud Carlos III (ISCIII) through the project PI22/00905 co-funded by the European Union and by the Spanish Ministry of Science and Innovation through the project PID2019-110686RB-I00.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

[1] J. Ferlay, M. Ervik, F. Lam, M. Laversanne, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, F. Bray, Global Cancer Observatory: Cancer Today, Lyon, France, 2024, URL: https://gco.iarc.who.int/today.

[2] L.Q.M. Chow, Head and neck cancer, New Engl. J. Med. 382 (1) (2020) 60–72, http://dx.doi.org/10.1056/nejmra1715715.

[3] I.-H. Bjerkli, O. Jetlund, G. Karevold, Á. Karlsdóttir, E. Jaatun, L. Uhlin-Hansen, O.G. Rikardsen, E. Hadler-Olsen, S.E. Steigen, Characteristics and prognosis of primary treatment-naive oral cavity squamous cell carcinoma in Norway, a descriptive retrospective study, PLoS One 15 (1) (2020) e0227738, http://dx.doi.org/10.1371/journal.pone.0227738.

[4] A. Sofi-Mahmudi, M. Masinaei, E. Shamsoddin, M.R. Tovani-Palone, M.-H. Heydari, S. Shoaee, E. Ghasemi, S. Azadnajafabad, S. Roshani, N. Rezaei, M.-M. Rashidi, R. Kalantar Mehrjardi, A.A. Hajebi, B. Larijani, F. Farzadfar, Global, regional, and national burden and quality of care index (QCI) of lip and oral cavity cancer: a systematic analysis of the Global Burden of Disease Study 199020132017, BMC Oral. Health 21 (1) (2021) http://dx.doi.org/10.1186/s12903-021-01918-0.

[5] Y. Yang, M. Zhou, X. Zeng, C. Wang, The burden of oral cancer in China, 199020132017: an analysis for the Global Burden of Disease, Injuries, and Risk Factors Study 2017, BMC Oral. Health 21 (1) (2021) http://dx.doi.org/10.1186/s12903-020-01386-y.

[6] A.M. Lima, I.A. Meira, M.S. Soares, P.R. Bonan, C.B. Mélo, C.S. Piagge, Delay in diagnosis of oral cancer: a systematic review, Med. Oral Patol. Oral Cirugia Bucal (2021) e815–e824, http://dx.doi.org/10.4317/medoral.24808.

[7] S. Warnakulasuriya, O. Kujan, J.M. Aguirre-Urizar, J.V. Bagan, M.Á. González-Moles, A.R. Kerr, G. Lodi, F.W. Mello, L. Monteiro, G.R. Ogden, et al., Oral potentially malignant disorders: A consensus report from an international seminar on nomenclature and classification, convened by the WHO Collaborating Centre for Oral Cancer, Oral Dis. 27 (8) (2020) 1862–1880, http://dx.doi.org/10.1111/odi.13704.

[8] T. Walsh, S. Warnakulasuriya, M.W. Lingen, A.R. Kerr, G.R. Ogden, A.-M. Glenny, R. Macey, Clinical assessment for the detection of oral cavity cancer and potentially malignant disorders in apparently healthy adults, Cochrane Libr. 2021 (12) (2021) http://dx.doi.org/10.1002/14651858.cd010173.pub3.

[9] F. Jäwert, J. Nyman, E. Olsson, C. Adok, M. Helmersson, J. Öhman, Regular clinical follow-up of oral potentially malignant disorders results in improved survival for patients who develop oral cancer, Oral Oncol. 121 (105469) (2021) 105469, http://dx.doi.org/10.1016/j.oraloncology.2021.105469.

[10] C. Grafton-Clarke, K.W. Chen, J. Wilcock, Diagnosis and referral delays in primary care for oral squamous cell cancer: a systematic review, Br. J. Gen. Pr.: J. R. Coll. Gen. Pr. 69 (679) (2019) e112–e126, http://dx.doi.org/10.3399/bjgp18x700205.

[11] P. Stathopoulos, W.P. Smith, Analysis of survival rates following primary surgery of 178 consecutive patients with oral cancer in a large district general hospital, J. Maxillofac. Oral Surg. 16 (2) (2017) 158–163, http://dx.doi.org/10.1007/s12663-016-0937-z.

[12] R. Morgan, J. Tsang, N. Harrington, L. Fook, Survey of hospital doctors2019 attitudes and knowledge of oral conditions in older patients, Postgrad. Med. J. 77 (908) (2001) 392–394, http://dx.doi.org/10.1136/pmj.77.908.392.

[13] R.A. Morelatto, C. Moretti, N.J. Bolesina, M.J. Zapata, M.F. Liandro, S. Warnakulasuriya, S.L. de Blanc, Reexamination of delays in diagnosis of oral cancer following an intervention program in Cordoba, Argentina, Oral Surg. Oral Med. Oral Pathol. Oral Radiol. 133 (3) (2022) 301–307, http://dx.doi.org/10.1016/j.oooo.2021.11.006.

[14] P.I. Varela-Centelles, D.P. López, J.L. López-Cedrún, Á. García-Rozado, P.C. Baz, A. Romero-Méndez, J. Seoane, Impact of the presenting symptom on time intervals and diagnostic routes of patients with symptomatic oral cancer, Cancers 13 (20) (2021) 5163, http://dx.doi.org/10.3390/cancers13205163.

[15] S. Kumar, How do primary care doctors compare to dentists in the referral of oral cancer? Evid.-Based Dent. 22 (3) (2021) 94–95, http://dx.doi.org/10.1038/s41432-021-0201-3.

[16] R. Czerninski, N. Mordekovich, J. Basile, Factors important in the correct evaluation of oral high2010risk lesions during the telehealth era, J. Oral Pathol. Med.: Off. Publ. Int. Assoc. Oral Pathol. Am. Acad. Oral Pathol. 51 (8) (2022) 747–754, http://dx.doi.org/10.1111/jop.13343.

[17] C. Bishop, Pattern Recognition and Machine Learning, vol. 2, Springer, 2006, pp. 5–43.

[18] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T.C. Thai, K. Moore, R.S. Mannel, H. Liu, B. Zheng, Y. Qiu, Recent advances and clinical applications of deep learning in medical image analysis, Med. Image Anal. 79 (2022) 102444, http://dx.doi.org/10.1016/j.media.2022.102444.

[19] T. Philipp, C. Noel, A. Bengü Nisa, A. Giusseppe, B. Ralph P, C. Horacio, G. David, H. Allan, H. Brian, H.-W. Rainer, L. Aimilios, L. Jan, L. Caterina, M. Josep, M. Michael A, M. Ashfaq, M. Scott, O. Amanda, P. John, P. Susana, R. Christoph, R. Cliff, S. Alon, S. Christoph, S. H Peter, T. Luc, Z. Iris, K. Harald, Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study, Lancet Oncol. 20 (7) (2019) 938–947, http://dx.doi.org/10.1016/S1470-2045(19)30333-X.

[20] B.R. Pandit, A. Alsadoon, P.W.C. Prasad, S. Al Aloussi, T.A. Rashid, O.H. Alsadoon, O.D. Jerew, Deep learning neural network for lung cancer classification: enhanced optimization function, Multimed. Tools Appl. 82 (2022) 6605–6623, http://dx.doi.org/10.1007/s11042-022-13566-9.

[21] S.M. Qader, B. Hassan, T. Rashid, An improved deep convolutional neural network by using hybrid optimization algorithms to detect and classify brain tumor using augmented MRI images, Multimed. Tools Appl. 81 (2022) 44059–44086, http://dx.doi.org/10.1007/s11042-022-13260-w.

[22] L.d.S. Lucas, P.F. Felipe, D.A. Anna Luiza, A.L. Marcio, A.V. Pablo, K. Syed Ali, K. Luiz Paulo, T.d.S. Harim, W. Saman, D. James, T.P. Alexander, R.S.-S. Alan, Machine learning for detection and classification of oral potentially malignant disorders: A conceptual review, Oral Pathol. Med. 53 (2023) 197–205, http://dx.doi.org/10.1111/jop.13414.

[23] A. Ferrer-Sánchez, J. Bagan, J. Vila-Francés, R. Magdalena-Benedito, L. Bagan-Debon, Prediction of the risk of cancer and the grade of dysplasia in leukoplakia lesions using deep learning, Oral Oncol. 132 (2022) http://dx.doi.org/10.1016/j.oraloncology.2022.105967.

[24] I. Elmakaty, M. Elmarasi, A. Amarah, R. Abdo, M.I. Malki, Accuracy of artificial intelligence-assisted detection of oral squamous cell carcinoma: a systematic review and meta-analysis, Crit. Rev. Oncol. Hematol. (2022) 103777, http://dx.doi.org/10.1016/j.critrevonc.2022.103777.

[25] M. García-Pola, E. Pons-Fuster, C. Suárez-Fernández, J. Seoane-Romero, A. Romero-Méndez, P. López-Jornet, Role of artificial intelligence in the early diagnosis of oral cancer. A scoping review, Cancers 13 (18) (2021) 4600, http://dx.doi.org/10.3390/cancers13184600.

[26] R. Omobolaji Alabi, O. Youssef, M. Pirinen, M. Elmusrati, A.a. Mäkitie, I. Leivo, A. Almangush, Machine learning in oral squamous cell carcinoma: Current status, clinical concerns and prospects for future—A systematic review, Artif. Intell. Med. 115 (2021) http://dx.doi.org/10.1016/j.artmed.2021.102060.

[27] R. Chauhan, K.K. Ghanshala, R. Joshi, Convolutional neural network (CNN) for image detection and recognition, in: 2018 First International Conference on Secure Cyber Computing and Communication, ICSCCC, IEEE, 2018, pp. 278–282, http://dx.doi.org/10.1109/ICSCCC.2018.8703316.

[28] S. Wang, B. Mo, J. Zhao, Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks, Transp. Res. Part B: Methodol. 146 (2021) 333–358, http://dx.doi.org/10.1016/j.trb.2021.03.002.

[29] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708, http://dx.doi.org/10.48550/arXiv.1608.06993.

[30] D. Rothman, Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and More, Packt Publishing Ltd, 2021.
[31] K. O'Shea, R. Nash, An introduction to convolutional neural networks, 2015, http://dx.doi.org/10.48550/arXiv.1511.08458, arXiv preprint arXiv:1511.08458.
[32] A.S. Shamsaldin, P. Fattah, T.A. Rashid, N.K. Al-Salihi, A study of the convolutional neural networks applications, UKH J. Sci. Eng. 3 (2) (2019) 31–40, http://dx.doi.org/10.25079/ukhjse.v3n2y2019.pp31-40.
[33] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015, http://dx.doi.org/10.48550/arXiv.1409.1556, arXiv preprint arXiv:1409.1556.
[34] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114, http://dx.doi.org/10.48550/arXiv.1905.11946.
[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions (2014), 2014, http://dx.doi.org/10.48550/arXiv.1409.4842, arXiv preprint arXiv:1409.4842. 10.
[36] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, http://dx.doi.org/10.48550/arXiv.1704.04861, arXiv preprint arXiv:1704.04861.
[37] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3349–3364, http://dx.doi.org/10.48550/arXiv.1908.07919.
[38] M.Z.M. Shamim, S. Syed, M. Shiblee, M. Usman, S.J. Ali, H.S. Hussein, M. Farrag, Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer, Comput. J. 65 (1) (2022) 91–104, http://dx.doi.org/10.1093/comjnl/bxaa136.
[39] R.A. Welikala, P. Remagnino, J.H. Lim, C.S. Chan, S. Rajendran, T.G. Kallarakkal, R.B. Zain, R.D. Jayasinghe, J. Rimal, A.R. Kerr, et al., Fine-tuning deep learning architectures for early detection of oral cancer, in: Mathematical and Computational Oncology: Second International Symposium, ISMCO 2020, San Diego, CA, USA, October 8–10, 2020, Proceedings 2, Springer, 2020, pp. 25–31, http://dx.doi.org/10.1007/978-3-030-64511-3_3.
[40] G. Tanriver, M. Soluk Tekkesin, O. Ergen, Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders, Cancers 13 (11) (2021) 2766, http://dx.doi.org/10.3390/cancers13112766.
[41] R.F.T. Gomes, J. Schmith, R.M.d. Figueiredo, S.A. Freitas, G.N. Machado, J. Romanini, V.C. Carrard, Use of artificial intelligence in the classification of elementary oral lesions from clinical images, Int. J. Environ. Res. Public Heal. 20 (5) (2023) 3894, http://dx.doi.org/10.3390/ijerph20053894.
[42] P. Birur N, B. Song, S.P. Sunny, P. Mendonca, N. Mukhia, S. Li, S. Patrick, S. AR, T. Imchen, S.T. Leivon, et al., Field validation of deep learning based Point-of-Care device for early detection of oral malignant and potentially malignant disorders, Sci. Rep. 12 (1) (2022) 14283, http://dx.doi.org/10.1038/s41598-022-18249-x.
[43] H. Lin, H. Chen, L. Weng, J. Shao, J. Lin, Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis, J. Biomed. Opt. 26 (8) (2021) 086007, http://dx.doi.org/10.1117/1.JBO.26.8.086007.
[44] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, P. Jantana, Performance of deep convolutional neural network for classification and detection of oral potentially malignant disorders in photographic images, Int. J. Oral Maxillofac. Surg. 51 (5) (2022) 699–704, http://dx.doi.org/10.1016/j.ijom.2021.09.001.
[45] Q. Fu, Y. Chen, Z. Li, Q. Jing, C. Hu, H. Liu, J. Bao, Y. Hong, T. Shi, K. Li, et al., A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study, EClinicalMedicine 27 (2020) http://dx.doi.org/10.1016/j.eclinm.2020.100558.
[46] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, P. Jantana, Automatic classification and detection of oral cancer in photographic images using deep learning algorithms, J. Oral Pathol. Med. 50 (9) (2021) 911–918, http://dx.doi.org/10.1111/jop.13227.
[47] V. Talwar, P. Singh, N. Mukhia, A. Shetty, P. Birur, K.M. Desai, C. Sunkavalli, K.S. Varma, R. Sethuraman, C. Jawahar, et al., AI-assisted screening of oral potentially malignant disorders using smartphone-based photographic images, Cancers 15 (16) (2023) 4120, http://dx.doi.org/10.3390/cancers15164120.
[48] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, P. Jantana, S. Vicharueang, AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer, Plos One 17 (8) (2022) e0273508, http://dx.doi.org/10.1371/journal.pone.0273508.
[49] R.A. Welikala, P. Remagnino, J.H. Lim, C.S. Chan, S. Rajendran, T.G. Kallarakkal, R.B. Zain, R.D. Jayasinghe, J. Rimal, A.R. Kerr, et al., Automated detection and classification of oral lesions using deep learning for early detection of oral cancer, IEEE Access 8 (2020) 132677–132693, http://dx.doi.org/10.1109/ACCESS.2020.3010180.
[50] R. Shaoqing, H. Kaiming, G. Ross, S. Jian, Faster R-CNN: Towards real-time object detection with region proposal networks, Comput. Vis. Pattern Recognit. (2016) http://dx.doi.org/10.48550/arXiv.1506.01497.
[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, http://dx.doi.org/10.48550/arXiv.2010.11929, arXiv preprint arXiv:2010.11929.
[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022, http://dx.doi.org/10.48550/arXiv.2103.14030.
[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255, http://dx.doi.org/10.1109/CVPR.2009.5206848.
[54] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I.S. Kweon, S. Xie, Convnext v2: Co-designing and scaling convnets with masked autoencoders, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16133–16142, http://dx.doi.org/10.48550/arXiv.2301.00808.
[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755, URL: https://doi.org/10.1007/978-3-319-10602-1_48.
[56] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ade20k dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 633–641, http://dx.doi.org/10.1109/CVPR.2017.544.
[57] M. Cantone, C. Marrocco, F. Tortorella, A. Bria, Convolutional networks and transformers for mammography classification: An experimental study, Sensors 23 (3) (2023) 1229, http://dx.doi.org/10.3390/s23031229.
[58] Z. Han, M. Jian, G.-G. Wang, ConvUNeXt: An efficient convolution neural network for medical image segmentation, Knowl.-Based Syst. 253 (2022) 109512, http://dx.doi.org/10.1016/j.knosys.2022.109512.
[59] S. Warnakulasuriya, O. Kujan, J.M. Aguirre-Urizar, J.V. Bagan, M.A. Gonzalez-Moles, A.R. Kerr, G. Lodi, F. Weber Mello, L. Monteiro, G. Ogden, P. Sloan, N.W. Johnson, Oral potentially malignant disorders: A consensus report from an international seminar on nomenclature and classification, convened by the WHO Collaborating Centre for Oral Cancer, Oral Dis. (2020) http://dx.doi.org/10.1111/odi.13704.
[60] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q.V. Le, H. Adam, Searching for MobileNetV3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2019.
[61] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, http://dx.doi.org/10.48550/arXiv.1512.03385.
[62] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986, http://dx.doi.org/10.48550/arXiv.2201.03545.
[63] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), 2016, http://dx.doi.org/10.48550/arXiv.1606.08415, arXiv preprint arXiv:1606.08415.
[64] T. Dozat, Incorporating nesterov momentum into adam, 2016, URL: https://api.semanticscholar.org/CorpusID:70293087.
[65] E. Gordon-Rodriguez, G. Loaiza-Ganem, G. Pleiss, J.P. Cunningham, Uses and abuses of the cross-entropy loss: Case studies in modern deep learning, in: J. Zosa Forde, F. Ruiz, M.F. Pradier, A. Schein (Eds.), Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops, in: Proceedings of Machine Learning Research, vol. 137, PMLR, 2020, pp. 1–10, URL: https://proceedings.mlr.press/v137/gordon-rodriguez20a.html.
[66] A. Al-Kababji, F. Bensaali, S.P. Dakua, Scheduling techniques for liver segmentation: Reducelronplateau vs onecyclelr, in: International Conference on Intelligent Systems and Pattern Recognition, Springer, 2022, pp. 204–212, http://dx.doi.org/10.48550/arXiv.2202.06373.
[67] A.Y. Ng, Feature selection, L1 vs. L2 regularization, and rotational invariance, in: Proceedings of the Twenty-First International Conference on Machine Learning, 2004, p. 78, http://dx.doi.org/10.1145/1015330.1015435.
[68] J. Terven, D.M. Cordova-Esparza, A. Ramirez-Pedraza, E.A. Chavez-Urbiola, Loss functions and metrics in deep learning. A review, 2023, http://dx.doi.org/10.48550/arXiv.2307.02694, arXiv preprint arXiv:2307.02694.
[69] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2009, URL: https://hastie.su.domains/ElemStatLearn/.
[70] S. Rabinovici-Cohen, N. Fridman, M. Weinbaum, E. Melul, E. Hexter, M. Rosen-Zvi, Y. Aizenberg, D. Porat Ben Amy, From pixels to diagnosis: Algorithmic analysis of clinical oral photos for early detection of oral squamous cell carcinoma, Cancers 16 (2024) 1019, http://dx.doi.org/10.3390/cancers16051019.