Smart Health 35 (2025) 100538

Contents lists available at ScienceDirect

Smart Health

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

Explainable screening of oral cancer via deep learning and case-based reasoning

Mario G.C.A. Cimino [a,*], Giuseppina Campisi [b,c], Federico A. Galatolo [a], Paolo Neri [d], Pietro Tozzo [e], Marco Parola [a], Gaetano La Mantia [c,f,g], Olga Di Fede [f,**]

[a] Department of Information Engineering, University of Pisa, l.go L. Lazzarino 1, Pisa, 56122, Italy
[b] Department of Biomedicine, Neuroscience and Advanced Diagnostics (BIND), University of Palermo, Palermo, Italy
[c] Unit of Oral Medicine and Dentistry for Fragile Patients, Department of Rehabilitation, Fragility, and Continuity of Care, University Hospital Palermo, Palermo, Italy
[d] Department of Civil and Industrial Engineering, University of Pisa, l.go L. Lazzarino 2, Pisa, 56122, Italy
[e] Hospital Agency "Villa Sofia-Cervello", via Trabucco 180, Palermo, 90146, Italy
[f] Department of Precision Medicine in Medical, Surgical and Critical Care (Me.Pre.C.C.), University of Palermo, via L. Giuffrè 5, Palermo, 90127, Italy
[g] Department of Biomedical and Dental Sciences and Morphofunctional Imaging, University of Messina, Messina, Italy

ARTICLE INFO

Keywords:
Oral cancer
OSCC
Convolutional neural network
Case-based reasoning
Explainable artificial intelligence
Medical imaging

ABSTRACT

Oral Squamous Cell Carcinoma is characterized by significant mortality and morbidity. Dental professionals can play an important role in its early detection, thanks to the availability of embedded smart cameras for oral photos and remote screening supported by Deep Learning (DL). Despite the promising results of DL for automated detection and classification of oral lesions, its effectiveness is based on a clearly defined protocol, on the explainability of results, and on periodic cases collection. This paper proposes a novel method, combining DL and Case-Based Reasoning (CBR), to allow the post-hoc explanation of the system answer. The method uses explainability tools organized in a protocol defined in the Business Process Model and Notation (BPMN) to allow its experimental validation. A redesign of the Faster-R-CNN Feature Pyramid Networks (FPN) + DL architecture is also proposed for lesions detection and classification, fine-tuned on 160 cases belonging to three classes of oral ulcers. The DL system achieves state-of-the-art performance, i.e., 83% detection and 92% classification rate (98% for neoplastic vs. non-neoplastic binary classification). A preliminary experimentation of the protocol involved both resident and specialized doctors over selected difficult cases. The system and cases have been publicly released to foster collaboration between research centers.

1. Introduction and background

Oral cancer is the 13th most common cancer globally, accounting for 377,713 new cases and 177,757 deaths in 2020. According to the Global Cancer Observatory (GCO), the incidence of Oral squamous cell carcinoma (OSCC) will rise by approximately 40% by 2040, accompanied by a growth in mortality (Tan et al., 2023).

* Corresponding author.
** Corresponding author.
E-mail addresses: mario.cimino@unipi.it (M.G.C.A. Cimino), olga.difede@unipa.it (O. Di Fede).

https://doi.org/10.1016/j.smhl.2024.100538
Received 6 January 2024; Received in revised form 23 December 2024; Accepted 31 December 2024
Available online 1 January 2025

Its multifactorial etiology includes both intrinsic factors (e.g. iron-deficiency anemia or malnutrition) and extrinsic factors (e.g. alcoholic substances or products of tobacco). Furthermore, the patient's economic and social condition has a non-negligible impact in oral cancer risk; indeed, the cases of disease diagnosed in advanced stages have a high probability of referring to the underprivileged groups.

The most common sites for the OSCC are tongue (25.4%), labial/buccal mucosa (21.7%), gingiva (14.0%), palate (9.9%), and alveolar mucosa (7.9%), respectively (Dhanuthai et al., 2018) (Fitzpatrick et al., 2019). Ulcerated pattern is the main clinical form of OSCC that is notorious for its ability to mimic benign ulcerative lesions.

The prevalence of OSCC is not high compared to other cancer entities, but it poses significant mortality and morbidity issues in patients, both because of its late diagnosis/misdiagnosis and of the impact on life quality of advanced oral cancer therapies, including extensive surgery and radiation therapy (American Cancer, 2018).

Upon cancer diagnosis, the stage of disease progression is a crucial prognostic parameter for life expectancy: five-year survival rates for more advanced stages ranging from 20% to 50%, and increasing to about 80% for stages I and II. Five-year mortality rates reach nearly 40%, and oral cancer accounts for 3.6% of all cancer deaths (García et al., 2020). Obviously, biopsy is the spearhead for OSCC diagnosis of oral cancer, but it is not applicable as a mass screening method, which, instead could benefit from conventional oral examination by an expert, especially in high-risk groups. Unfortunately, most high-risk subjects are currently unable to access specialized physicians or dentists in developing due to inadequate health services and insufficiently well-trained medical personnel (Coelho, 2012).

Although the oral cavity is very easily accessed for examination, it often receives minimal attention in routine practice. Moreover, the identification of OSCC or oral potentially malignant disorders (OPMD) often require experienced examiners and specialized medical staff (World Health Organization (WHO)). Dentists are also responsible for the basic prevention of oral cancer by providing information on the main risk factors and advice on avoiding smoking, reducing alcohol consumption and protecting against sunlight. For secondary prevention of OSCC, general practitioners and dentists play a vital role, but the lack of routine inspection of the oral mucosa during dental visits is noticed by many patients whose oral cancer is only identified at an advanced phase. One of the most important variables associated with a long time to treatment initiation is the advanced-staged cancer. The literature shows significantly different values of actual physician and treatment time delay, but excessive duration emerges from all. When the disease is found to be in the early stages, mortality rates are approximately 16 %, while for more advanced stages this value rises to 61% (Gigliotti et al., 2019). In addition to liability attributable to the physician ("provider delay"), we can identify another form of responsibility entirely attributable to the patient: patient delay. Patient delay can be characterized as the period between the first symptoms occurring and the first follow-up visit with a health care professional (Nagao & Warnakulasuriya, 2020).

It is worth to mention that patients with higher risk of developing OSCC, which are the socioeconomically disadvantaged people, are those who least likely go to visit a dentist for intra-oral examination and who are more likely to be visited by less specialized health care providers, such as general practitioners (Conway et al., 2002, pp. 119–123).

Since early cancerous lesions often resemble common oral diseases, early detection of malignant lesions can also improve patient survival. It would be desirable that an accurate mouth inspection, performed by trained and experienced doctors with great expertise for OSCC detection, could be part of cancer screening, but it is not always feasible, especially in developing countries or remote areas. Indeed, over the past two decades, the ratio of dentists-to-population in urban and rural regions still shows a large gap, although there is a decreasing trend. Remote regional locations receive less dental care services than more accessible ones. This situation causes excluded segments of the population to rely on non-dental care centers as doctors, pharmacies or emergency facilities in hospitals (Cohen et al., 2011) (Okunseri et al., 2011). Considering this, and due to improvements in information and communication technology's (ICT) declining costs, e-learning for professional education, telemedicine services and AI-based medical platform usage is increasing (Sun et al., 2023).

1.1. AI in healthcare system

Machine learning systems are progressively being implemented in modern oncology practice for categorization and prediction. In particular, the adoption of Deep Learning (DL) architectures, especially Convolutional Neural Networks (CNN) achieved promising results in computer vision challenges, such as image classification, object detection, and instance/semantic image segmentation (Adeoye et al., 2021) (Fabbrizzi et al., 2022).

The major issues related to DL are the following: (i) the difficulty of collecting large amounts of data in collaboration with hospital centers; (ii) the understanding of the model output, as expected by eXplainable Artificial Intelligence (XAI). Indeed, the number and quality of healthcare data collected by a single organization is usually limited with respect to what is required by DL. Non-representative training sets can bias a DL architecture, causing inaccurate classification. Moreover, various factors can cause a DL system to become less effective over time, necessitating regular updates and retraining to maintain its performance. First, *technological evolvement*: as new devices come on the market, the data and methods that a DL system was originally trained on may become outdated. This can lead to a mismatch between the system's capabilities and the current technological setting. Second, *change of normative or clinical practices*: in fields such as healthcare, practices and standards evolve; if a DL system was trained on slightly different practices, it might not perform well under new standards or guidelines. Third, *socio-economic or cultural trends*: changes in society, economy, or culture may influence people's behavior; for instance, a system trained on oral lesions caused by smoking one type of tobacco may struggle to adapt when a new tobacco is introduced for cultural reasons. Other contextual factors can easily affect the performance of a DL system in the medium term. Consequently, the statistical properties of input-output relationships change, producing an effect called concept drift (Priya & Uthra, 2023). In general, the effectiveness of DL technology is based on a collaborative, systematic and massive

data ingestion, managed by a global health data consortium (Chen et al., 2019).

Another important barrier to the achievement of DL-based solutions is related to their explainability, i.e., the understanding of how a result has been generated. For medical and legal reasons, any diagnosis system should provide explanations to support human decision. Such explanations can also highlight unexpected training bias or drift of the model. Moreover, the recent European General Data Protection Regulation (GDPR) includes retractability of decisions as a requirement of all businesses. As a such, humans should keep track of motivations and use their judgement in decisions (Holzinger et al., 2017).

1.2. Goal and purpose

This paper introduces a novel method for the explainable classification of oral lesions, by combining Deep Learning (DL) with Case-Based Reasoning (CBR). CBR is a problem-solving method commonly used by doctors to classify a new case by referring to similar past cases (Lamy et al., 2019a). With regard to the proposed diagnostic system, a case refers to an oral lesion, with the class representing the name of a disease or condition that explains the lesion. In a therapeutic system the classes correspond to the available treatment options. CBR is traditionally combined with symbolic AI to execute a set of logical rules, which is effective for standard cases by formalizing evidence-based clinical practice guidelines. CBR is essential for handling cases that fall outside these guidelines due to contraindications or other patient-specific variables. These non-standard cases can account for up to 45% of all cases (Bouaud et al., 2009). For this purpose, in the proposed method, the DL architecture is designed to provide, after training, a measure of similarity between cases, thus facilitating the explainability of the classification results. Such similarity is used as a basis for CBR. Consequently, the CBR process carried out by the decision maker explains the CNN behavior (Singh et al., 2020a). Specifically, for a given problem, the most similar cases are used as samples for justifying the response of the CNN. The similarity is also used to visualize, via dimension reduction, a scatter plot of the training set in which the new case is located. In the scatter plot, several similar cases can be determined for each class in order to identify the class with the most similar samples. As a result, the physician can retrace the decision and use their judgement, which is a key for a safe and trustworthy process based on DL (Lamy et al., 2019a).

The paper illustrates the motivation for this approach, the design of the decision process, and the development of the DL-CBR decision support system, which is able to perform the task of lesion detection and classification for an effective diagnosis. A pilot system has been implemented and publicly released, in terms of both source code and online service. The DL engine is based on Feature Pyramid Network Resnet-50 Faster R-CNN for object detection and classification, pre-trained with the Microsoft Common Objects in Context (COCO) data. The CBR is based on 221 cases belonging to the main three classes of oral ulcers: neoplastic (N), aphthous (A) and traumatic (T). The DL achieves the state-of-the-art performance, i.e., 81.6% detection and 90% classification rate (98% for neoplastic vs. no neoplastic binary classification). The DL-CBR based decision process, carried out by humans with a small set of decision rules, is able to address even resident doctors to correctly carry out the correct diagnosis. The system is intended to support doctors to correctly address the patient to specialized centers or to second-level inspection by experienced staff, implementing the early detection and treatment of OSCC.

An important aspect of the presented approach is its complete reproducibility: the data set of lesions, the source code, the complete hyper-parametrization method, as well as the online service are publicly provided (Galatolo, 2024). We aim to foster: (i) the collaboration between clinical centers in collecting more data on regional/national basis, (ii) the collaboration between deep learning development centers to achieve the best networks in this field.

The paper is structured as follows. Section 2 presents related work. Section 3 focuses on the design and evaluation methodologies, while in Section 4 the case study and the experimental results are covered. The final discussion is described in Section 5. Section 6 draws conclusions and future work.

2. Related work

Explainable diagnosis is a thriving research area in the healthcare domain. To address this task, several studies proposed AI-based solutions. This section summarizes the most relevant contributions on this topic, considering the two pillars of the proposed approach: (i) deep learning and (ii) case-based reasoning.

2.1. Deep learning for medical imaging

As mentioned in the previous section, DL is increasingly being adopted in the healthcare context. The most popular DL architectures employed for medical imaging are CNNs (Krizhevsky et al., 2017). The scientific community has effectively exploited CNNs for many case studies in several medical branches for clinical image analysis, including cancer detection and diagnosis into both malignant/benign and mass/calcification (Arevalo et al., 2016) (Suh et al., 2020), skin lesion investigation (Nasr-et al., 2016), cardiovascular context for blood cancer and heart anomalies (Kiranyaz et al., 2015), and, of course, numerous pulmonary studies on the effects of coronavirus (COVID-19) (Bhattacharya et al., 2021) (Soomro et al., 2022).

Regarding the implementation, a CNN is based on multiple layers employing a mathematical operation called convolution, applied in different ways to the original image to generate feature maps. Each feature map is easier to classify: it represents a separated concept of an image, like texture and shape, which can be measured by a real-valued function. By extracting many features, the initial image can be transformed from a real-valued matrix (pixels) into a numerical feature vector. The transformation of image pixels into a numerical feature vector is called image embedding. The final segment of a CNN is called classification layer, which transforms the feature vector into a numerical coding of a class. For example, “1” is the numerical code associated to a picture with a “malignant”

lesion, and "0" is associated to "benign" lesion.

Compared to conventional image processing algorithms, CNNs have the first benefit of not requiring manually engineered filters based on past knowledge and human labor. This means that a training algorithm (solving an optimization problem) can be run to iteratively change the NN parameters to decrease the output error value, by providing a set of images with the related numerical coding of classes, i.e., to predict a continuous value very close to the coding. Additionally, this coding may be general, allowing the network to accurately classify new images.

The second important advantage of CNNs is that the major private and public research centers, having large image databases of various objects, publicly provide pre-trained CNNs. For example, the VGG-19 network is a 19-layer CNN trained using more than 1 million of 224 $\times$ 224 pixels colored images from the ImageNet database, classifying up to 1000 objects with 71% of accuracy (Simonyan & Zisserman, 2015). Similarly, the GoogleNet is a 50 layers CNN with 78% of accuracy, fed with 299 $\times$ 299 pixels images of ImageNet (Szegedy et al., 2015). ResNet50 is a 50-layers CNN trained by Microsoft, with 75% of accuracy (He et al., 2016, pp. 770–778), followed by ResNet101 and resNet152 models. More recently, EfficientNet is a CNN released in 2019 by Google with 77% of accuracy (Mingxing et al., 1905).

The importance of pre-trained networks is that they provide an efficient and robust image embedding. Their classification layer can be replaced with a new classification layer that can be efficiently trained (fine-tuning) for the specific classes to learn. The reuse of the embedding layers is called transfer learning. A very important aspect of fine-tuning from pre-trained models is to select the best hyperparameters of the classification layer. The learning rate, which measures how fast the model learns, is an example of a hyper-parameter. A very low value can result in a very drawn-out training process that may become stuck, whereas a very high value can result in an unstable training process. The most sensitive hyperparameters are tuned by introducing an optimization approach based on a Tree Parzen Estimator (TPE) for identifying the optimal option in order to ensure the reproducibility of the training (Bergstra et al., 2011).

The hyperparameters optimization introduces a new order of magnitude in computational cost, in relation to the large number of hyperparameters of DL models, and to their dynamic lifecycle. Consequently, despite the improvement of computational capabilities and costs, a training infrastructure for DL is not convenient nor effective to maintain for a single healthcare organization, with respect to a knowledge-based decision support system.

In (Jeyaraj et al., 2019a) the oral cancer diagnosis is performed by solving a classification problem. Specifically, a DL regression based on partitioned CNNs is experimented to process multidimensional medical images and automatically perform diagnosis. The methodology has been validated on hyperspectral images from standard datasets from the BioGPS UCI repository, in which healthy tissue is distinguished from diseased tissue, characterized by a malignant/benign tumor. The authors compared the performance of the convolutional architecture with other traditional classification models i.e., Support Vector Machine (SVM) model and Deep Belief Network (DBN). The accuracy achieved by this regression-based partitioned approach was 94.5% compared with SVM and DBN, which achieved 82.4% and 84.5%, respectively.

Qiuyun Fu et al. (Fu et al., 2020) designed a deep architecture exploiting convolutional networks to deal with photographic images. A detector model received an oral photo as input and produced a bounding box where it placed the potential lesion. Then, the candidate area is sent to a neural classifier based on a model pre-trained on the ImageNet dataset and further fine-tuned on the available dataset. The methodology was developed on an internal dataset consisting of 6176 images, randomly split by hold-out method into training set (5775) and test set (401). The related reports served as ground truth for the methodology's development and verification. The authors also performed a verification phase with external data comprising 420 clinical photographs from six sources in the dental area.

Recently, Tanriver et al. (Dinesh et al., 2023) suggested a two-stage approach to identify oral lesions using a detector model. The proposed approach categorizes the detected region into three distinct categories (benign, OPMD, cancer) with a second-stage classifier. The authors of this work showed that DL-based methods may be used to automatically detect and classify oral lesions in real time, opening up the possibility of a highly effective, inexpensive, and non-invasive tool for the early identification of OPMD.

In (Srivastava et al., 2021) Srivastava et al. made a comparative analysis of deep learning image detection algorithms, focusing on Faster R-CNN, SSD (Single Shot MultiBox Detector) and YOLO v3 (You Only Look Once). Faster R-CNN is considered superior for photographic detection with a relatively small dataset and without real-time requirements. Nevertheless, it is worth noting that use case and dataset size can sensibly influence algorithmic performance. Indeed, oral cancer images are characterized by shapes and patterns that are very different with respect to the object detection benchmarks available in the literature, such as ImageNet and COCO. Specifically, Faster R-CNN is known for its high accuracy and precision in object detection tasks. It uses a region proposal network (RPN) to generate high-quality region proposals, which significantly improves detection performance. It is faster than its predecessors like R-CNN and Fast R-CNN: the RPN allows for end-to-end training and faster inference times. It has shown robustness in handling various scales and aspect ratios of objects, which is crucial for detecting diverse manifestations of oral cancer in photographic images. Overall, one-stage detectors, such as YOLO, tend to be faster but generally offer lower accuracy compared to two-stage detectors like Faster R-CNN, which are usually slower but more accurate.

2.2. Case-based reasoning in the medical context

Case-based Reasoning is a general AI paradigm for problem solving, with applications to a large variety of domains. CBR is based on reasoning from cases, exploiting a four-step cycle: retrieval, reuse, revision, and retraining (Sørmo et al., 2005) (Johs et al., 2018). For a new problem (i.e., case) to solve, the most similar problems (cases) are first retrieved from a history of solved problems (cases). The retrieval phase finds cases by comparing the features of the new case with the history of cases, using a search algorithm.

In (Keane & Kenny, 2019), Keane and Kenny proposed an integration between CBR and Artificial Neural network (ANN) as a "twin" system. Specifically, an ANN-CBR Twin-System is defined by the following main aspects: (i) *hybrid system*, i.e., made by the ANN and CBR methods, combined to comply with accuracy and interpretability requirements of the overall system; (ii) *modules separation*, i.e., The above methods are executed as separate, independent but side-by-side modules; (iii) *shared datasets*, i.e., the two techniques are based on the same dataset; (iv) *bipartite split of work*, i.e., the ANN module makes predictions, while the CBR provides interpretations, explaining the ANN results. In the context of CBR, an important aspect is the similarity relationships between cases. An adopted solution in this regard is the Weighted Euclidean Distance (WED) (Ahn et al., 2020) (Ragnemalm, 1993) (Kwon et al., 2019), formally defined in Equation (1):

$$WED(x,y) = \sqrt{\sum_{i=1}^{n} w_i \,(x_i - y_i)^2} \tag{1}$$

In the health care domain, Chun-Ling Chuang (Chuang, 2011) designed and implemented CBR systems for the diagnosis of liver disease, able to solve a binary classification problem to distinguish diseased patients from healthy individuals. The author proposed four CBR solutions based on different strategies: (i) a three-layer feed-forward NN architecture; (ii) a tree model, built with a classification and regression tree algorithm; (iii) a logistic regression model; (iv) a discriminatory analysis approach. The CBR systems were trained and tested on a dataset of 166 observations: 91 belonging to the diseased class and 75 to the healthy one. The NN architecture achieved 95% accuracy, followed by the tree model with 91%, while the other two approaches performed below 90%.

Regarding the breast cancer diagnosis problem, Dongxiao Gu et al. (Gu et al., 2020) proposed a decision support system that combines the case-based paradigm with ensemble learning, to solve a binary classification problem. They designed and implemented the CBR paradigm based on Extreme Gradient Boost (XGBoost), a state-of-the-art ensemble learning method, after comparing the performance with other popular classification methods: logistic regressor, SVM, kNN, decision tree, random forest, and a feed forward NN architecture. XGBoost was selected as the best classifier, achieving 91.62% accuracy, followed by NN with 90.4%, and logistic regression with 90.13%; the other models achieved values below 90%.

In the literature of oral cancer, the case-based approach has been successfully experimented for both learning (Shrestha et al., 2021) and reasoning (Ehtesham et al., 2019). Specifically in Shrestha et al. (Shrestha et al., 2021) about 60 third-year undergraduate students were involved in a Case-Based Learning (CBL) module on OSCC, divided into CBL and conventional learning groups. The cases considered included clinical and habit history, lesion descriptions, radiographs, laboratory investigations, histopathological features and diagnosis. The CBL group achieved significant improvements in performance. Similar findings were observed in 40 fourth-year dental students in oral medicine (Du et al., 2013), and in about 70 first-year dental students in physiology (Omprakash et al., 2018). Significant research is presented in (Ehtesham et al., 2019), where a CBR system was extensively experimented for the diagnosis of different oral diseases: ulcerative, vesicular, and bullous lesions of the oral mucosa; red and white lesions of the oral mucosa; pigmented lesions of the oral mucosa; benign lesions of the oral cavity; oral cancer; salivary gland diseases. A Delphi study was used to determine the parametric values for calculating the similarity rates. Cases were compared with each other based on symptoms. For a new case, the system was able to suggest a diagnosis for the dentist, based on its similarity with the available cases, filtered by a similarity threshold. A database of about 500 cases was used. The system evaluation, involving specialists and about 40 patients, indicated the potential to improve the quality and efficiency of oral disease diagnosis (Ehtesham et al., 2019).

2.3. Case-based reasoning fundamentals

Let us formally define a specific task of CBR, namely Case-Based Classification (CBC). Given a predefined set P of problems, and a set S of predefined solutions, then a *case* c_i is an ordered pair:

$$c_i \equiv (p, s) \tag{2}$$

where $p \in P, s \in S$. In medicine, p and s correspond to *symptoms* and *diagnosis*, respectively. A *classifier* $C(\cdot)$ models the mapping from p to s, where p is the input medical image, and s is a corresponding label:

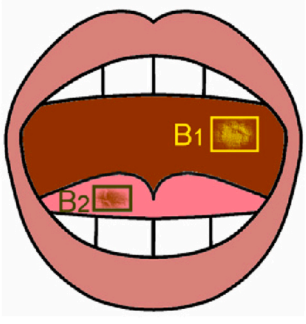

Fig. 1. – Object detection and classification: schematic representation of an oral cavity with two bounding boxes (B1 and B2).

$$s \equiv C(p) \tag{3}$$

The image is classified if one or more target objects are detected. The region of each instance is identified by one or more (partially overlapping) bounding boxes B_i. Fig. 1 shows an example of schematic drawing of an oral cavity, with two lesions and corresponding boxes B_1 and B_2. For a given input image, a *detector* $D(\cdot)$ finds out lesions providing a set of bounding boxes:

$$\{B_i\} = D(p) \tag{4}$$

A similarity function between cases is then defined as follows:

$$\begin{aligned} 0 \leq sim(c_i, c_j) \leq 1 \\ sim(c_i, c_i) = 1 \end{aligned} \tag{5}$$

In CBC, the solution of a new problem p starts with the retrieval of the most similar cases. The hypothesis is that similar problems have similar solutions. Given a new problem, when considering a solution associated to a similar problem, it can either be accepted in the given form or modified. This process is called *case adaptation*. An external validation, called *case revision*, is finally carried out to ensure that any discrepancies or failures in the adapted solution are corrected, improving the case for future use (Szczepaniak & Duraj, 2018). As an example, let us consider a new patient for which the CBR system finds, as the most similar past case, a treatment protocol with a drug to which the new patient is allergic. During the case adaptation, it is included an alternative medication that is effective for the disease but is safe for the patient given their allergy history. During the case revision, the patient initially improves the modified treatment plan but then develops new symptoms not seen in the original case. The case is then revisited to assess why the patient's

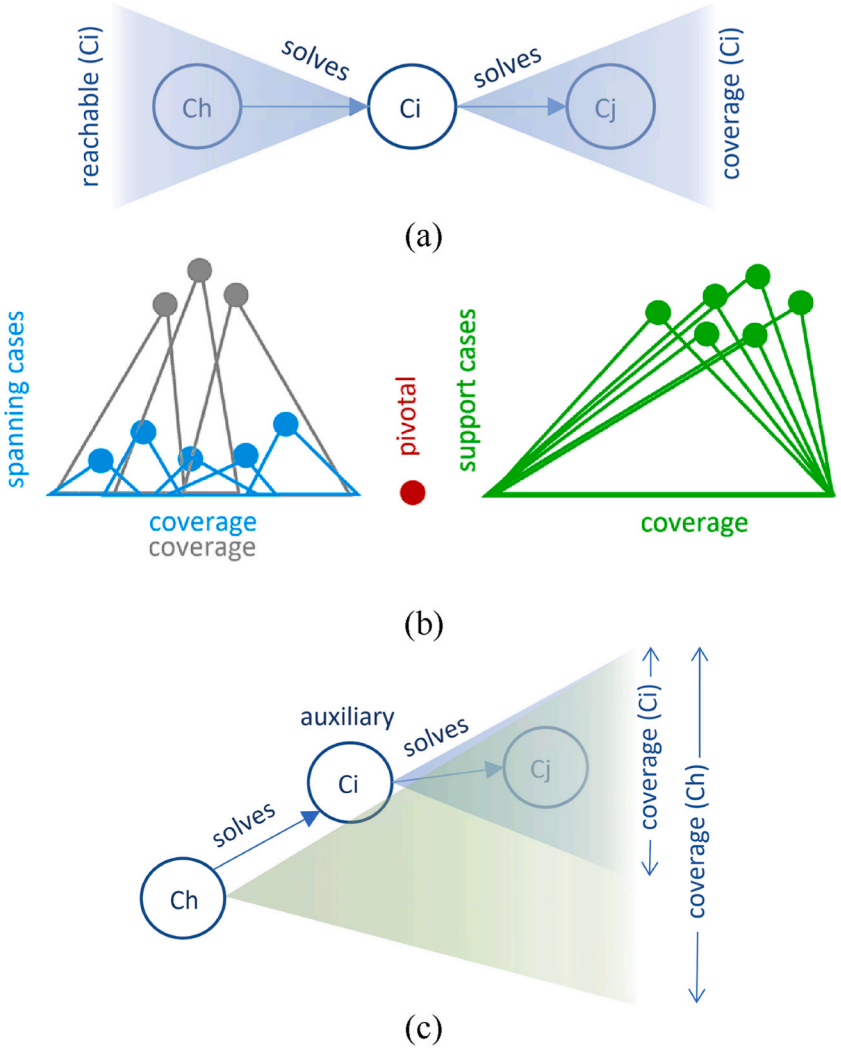

Fig. 2. – Representation of case-based reasoning concepts: (a) reachability and coverage; (b) pivotal, supporting and spanning cases; (c) auxiliary case.

condition worsened. It identifies that the substitute drug, although safer, is less effective. The case is revised by combining the substitute drug with another medication to enhance its effectiveness while avoiding the allergy risk. This revised case is then stored in the case base for future reference.

Given a case base $C = \{c_i\}$, the *coverage* of a known case c_i is the set of *new solvable cases* c_j via c_i, i.e., new cases c_j whose solution is found by adapting c_i's solution to c_j:

$$coverage(c_i) = \{c_j \in C : adaptable(c_i, c_j)\} \tag{6}$$

whereas the *reachability* of a target problem c_i is the set of cases c_h whose solution can be used to solve c_i, i.e., by adapting c_h's solution to c_i:

$$reachable(c_i) = \{c_h \in C : adaptable(c_h, c_i)\} \tag{7}$$

Fig. 2a illustrates the concepts of reachability and coverage. Here, cases are represented by circles; an arrow between two cases represents the relationship of adapting one case to solve another case; a colored triangle represents a beam of arrows.

To determine the importance of cases in terms of competence contributions, there are four different categories. In Fig. 2b, colored circles represent cases, each category with a different color. Each case is positioned at the vertex of a triangle. The coverage of the case is represented by the base of the triangle.

A pivotal case is reachable by no other case except itself. It is the isolated red circle in the middle of Fig. 2b. As such, its deletion directly reduces the competence of the system. Support cases are a group of cases each providing coverage similar to the others. The deletion of one or more support cases does not reduce competence, but the deletion of a pivot is like removing the group as a whole. On the right of Fig. 2b, support cases are represented as green circles with the same coverage. Spanning cases are cases covering regions of the problem space that are independently covered by other cases. As a such, spanning cases can be entirely deleted without affecting the competence of the system. On the left of Fig. 2b, spanning cases are represented as blue circles covering the same area of another group of grey circles. Finally, a case is an auxiliary case if the coverage it provides is included by the coverage of one of its reachable cases. Fig. 2c illustrates the concept. Auxiliary cases form groups such that the deletion of an auxiliary case does not reduce the competence of the system, because another reachable case can be considered to solve any target that the deleted auxiliary could solve.

Let us define an *outlier* as a *pivotal case* (Shrestha et al., 2021), i.e. cases that are too isolated to be solved by any other case:

$$c_{out} = pivot(c), iff \;\; reachable(c) - \{c\} = \varnothing \tag{8}$$

Considering similarity, the outlier is defined as a case for which no more than k cases in the case base have similarity l or more:

$$outlier(c_i) = true \; if \; |\{c : sim(c, c_i) \geq l\}| \leq k \tag{9}$$

As a result, with an outlier the CBR system cannot generate an effective response. In this scenario, an alternative solution must be found, and the new case must then be added to the case base. In contrast, for a normal case with CBC, the target class of a new object is assigned by taking the k most similar cases already labelled, and voting for their majority class.

3. Design and evaluation methodologies

This section is organized as follows: in Section 3.1, the problem and the design of the DL architecture are detailed; Section 3.2 covers the design of the DL-CBR system; finally, in Section 3.3, an overall DL-CBR workflow is proposed.

3.1. Problem definition and design of the deep learning architecture

To solve an object detection task, all occurrences of objects belonging to a predefined set should be detected, delimiting as best as possible the portion of the image in which they are located. For instance, scattered lesions have a high number of locations with lesions (Amit et al., 2020, pp. 1–9). Let us define the *detection rate* as the number of correctly detected lesions divided by the number of total lesions. To determine a correctly detected lesion, a bounding box proposed by the network, BB_p, is compared with the actual box, BB_a, drawn by a specialist doctor, via the Intersection over Union (IoU) metric (Padilla et al., 2020). IOU measures the similarity between the two boxes, as (size of) their overlapping area divided by (size of) their union area, as shown by Equation (10):

$$IoU(BB_a, BB_p) = \frac{size\ of\ area\ (BB_a \cap BB_p)}{size\ of\ area\ (BB_a \cup BB_p)} \tag{10}$$

A threshold IoU_{TH} is finally used to determine positive detections: if $IoU(BB_a, BB_p) > IoU_{TH}$, then the lesion is correctly detected. In all experiments, IoU_{TH} was set to the standard value of 0.5.

To evaluate the classification task, let us define the *classification rate* as the number of correctly detected and classified lesions divided by the number of correctly detected lesions, as shown in Equation (11):

$$classification\ rate = \frac{no.\ of\ correctly\ detected\ and\ classified\ lesions}{no.\ of\ correctly\ detected\ lesions} \tag{11}$$

As an architectural model for the detector $D(\cdot)$, the classifier $C(\cdot)$, as well as the similarity $sim(\cdot,\cdot)$, a pre-trained convolutional DL

architecture is leveraged. The architecture has been extended to manage similarity, and it is based on a CNN architecture called Detectron 2, an object detection system developed by Facebook AI Research (Wu et al., 2019). Fig. 3 shows the overall blocks of the extended architecture.

In particular

(a) *Backbone Network* – Base CNN model for feature extraction. Its role is to extract feature maps from the input image, characterized by H (height) × W (width), and BGR colors (Blue, Green and Red). Its structure is made by a basic 'stem' block, a maxpool, and 4 stages (res2, …, res5) to downsample convolution layers. Each stage is made by a lateral convolution layer, an output convolution layer, a forward process, and a last level max pool. The detector is based on a Faster R-CNN with Feature Pyramid Network (FPN). It detects different scale objects (from P2 to P6). As a backbone network, a pretrained Mask R-CNN model using ResNet-50 with FPN has been used. The network was pretrained using the COCO 2017 dataset.

(b) *Region Proposal Network (RPN)* – Identification of regions in images that have objects, called proposals. It detects object regions from the multiscale features, providing 1000 box proposals. It consists of a NN (RPN Head) and non-neural-network functionalities. The RPN Head consists of three convolution layers. RPN has an *object/non-object* binary classifier to propose regions with objects. In the following, the non-object (complement) class will be called *healthy* tissue.

(c) *Similarity Features (SIMF)* – It is based on the ROI Pooler, it is compressed via an AVG Pooling 7x7, and then flattened; it provides an array of 256 features, which is adopted as feature vector for calculating similarity via cosine distance.

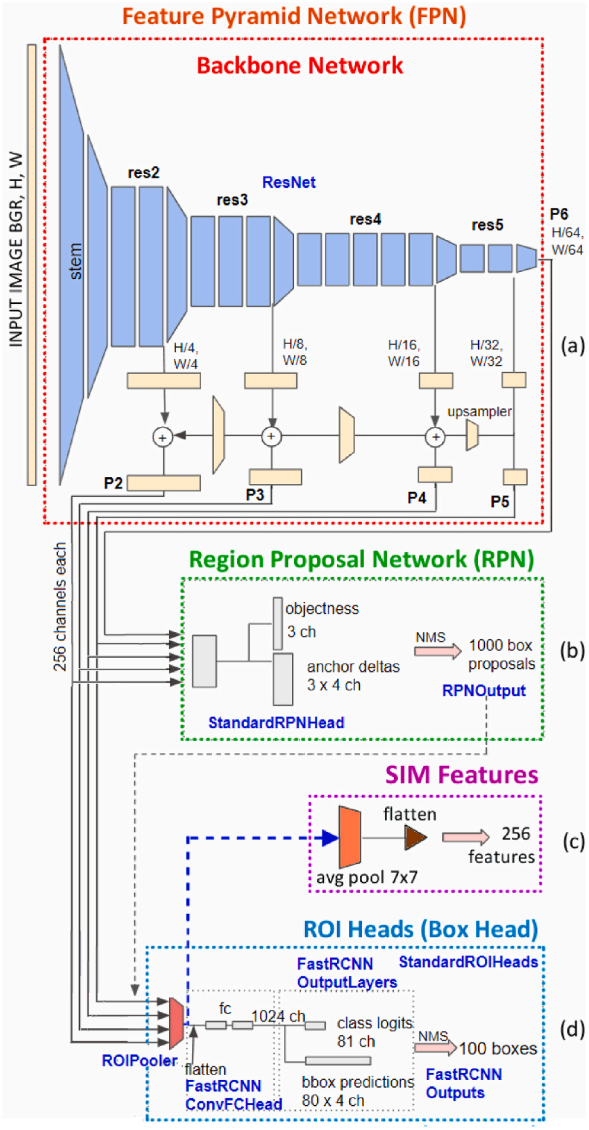

Fig. 3. – The proposed extension of the Faster-R-CNN FPN + architecture (adapted from (Omprakash et al., 2018)).

(d) *ROI (Box) Head* – Prediction of final bounding boxes and classes based on multi-task loss. Mask R-CNN also predicts masks via an additional head using ROI Align output. It takes the feature maps from FPN, and the proposal boxes, and releases box location and classification. The classification is provided as a membership degree of the input image for each class. Each membership degree is between 0 and 1, and the sum of the membership degrees of all classes to which the input image belongs is 1. Finally, the predicted class is the one corresponding to the highest membership degree.

3.2. *Design of the case based reasoning system*

The CBR system is based on a data set of cases selected by domain experts and mapped into an n-dimensional space by the DL system. Specifically, for a new image submitted to the system, the DL system extracts a set of features. The CBR system performs a retrieval to get similar cases (Kolodner, 2014). This retrieval phase is made via the k-Nearest Neighbors (kNN) search, using the cosine distance similarity for calculating the nearest neighbors (Juarez et al., 2018) (Sun & Huang, 2010) (Ray, 2019).

For reasons of readability, explainability, consistency and cost, a CBR system tends to achieve maximum competence with a small case base that is also complete in terms of coverage. Indeed, a CBR system is most effective when it has a small, but comprehensive, collection of cases, which guarantees: (a) *readability* – a smaller case base is easier to manage, navigate, and understand. It makes it

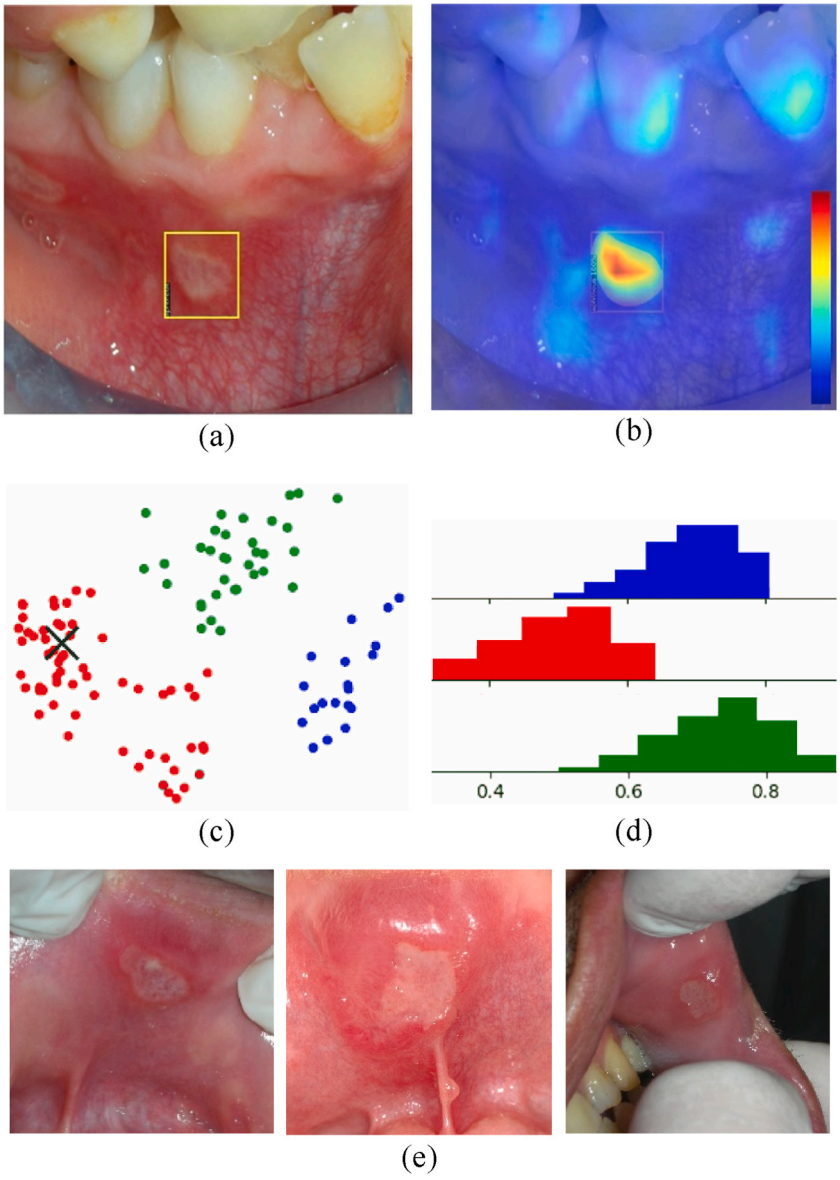

Fig. 4. – Visual explanation tools: (a) bounding box; (b) saliency map generated via Grad-Cam++ (Juarez et al., 2018), (c) scatter plot, (d) similarity histogram (e) case-based explanation, $k = 3$.

simpler to find relevant cases quickly without being overwhelmed by too many options; (b) *explainability* – when the case base is small and well-curated, it is easier to explain how and why the system reached a particular decision. This transparency is crucial for users to trust the system's recommendations; (c) *consistency* – a smaller, complete case base ensures that similar problems are solved in similar ways, which helps maintain consistent decision-making across the system; (d) *cost* – managing and updating a large case base can be expensive and time-consuming. By keeping the case base small and focused, the system reduces these costs while still performing effectively; (e) *completeness in terms of coverage* – the case base should be comprehensive enough to cover all possible scenarios the system might encounter. This ensures that, even with a smaller number of cases, the system can still find a relevant case for a new problem.

On the other side, a DL system tends to achieve maximum competence with a large training set of cases that is also sufficiently various. In this research, an integrated workflow for collecting cases is developed, for achieving coherence in both systems. The goal is to provide cases for the CBR system that can also explain and validate the DL behavior (Lamy et al., 2019b). An essential aspect of this integration is the similarity metric. In this work the cosine similarity is used. Formally, let $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$ be the n-dimensional vector representation of two feature vectors generated by the DL system for two corresponding images. The cosine similarity of $\boldsymbol{X}$ and $\boldsymbol{Y}$ represents the angle α between them:

$$sim(x, y) = \cos(\alpha) = \frac{\boldsymbol{X} \bullet \boldsymbol{Y}}{||\boldsymbol{X}|| \bullet ||\boldsymbol{Y}||} \tag{12}$$

3.3. Visual explainability of the deep learning system

The goal of the DL-CBR integrated workflow is to exploit the DL system automation for cases that are reliably covered by the training set, and the CBR system for the other cases. Indeed, the CBR system can also manage an outlier, which is solved via highly specialized doctors and placed in the case base. Subsequently, when sufficient variants cover the initial outlier, they can integrate the training set of the DL, enabling a new training phase. An essential aspect of this workflow is to assess the reliability of the DL system for a given case, possibly without highly specialized doctors. For this purpose, different visual explainability tools have been integrated in the workflow (Singh et al., 2020b). Specifically, Fig. 4 shows four tools used in the proposed approach:

(a) *bounding boxes*: a visualization of a rectangular bounding box, with the associated label, as in Fig. 4a, allows the human decision maker to understand whether the detected objects correspond to real lesions of the oral mucosa or not. In particular, if the box encloses photo artifacts (for example, light reflected by saliva, hard tissue such as teeth or dental prosthesis), then the output of the DL is considered not reliable.
(b) *saliency maps*: a visualization of the image that highlights via warm colors (red/yellow) the pixels of the image which are relevant for the classification. It is the most used explanation method for interpreting CNNs (Chattopadhay et al., 2018). Fig. 4b shows the saliency maps related to the image of Fig. 4a. Here, it is noticeable that the pixels related to the detected lesion are the most relevant for the classification.
(c) *scatter plot*: a visualization of a scatter plot allows to check the similarity of a new case against the case base. Fig. 4c shows the bidimensional representation of the similarity between cases, in which a different color means a different class and the "X" symbol identifies the new case. The scatter plot can be generated using a multi-dimensional scaling of the considered feature space of the similarity, such as Principal Component Analysis (PCA), or T-SNE (t-distributed stochastic neighbor embedding).

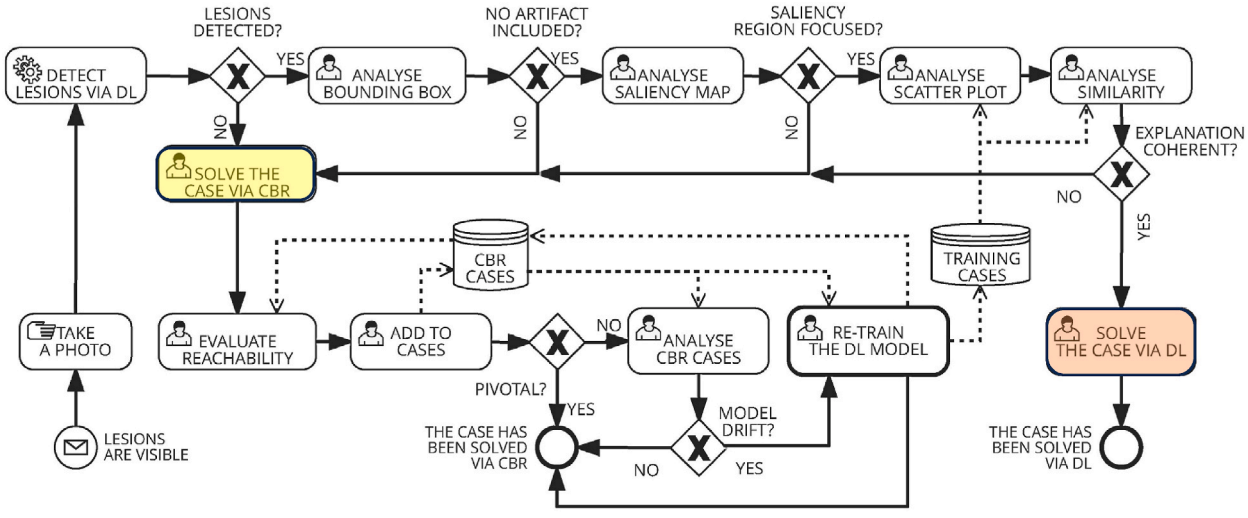

Fig. 5. – Workflow design of the integrated DL-CBR screening protocol.

(d) *similarity histogram*: since the scatter plot can remove relevant information about similarity, a visualization of the similarity of the new case with all case base complements similarity information. Fig. 4d shows, for the new case and for each class, the number of cases falling in a given range of dissimilarity. The concept of dissimilarity is similar to a distance, and can be considered as the complementary similarity value (i.e., 1−similarity). The figure clearly shows the low distance of the new case with the red class, with respect to the other two classes.

(e) *case-based explanation*: a visualization of the most similar k cases (Fig. 4e), according to the CBR methodology, allows to determine the reachability set of the new case, according to Formula (7) and (8), thus determining if an outlier has been found.

3.4. Workflow model of the integrated DL-CBR system

To provide evidence of the integrated approach, the proposed workflow has been modeled and experimented (Cimino et al., 2017). The purpose is to reduce costs by decreasing the involvement of experienced doctors in the screening process, fostering a greater involvement of non-specialized medical personnel. The language for workflow modeling is the Business Process Modeling Notation (BPMN), an international standard with a strong mathematical basis. In the BPMN, a circle, a rounded box, and a diamond, represent an event, an activity, and a decision/merge node, respectively. Sequence flow, data flow and data storage are represented by a solid arrow, a dotted arrow, and a cylindric shape, respectively.

Fig. 5 shows the BPMN workflow model of the overall methodology (Cimino et al., 2017). The workflow starts when there are lesions visible (event on bottom-left). After taking a photo, the DL architecture is used for detecting lesions. If lesions are not detected, the case are solved via CBR. Its reachability is then evaluated, and the case is added to the CBR solved cases. If the case is pivotal the process ends without other activities, because it is an outlier which does not affect the DL architecture. Otherwise, the case base is analyzed for assessing whether there is a high number of cases that have been unsolved via DL, which means that a model drift is occurring. In case of model drift, the DL model is re-trained, which involves an update of the training cases via the CBR cases. The latter are also updated according to the capabilities of the new DL architecture. On the other hand, if lesions are detected by the DL, the bounding boxes are first analyzed: for each bounding box, if an artifact is included, then the DL result is not reliable, and the case must be solved via CBR, according to the above flow. Otherwise, the saliency map is analyzed. If the saliency map is not focused, i.e., the map does not highlight the lesion region, then the case must be solved via CBR. If the saliency map is focused, then the similarity plots (i.e. scatter and histogram) are analyzed, together with the most similar/dissimilar training cases. If the explanation is coherent, the case is finally solved via DL, otherwise it is solved via CBR.

4. Case study and experimental results

4.1. The oral case study

The DL-CBR system takes as an input a photo of the oral cavity, and identifies as an output the bounding boxes where the lesion is located. Finally, it provides a label representing an oral lesion class such as “neoplastic”, “aphthous”, or “traumatic”. The DL-CBR system was developed with a dataset of oral ulcerated lesions. The images set adopted to train the CNN was initially based on 220 simple plain photographs of clinical cases, after histological confirmation (whereas necessary). All images were taken by Oral Medicine Staff (i.e. consultants, dental hygienists and students) from 2015 to 2020 in the Oral Medicine Unit of University Hospital P. Giaccone of Palermo (Italy). The cases set is publicly available (Galatolo, 2024). The full concordance of histological exams guarantees the correct identification of neoplastic lesions. Aphthous lesions are categorized after the evaluation of their clinical features, location, and course (Baccaglini et al., 2013). Traumatic ulcers can be confirmed whether an inciting or triggering trauma, condition, or medication can be identified.

The training, validation and test sets were generated using the holdout method, with percentages of 70%, 15% and 15%, respectively. Table 1 summarizes the sample sizes of the dataset and the frequencies by class. All selected images include one or more lesions belonging to the same class. An input image is made by 800 $\times$800 pixels $\times$ 3 colors (BGR). No images of mixed classes were included in the dataset. Regions of interest of interest containing the oral lesion were manually segmented.

Due to the limited sample size, it was implemented a balanced sampler over an imbalanced dataset. Each image was assigned to a weight as the frequency of its class and those were used as weights of a multinomial distribution from which the training images were sampled. This technique is called Mask R-CNN with Repeat Factor Sampling (Wang et al., 2019). The frequency threshold t controls the degree of resampling of rare categories.

For each proposed scenario, a hyperparameter optimization is carried out. The TPE is used as hyperparameters sampler, and the Successive Halving Pruner to early stop unpromising runs (Bergstra et al., 2011). The objective function for the hyperparameters

Table 1
Dataset summary.

Dataset	# samples	# aphthous	# traumatic	# neoplastic
Whole	318	110	111	97
Train	220	76	77	67
Validation	50	18	17	15
Test	48	16	17	15

optimization is the Overall Accuracy rate defined as Detection rate × Classification rate, computed on the validation set. The run limit for each hyperparameter optimization was set to 130.

In order to assess the effectiveness of the proposed approach, an operational environment has been developed to support the methodological workflow (Coelho, 2012). Table 2 shows the available web-based resources, which are publicly released to foster a research initiative called "DoctOralAi", based on two pillars: (i) the collaboration between clinical centers in collecting more data on regional/national basis, (ii) the collaboration between deep learning development centers in order to further develop and achieve the best DL performance in this field.

4.2. The training of the deep learning system

To ensure reproducible trials, the most sensitive hyperparameters have been selected by adopting an optimization algorithm based on a TPE that selects the best choice (Bergstra et al., 2011). As a result, Table 3 summarizes the bests hyperparameters determined by the optimization, whereas Table 4 summarizes the performance evaluation of the DL-based lesion detection and classification system. On average, the hyperparameters optimization has been carried out in 35 trials and 8.13 h, using the following hardware resources: GPU NVIDIA™ GeForce RTX 2080; CPU Intel® Core™ i9-9900K @ 3.60 GHz; CACHE L1 512 KB, L2 2 MB, L3 16 MB. On average, a training session is carried out in 13.94 min. Fig. 6 shows the classification loss against the number of iterations over 10 trials. In terms of accuracy (Table 4), the detection and classification rates achieved are 82.9% and 92.0%, respectively.

4.3. Pilot cases discussion

This section illustrates two pilot cases (Figs. 15–21). For a detailed reporting of all cases, the interested reader is referred to reference (DoctOralAi research initiative, 2021).

Case A relates to a neoplastic lesion, clearly identified by the Detector (Fig. 10). The saliency map, shown in Fig. 11, shows that the most important regions are those related to the lesion. Fig. 12 shows the most similar neoplastic case. The position of Case A in the neoplastic class is also clear in the scatter plot of Fig. 7. The similarity histogram of Fig. 8 shows that the distance of case A from the neoplastic cases is lower than from the aphthous and traumatic ones, thus confirming the provided classification. Not surprisingly, the membership degrees of Case A to the aphthous, traumatic and healthy (non-object) classes are not significant (Table 5). As a result, Case A is fully detected, classified, and explained.

Case B, illustrated in Fig. 13, is characterized by three bounding boxes. Only the largest box B3 includes the complete lesion. Indeed, the saliency map in Fig. 14 shows that the relevant region is mostly included in Box B3. Both the scatter plot (Fig. 9) and the similarity histogram (Fig. 23) show that the case is neoplastic. The membership degree of Case B to the neoplastic class is 84.8% (Table 6), lower than Case A, mostly because there is a higher membership degree to the healthy class. In general, it can be realized that this difference of about 10% in membership degree can be neglected.

4.4. Screening protocol experimentation

To carry out an experimentation of the screening protocol illustrated in Fig. 5, the following procedure has been carried out. A group of 9 resident doctors and a group of 6 specialized doctors have been involved in the screening of 20 cases using DoctOralAi. The 20 cases have been purposely selected to include a variety of border line cases, difficult to solve for a resident doctor. In the experiment, each doctor carries out the workflow of Fig. 5 to each case, via an illustrated questionnaire. The goal is to show that both the specialized and the resident doctor can understand if the DL is not reliable, by moving to a human-driven diagnosis based on CBR. Fig. 22 shows the heatmap of the experiment made by the group of specialized doctors. Here, a hot/cold color denotes a frequent/infrequent activity, respectively, according to the colorbar on the bottom left.

Overall, it is clear (from the red tasks at the top) that most cases were detected by the DL without artifacts and with saliency region focused. Since the selected cases are borderline, the DL was able to solve about the 64% of them (colored in orange on the right), whereas the remaining 36% were solved via CBR (colored in yellow on the left).It is expected that for increasing number of cases in the machine learning sets, the system improves its capability of problem solving, thus moving the workflow on the right side of the picture, supported by the DL. Table 7 shows the resulting percentages of questions answered with "NO", i.e. questions causing an execution of the CBR in Fig. 22. Not surprisingly, the resident doctors are more prone to rely on CBR than on DL. Specifically, the tools that are more sensitive to the specialization level are the scatter plot (12.22% vs 26.42%) and the group of similarity tools (36.67% vs 54.72%).

In order to compare the experimental results with other approaches in the literature, it is worth to highlight some aspects related to

Table 2
Web-based toolchain for managing the DoctOralAi initiative.

Tool	Description
Cases	The complete set of 318 cases.
Annotator	A web-based image annotation tool.
DL Demonstrator	A web-based detector and classifier.
Reporting	Reports generated from cases.
Source code	A Repository with the python code.

Table 3
Best hyperparameters determined by the optimization.

Hyperparameter	value
Learning rate	0.00108904
RPN loss weight	0.95390
ROI heads loss weight	0.82677
ROIs per image	256
Repeat factor threshold	0.33561

Table 4
Performance evaluation of the DL-based lesion detection and classification.

Performance metric	Conf. interval
Hyperparameterization time (hrs, 35 trials)	11.56 ± 0.023
Training time (min.)	15.71 ± 0.067
Detection rate (%)	0.829 ± 0.028
Classification rate (%)	0.920 ± 0.010

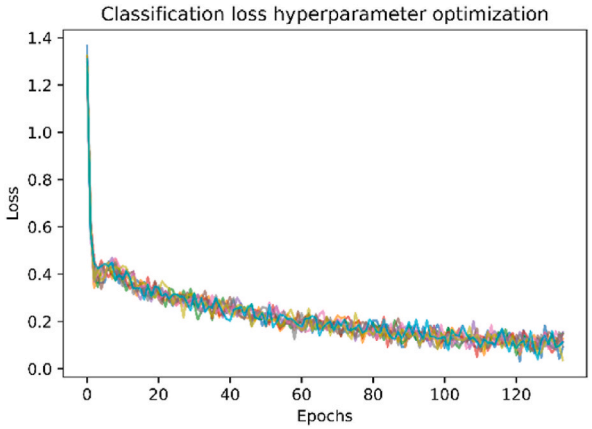

Fig. 6. – Classification loss against number of iterations, over 10 trials.

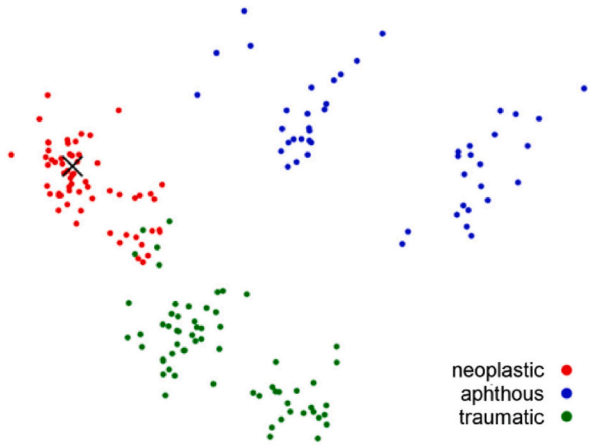

Fig. 7. – Case A – scatter plot (neoplastic 98.8%).

cases. This study includes only histologically proven OSCC lesions of patients of the Unit of Oral Medicine at the University Hospital of Palermo, which were followed and treated when needed. Each patient underwent a biopsy after their pictures were taken, and the histology results from the biopsy were used to label the corresponding images in the training set. Moreover, experiments have been based on only simple plain photographs, with a great degree of clinical practice acceptance also in developing countries, or simply

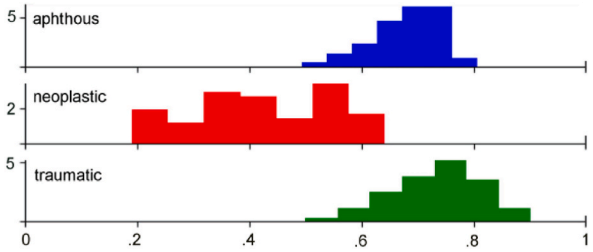

Fig. 8. Case A – (dis-)similarity histogram (neoplastic 98.8%).

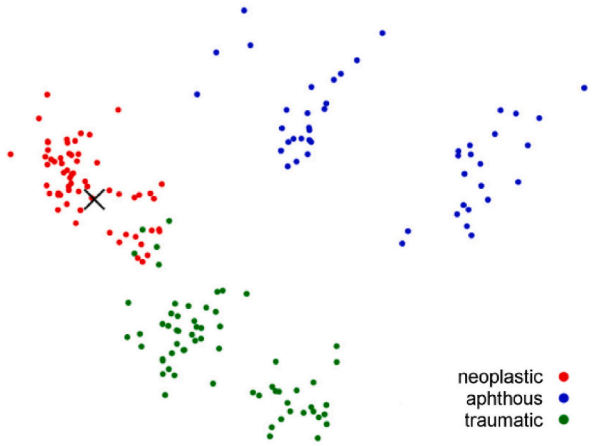

Fig. 9. – Case B – scatter plot (neoplastic 84.8%).

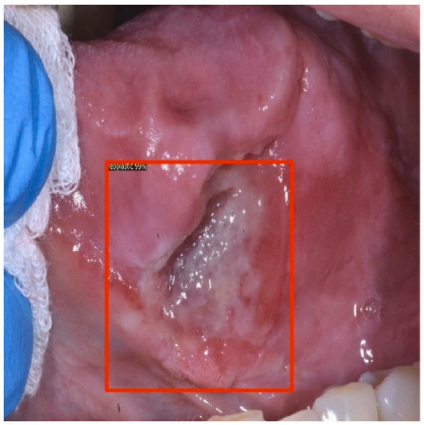

Fig. 10. – Case A, bounding box (neoplastic 98.8%).

when other oral cavity imaging systems evaluation is not available (e.g., hyperspectral images, autofluorescence based images and OCT images).

5. Discussion

Table 8 defines the performance measures adopted for a comparison with the state-of-the art. Specifically, the "neoplastic" class has been represented with the symbol "+", whereas the remaining classes have been represented with "−". Table 8(b) shows the definitions of: (i) *true positive* (T^+), if the classifier correctly detects a neoplastic lesion; (ii) *true negative* (T^-), if it correctly detects an "aphthous" or "traumatic" lesion; (iii) *false negative* (F^-), if the classifier erroneously detects an "aphthous" or "traumatic" lesion that is actually "neoplastic"; (iv) finally *false positive* (F^+) if the classifier erroneously detects a "neoplastic" lesion that is actually an "aphthous" or "traumatic". According to the above definitions, Table 8(a) formally expresses the following measures: Accuracy (A),

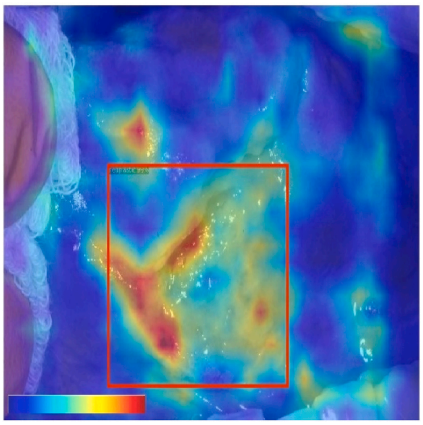

Fig. 11. – Case A, saliency map (neoplastic 98.8%).

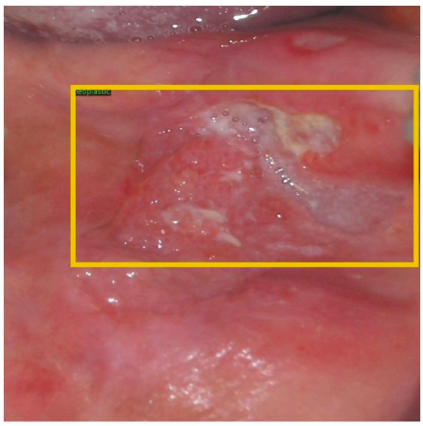

Fig. 12. – Case A, the most similar neoplastic case (dist 0.191).

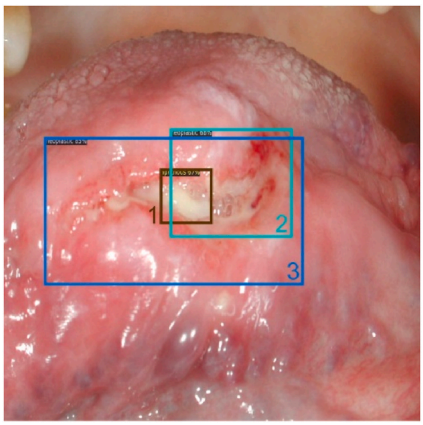

Fig. 13. – Case B: B1 aphthous 66.6%, B2 neoplastic 68.2%, B3 neoplastic 84.8%.

Precision (P), Recall/Sensitivity (R), Specificity (S), and F_1 score.

Table 9 shows the confusion matrix on the test set. Specifically, accuracy values per class are the following: (A)phthous 100%, (T) raumatic 100%, and (N)eoplastic 94%.

In the literature, different deep networks and data sets are used. However, there is still a lack of available benchmark data. For this reason, a quantitative comparison is often not feasible. To assess the effectiveness of our approach, the research work called MeMoSA,

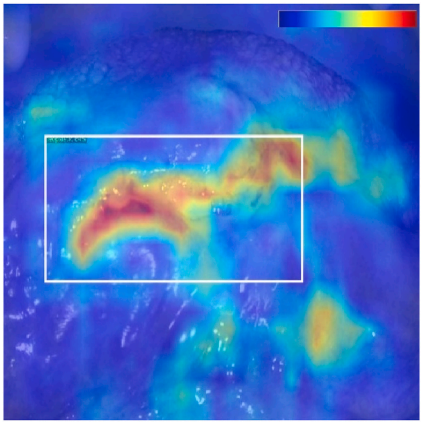

Fig. 14. – Case B3, saliency map (neoplastic 84.8%).

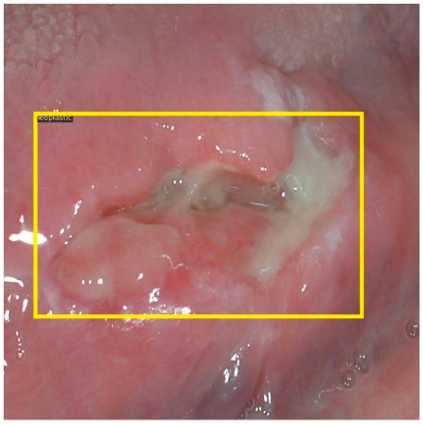

Fig. 15. – Case B3, the most similar neoplastic case (dist 0.101).

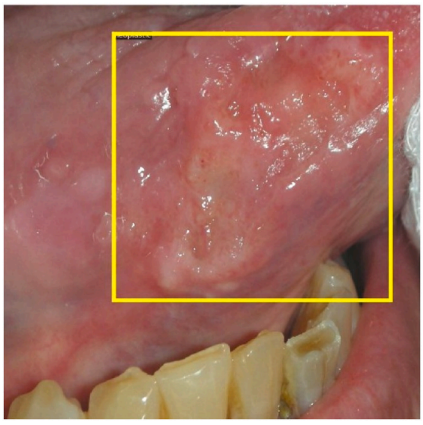

Fig. 16. – Case B3, a similar neoplastic case (dist 0.168).

published in Welikala et al. (Welikala et al., 2020), has been considered as a reference. The reason is that the base DL architecture is similar to OralCancerAi, although the cases are different. Table 10 shows that the performance of the two approaches, i.e., DoctoralAi and MeMoSA, which for the detection task are very similar. Finally, Table 11 shows the performance for the classification tasks: although they are not directly comparable, because calculated on different datasets, the effectiveness of DoctOralAi is clear with respect to MeMoSA.

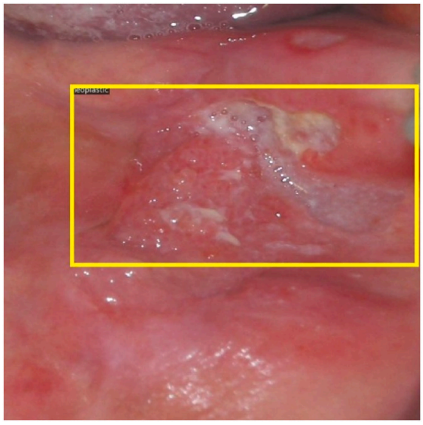

Fig. 17. – Case B3, a similar neoplastic case (dist 0.175).

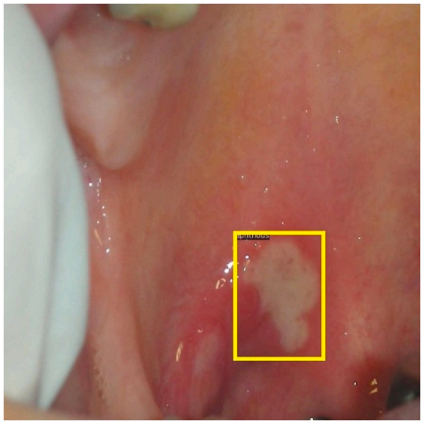

Fig. 18. – Case B3, the most similar aphthous case (dist 0.520).

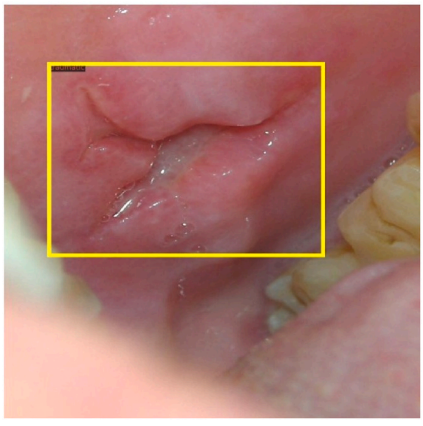

Fig. 19. – Case B3, the most similar traumatic case (dist. 0.389).

Some other studies have been published regarding the use of AI technology in order to classify medical images, both coming from indirect images, such as pathology specimens and radiological images, or from direct images, such as dermoscopic or fundus oculi or oral simple photographs (Ilhan et al., 2020) (Lin et al., 2021). A study by Bofan Son et al. (Song et al., 2018) presented an image classification approach based on autofluorescence and white light images using DL methods. Images of 170 cases were captured from buccal mucosa, gingiva, soft palate, vestibule, floor of mouth and tongue and the information from both the autofluorescence and

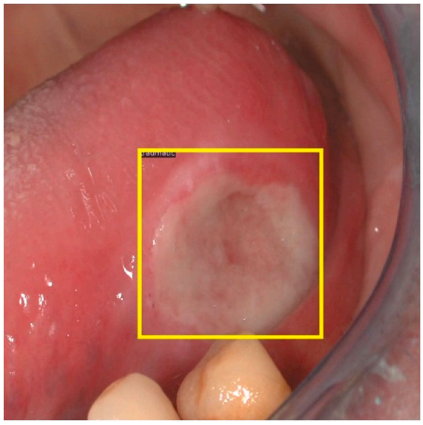

Fig. 20. – Case B3, a similar traumatic case (dist. 0.419).

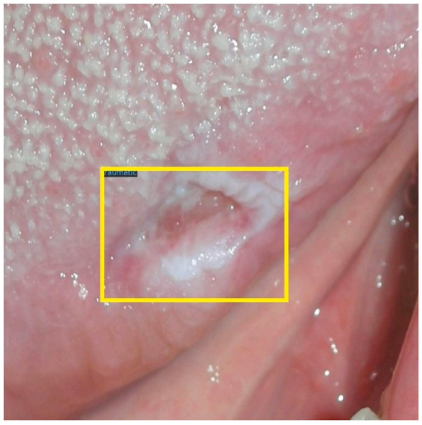

Fig. 21. – Case B3, a similar traumatic case (dist 0.431).

Table 5
Case A – Membership degrees.

Class	Membership degree
Neoplastic	98.807
Aphthous	0.516
Traumatic	0.242
Healthy	0.435

Table 6
Case B – Membership degrees.

Class	Membership degree
Neoplastic	84.848
Aphthous	2.135
Traumatic	1.819
Healthy	11.198

white light image pair were extracted, calculated, and fused to feed the DL model. Three different architectures were used: VGG-CNN-M, VGG-CNN-S, and VGG-16. Then, data augmentation was implemented due to small dataset size, achieving the best result using transfer learning of the VGG-CNN-M module along with the data augmentation of the dual-modal images, resulting in a 4-fold cross-validation average accuracy of 86.9% at a sensitivity of 85.0% and specificity of 88.7%.

Other papers that are worth mentioning are two studies conducted by Ling Ma et al. (Ma et al., 2017) and Jeyaraj et al. (Jeyaraj et al., 2019b). With an average accuracy of 91.36%, the first study demonstrated an OSCC classification model utilizing CNNs to

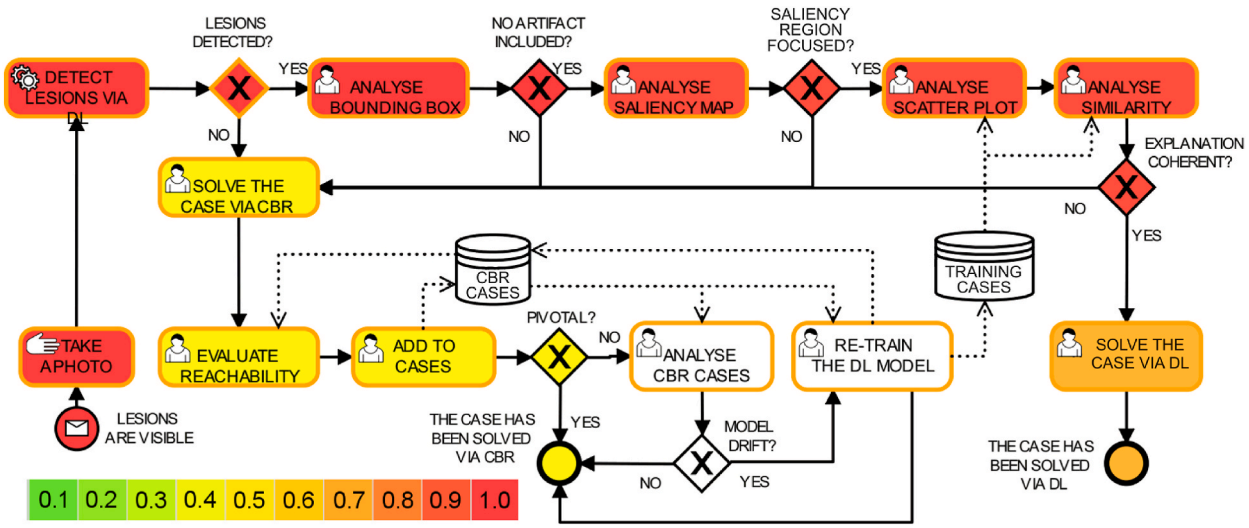

Fig. 22. – Heatmap based on counts of executions of the DoctOralAi.

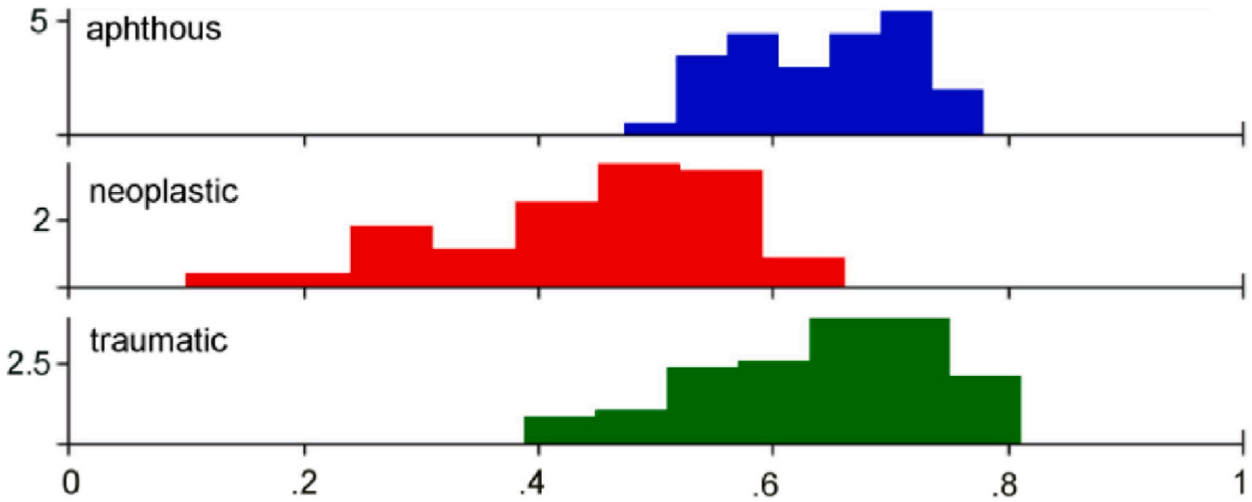

Fig. 23. – Case B3 – (dis-)similarity histogram (neoplastic 84.8%).

Table 7
Gateways and related percentages.

Question	NO (%) specialized doctor	NO (%) resident doctor
a) Lesions detected?	0.00	0.00
b) No artifact included?	5.56	5.00
c) Saliency region focused?	5.88	5.26
d) Scatter plot coherent?	12.22	26.42
e) Dissimilarity histogram coherent?	25.32	30.19
f) Inter-class similarity coherent?	3.39	2.70
g) Extra-class similarity coherent?	19.44	27.59
h) (e) or (f) or (g)	36.67	54.72
i) Pivotal?	1.00	1.00
j) Model drift?	99.50	99.50

identify head and neck cancer in mice using hyperspectral images (HSIs). The accuracy, specificity, and sensitivity of the second paper's analysis of HSIs of human oral cancer case studies were 94.5%, 98%, and 94%, respectively. Nevertheless, despite the very interesting results achieved in the last two studies, algorithms were developed on hyperspectral images only (HSI collects

Table 8
– Performance measures.

		Predicted class	
		+	–
Actual Class	+	T^+	F^-
	–	F^+	T^-

LEGEND.
•+ *neoplastic*.
•– *aphthous* or *traumatic*.
•Accuracy: A = T/(T + F).
•Precision: $P = T^+/(T^+ + F^+)$.
•Recall (Sensitivity): $R = T^+/(T^+ + F^-)$.
•Specificity $S = T^-/(T^- + F^+)$.
•F_1 score = (2P·R)/(P + R).

Table 9
Confusion matrix on the test set: Predicted (P) and Actual (A) cases on (A)phthous, (T)raumatic, and (N)eoplastic classes.

	A^A	T^A	N^A
A^P	15	0	1
T^P	0	17	0
N^P	0	0	15

Table 10
Detection performance comparison.

Approach	A
DoctOralAi	0.83
MeMoSA	0.81

Table 11
Classification performance comparison *neoplastic* vs. *no neoplastic*.

Approach	A	P	R	S	F_1
DoctOralAi	0.98	0.94	1.00	0.97	0.97
MeMoSA	0.75	0.67	0.94	0.57	0.78

high-resolution images at numerous spectral bands, providing big data to distinguish between various tissue types), which are not easily accessible. In contrast, the approach proposed in this paper is based on simple clinical photographs, to evaluate the potential impact of a cost effective and widely available screening method via DL techniques.

6. Conclusions and future work

In this research work, an explainable approach to the screening of oral cancer based on Convolutional Neural Networks (CNN) and Case-Based Reasoning (CBR) is discussed. The main contribution of this paper is the combination of Deep Learning (DL) and CBR supporting the post-hoc explanation of the system answer. This explanation allows for an easy integration of human decision makers and DL. On one side, explainable DL assesses the reliability of the DL result, and from the other side CBR guided by human expertise addresses unreliable results provided by the DL.

A CNN and a related workflow have been developed and publicly released, to foster the collaboration between clinical centers in collecting more data on regional/national basis, as well as the collaboration between DL development centers in order to achieve the best networks in this field. The DL system is able to achieve state-of-the-art performance, and the proposed decision process is able to correctly address even resident doctors when dealing with border line cases. With an increasing number of cases, the system is expected to be increasingly able to rely on DL instead of on human expertise. Moreover, the human-driven CBR is able to solve the cases currently not solved by DL, which can be distinguished via the explainability techniques.

An alternative strategy to the object detection performed with the Detectron 2 framework is to exploit YOLO, an architecture proposed by Redmon et al. (Redmon et al., 2016) in 2016, which carries out the task of object detection and classification in a one-step process, considerably speeding up the execution of the analysis. YOLO has been used in the clinical context, and recently also for the

oral cavity (Sonavane & Kohar, 2022), however it has not yet been integrated within a CBR system. In recent studies Transformers have shown promising results as advanced detection methods, although they typically require larger datasets to achieve high detection accuracy (Liu et al., 2023). In future works, it may be necessary to expand the dataset to optimize performance and fully leverage the capabilities of additional detection techniques.

In conclusion, the proposed methodology shows effectiveness and potentially significant advancements in early detection of oral cancer. However, to further validate and enhance our findings, future work will focus on conducting a comprehensive comparative analysis with other existing approaches. This will be achieved by leveraging publicly available datasets, for a more extensive evaluation of our approach in diverse clinical scenarios.

CRediT authorship contribution statement

Mario G.C.A. Cimino: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization. **Giuseppina Campisi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Federico A. Galatolo:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Investigation, Conceptualization. **Paolo Neri:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Pietro Tozzo:** Data curation. **Marco Parola:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Gaetano La Mantia:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation. **Olga Di Fede:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation, Conceptualization.

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.

Acknowledgments

The Authors thank Dr. Chiara Terranova for her valuable help in data collection. Work partially supported by: (i) the University of Pisa, in the framework of the PRA 2022 101 project “Decision Support Systems for territorial networks for managing ecosystem services”; (ii) the European Commission under the NextGenerationEU program, Partenariato Esteso PNRR PE1 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 ″Human-centered AI"; (iii) the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence), in the framework of the "Reasoning" project, PRIN 2020 LS Programme, Project number 2493 04-11-2021, and in the framework of the project "OCAX - Oral CAncer eXplained by DL-enhanced case-based classification" PRIN 2022 code P2022KMWX3. Work partly funded by the European Commission under the NextGeneration EU program, PNRR - M4 C2, Investment 1.5 ″Creating and strengthening of "innovation ecosystems", building "territorial R&D leaders", project "THE - Tuscany Health Ecosystem", Spoke 6 ″Precision Medicine and Personalized Healthcare". Work partially funded by the European Union—NextGenerationEU (National Sustainable Mobility Center CN00000023, Italian Ministry of University and Research Decree n. 1033–17/06/2022, Spoke 10)".

Data availability

Data and code have been shared as a reference link in the manuscript

References

Adeoye, J., Tan, J. Y., Choi, S. W., & Thomson, P. (2021). Prediction models applying machine learning to oral cavity cancer outcomes: A systematic review. *International Journal of Medical Informatics, 154*, Article 104557, 10.1016/j.ijmedinf.2021. 104557. Epub 2021 Aug 18. PMID: 34455119.

Ahn, J., Ji, S. H., Ahn, S. J., Park, M., Lee, H. S., Kwon, N., … Kim, Y. (2020). Performance evaluation of normalization-based CBR models for improving construction cost estimation. *Automation in Construction, 119*, Article 103329.

American Cancer, Society.. Treating Oral Cavity and Oropharyngeal Cancer. cancer.org | 1.800.227.2345. https://www.cancer.org/cancer/types/oral-cavity-and-oropharyngeal-cancer/treating.html.

Amit, Y., Felzenszwalb, P., & Girshick, R. (2020). Object detection. *Computer vision: A reference guide*.

Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine, 127*, 248–257.

Baccaglini, L., Theriaque, D. W., Shuster, J. J., Serrano, G., & Lalla, R. V. (2013). Validation of anamnestic diagnostic criteria for recurrent aphthous stomatitis. *Journal of Oral Pathology & Medicine, 42*(4), 290–294.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyperparameter optimization. *Advances in Neural Information Processing Systems, 24*.

Bhattacharya, S., Maddikunta, P. K. R., Pham, Q. V., Gadekallu, T. R., Chowdhary, C. L., Alazab, M., & Piran, M. J. (2021). Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey. *Sustainable Cities and Society, 65*, Article 102589.

Bouaud, J., Séroussi, B., Falcoff, H., Julien, J., Simon, C., & Denké, D. L. (2009). Consequences of the verification of completeness in clinical practice guideline modeling: A theoretical and empirical study with hypertension. *AMIA symposium*, 60–64.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE winter conference on applications of computer vision (WACV 2018)* (pp. 839–847). IEEE. https://christophm.github.io/interpretable-ml-book/pixel-attribution.html.

Chen, D., Liu, S., Kingsbury, P., et al. (2019). Deep learning and alternative learning strategies for retrospective real-world clinical data. *Nature Partner Journal Digit. Med., 2*, 43.

Chuang, C. L. (2011). Case-based reasoning support for liver disease diagnosis. *Artificial Intelligence in Medicine, 53*(1), 15–23.
Cimino, M. G. C. A., Palumbo, F., Vaglini, G., et al. (2017). Evaluating the impact of smart technologies on harbor's logistics via BPMN modeling and simulation. *Information Technology and Management, 18*, 223–239.
Coelho, K. R. (2012). Challenges of the oral cancer burden in India. *J Cancer Epidemiol, 2012*, Article 701932. https://doi.org/10.1155/2012/701932. Epub 2012 Oct 4. PMID: 23093961; PMCID: PMC3471448.
Cohen, L. A., Bonito, A. J., Eicheldinger, C., Manski, R. J., Macek, M. D., Edwards, R. R., et al. (2011). Comparison of patient visits to emergency departments, physician offices, and dental offices for dental problems and injuries. *Journal of Public Health Dentistry, 71*(1), 13–22. https://doi.org/10.1111/j.1752-7325.2010.00195.x
Conway, D. I., et al. (2002). *Oral cancer: Prevention and detection in primary dental healthcare", primary dental care 4*.
Dhanuthai, K., Rojanawatsirivej, S., Thosaporn, W., Kintarak, S., Subarnbhesaj, A., Darling, M., Kryshtalskyj, E., Chiang, C.-P., Shin, H.-I., Choi, S.-Y., Lee, S.-S., & Aminishakib, P. (2018). Oral cancer: A multicenter study. *Med Oral Patol Oral Cir Bucal, 23*(1), e23–e29.
Dinesh, Y., Ramalingam, K., Ramani, P., & Mohan Deepak, R. (2023). Machine learning in the detection of oral lesions with clinical intraoral images. *Cureus, 15*(8), Article e44018. https://doi.org/10.7759/cureus.44018. PMID: 37753028; PMCID: PMC10519616.
DoctOralAi research initiative. https://mlpi.ing.unipi.it/doctoralai, (2021).
Du, G. F., Li, C. Z., Shang, S. H., Xu, X. Y., Chen, H. Z., & Zhou, G. (2013). Practising case-based learning in oral medicine for dental students in C hina. *European Journal of Dental Education, 17*(4), 225–228.
Ehtesham, H., Safdari, R., Mansourian, A., Tahmasebian, S., Mohammadzadeh, N., & Pourshahidi, S. (2019). Developing a new intelligent system for the diagnosis of oral medicine with case-based reasoning approach. *Oral Diseases, 25*(6), 1555–1563.
Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2022). A survey on bias in visual datasets. In *Computer vision and image understanding* (Vol. 223), Article 103552.
Fitzpatrick, S. G., Cohen, D. M., & Clark, A. N. (2019). Ulcerated lesions of the oral mucosa: Clinical and histologic review. *Head Neck Pathol, 13*(1), 91–102. https://doi.org/10.1007/s12105-018-0981-8. Epub 2019 Mar 7. PMID: 30701449; PMCID: PMC6405793.
Fu, Q., Chen, Y., Li, Z., Jing, Q., Hu, C., Liu, H., … Xiong, X. (2020). A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study. *EClinicalMedicine, 27*, Article 100558.
Galatolo, F. A. (2024). Github repository, oral lesions detection. https://github.com/galatolofederico/oral-lesions-detection.
García-Martín, J. M., García-Pola, M. J., Varela-Centelles, P., & Seoane-Romero, J. M. (2020). On the role of physicians in oral cancer diagnosis. *Oral Oncology, 108*, Article 104843. https://doi.org/10.1016/j.oraloncology.2020.104843. ISSN 1368-8375.
Gigliotti, J., Madathil, S., & Makhoul, N. (2019). Delays in oral cavity cancer. *International Journal of Oral and Maxillofacial Surgery, 48*(Issue 9).
Gu, D., Su, K., & Zhao, H. (2020). A case-based ensemble learning system for explainable breast cancer recurrence prediction. *Artificial Intelligence in Medicine, 107*, Article 101858.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV*. https://doi.org/10.1109/CVPR.2016.90
Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). *What do we need to build explainable AI systems for the medical domain? arXiv*, Article 09923. arXiv:1712.
Ilhan, B., Lin, K., Guneri, P., & Wilder-Smith, P. (2020). Improving oral cancer outcomes with imaging and artificial intelligence. *Journal of Dental Research, 99*(3), 241–248. https://doi.org/10.1177/0022034520902128. PMID: 32077795; PMCID: PMC7036512.
Jeyaraj, P. R., & Nadar, E. R. S. (2019a). Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. *Journal of Cancer Research and Clinical Oncology, 145*, 829–837.
Jeyaraj, P. R., & Nadar, E. R. S. (2019b). Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. *Journal of Cancer Research and Clinical Oncology, 145*(4), 829–837.
Johs, A. J., Lutts, M., & Weber, R. O. (2018). Measuring explanation quality in XCBR. In *Proceedings of ICCBR 2018* (p. 75).
Juarez, J. M., Craw, S., Lopez-Delgado, J. R., & Campos, M. (2018). *Maintenance of case bases: Current algorithms after fifty years*. IJCAI International Joint Conferences on Artificial Intelligence Organization.
Keane, M. T., & Kenny, E. M. (2019). How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems. In *International conference on case-based reasoning* (pp. 155–171). Cham: Springer.
Kiranyaz, S., Ince, T., & Gabbouj, M. (2015). Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Transactions on Biomedical Engineering, 63*(3), 664–675.
Kolodner, J. (2014). *Case-based reasoning*. Morgan Kaufmann.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90.
Kwon, N., Lee, J., Park, M., Yoon, I., & Ahn, Y. (2019). Performance evaluation of distance measurement methods for construction noise prediction using case-based reasoning. *Sustainability, 11*(3), 871.
Lamy, J. B., Sekar, B., Guezennec, G., Bouaud, J., & Séroussi, B. (2019a). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine, 94*, 42–53.
Lamy, J. B., Sekar, B., Guezennec, G., Bouaud, J., & Séroussi, B. (2019b). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine, 94*, 42–53.
Lin, H., Chen, H., Weng, L., Shao, J., & Lin, J. (2021). Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis. *Journal of Biomedical Optics, 26*(8), Article 086007. https://doi.org/10.1117/1.JBO.26.8.086007. PMID: 34453419; PMCID: PMC8397787.
Liu, Z., Lv, Q., Yang, Z., Li, Y., Lee, C. H., & Shen, L. (2023). Recent progress in transformer-based medical image analysis. *Computers in Biology and Medicine*, Article 107268.
Ma, L., et al. (2017). Deep learning-based classification for head and neck cancer detection with hyperspectral imaging in an animal model. In *Medical imaging 2017: Biomedical applications in molecular, structural, and functional imaging* (Vol. 10137). International Society for Optics and Photonics.
Mingxing, T., & Le, Q. V. (1905). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. Of int. Conf. On machine learning (ICML) 2019, arXiv*, Article 11946v5.
Nagao, T., & Warnakulasuriya, S. (2020). Screening for oral cancer: Future prospects, research and policy development for Asia. *Oral Oncology, 105*, Article 104632. https://doi.org/10.1016/j.oraloncology.2020.104632. ISSN 1368-8375.
Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, S. M. R., Jafari, M. H., Ward, K., & Najarian, K. (2016). Melanoma detection by analysis of clinical images using convolutional neural network. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 1373–1376). IEEE.
Okunseri, C., Pajewski, N. M., Jackson, S., & Szabo, A. (2011). Wisconsin Medicaid enrollees' recurrent use of emergency departments and physicians' offices for treatment of nontraumatic dental conditions. *Journal of the American Dental Association, 142*(5), 540–550. https://doi.org/10.14219/jada.archive.2011.0224
Omprakash, A., Kumar, A. P., & Padmavathi, R. (2018). Perceptions of first year dental students on case based learning in Physiology. *International Archives of Integrated Medicine, 5*(4).
Padilla, R., Netto, S. L., & Da Silva, E. A. (2020). A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (IWSSIP)* (pp. 237–242). IEEE.
Priya, S., & Uthra, R. A. (2023). Deep learning framework for handling concept drift and class imbalanced complex decision-making on streaming data. *Complex & Intelligent Systems, 9*, 3499–3515. https://doi.org/10.1007/s40747-021-00456-0
Ragnemalm, I. (1993). The Euclidean distance transform in arbitrary dimensions. *Pattern Recognition Letters, 14*(11), 883–888.
Ray, S. (2019). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35–39). IEEE.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).
Shrestha, A., Marla, V., Rimal, J., Shrestha, S., Keshwar, S., & Zhimin, J. (2021). Case based learning as a dynamic approach towards learning oral pathology. *bioRxiv, 2021–04*.
Simonyan, K., & Zisserman, A.. Very deep convolutional networks for large-scale image recognition. http://arxiv.org/abs/1409.1556.
Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020a). Explainable deep learning models in medical image analysis. *Journal of Imaging, 6*(6), 52.
Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020b). Explainable deep learning models in medical image analysis. *Journal of Imaging, 6*(6), 52.
Sonavane, A., & Kohar, R. (2022). Dental cavity detection using YOLO. In *Proceedings of data analytics and management* (pp. 141–152). Singapore: Springer.
Song, B., et al. (2018). Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning. *Biomedical Optics Express, 9*(11), 5318–5329.
Soomro, T. A., Zheng, L., Afifi, A. J., Ali, A., Yin, M., & Gao, J. (2022). Artificial intelligence (AI) for medical imaging to combat coronavirus disease (COVID-19): A detailed review with direction for future research. *Artificial Intelligence Review, 55*(2), 1409–1439.
Sørmo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review, 24*(2), 109–143.
Srivastava, S., Divekar, A. V., Anilkumar, C., et al. (2021). Comparative analysis of deep learning image detection algorithms. *J Big Data, 8*, 66. https://doi.org/10.1186/s40537-021-00434-w
Suh, Y. J., Jung, J., & Cho, B. J. (2020). Automated breast cancer detection in digital mammograms of various densities via deep learning. *Journal of Personalized Medicine, 10*(4), 211.
Sun, S., & Huang, R. (2010). An adaptive k-nearest neighbor algorithm. In *2010 seventh international conference on fuzzy systems and knowledge discovery* (Vol. 1, pp. 91–94). IEEE.
Sun, L., Yin, C., Xu, Q., & Zhao, W. (2023). Artificial intelligence for healthcare and medical education: A systematic review. *Am J Transl Res, 15*(7), 4820–4828. PMID: 37560249; PMCID: PMC10408516.
Szczepaniak, P. S., & Duraj, A. (2018). Case-based reasoning: The search for similar solutions and identification of outliers. *Complexity, 2018*. Article ID 9280787, 12 pages.
Szegedy, C., et al. (2015). "Going deeper with convolutions. In *" 2015 IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA* (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594
Tan, Y., Wang, Z., Xu, M., et al. (2023). Oral squamous cell carcinomas: State of the field and emerging directions. *International Journal of Oral Science, 15*, 44. https://doi.org/10.1038/s41368-023-00249-w
Wang, T., Li, Y., Kang, B., Li, J., Jun Hao Liew, Tang, S., Hoi, S., & Feng, J. (2019). *Classification calibration for long-tail instance segmentation*. arXiv preprint arXiv: 1910.13081.
Welikala, R. A., Remagnino, P., Lim, J. H., Chan, C. S., Rajendran, S., Kallarakkal, T. G., … Barman, S. A. (2020). Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. *IEEE Access, 8*, 132677–132693.
World Health Organization (WHO), Oral cancer: early diagnosis and screeningwww.who.int/cancer/prevention/diagnosis-screening/oral-cancer/en/, accessed January 2021.[.
Wu, Y., Kirillov, A., Massa, F., Wan-Yen Lo, & Girshick, R. (2019). *Detectron2*. https://github.com/facebookresearch/detectron2. https://research.fb.com/wp-content/uploads/2019/12/4.-detectron2.pdf.