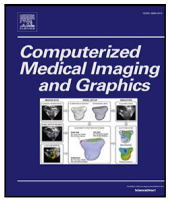
Computerized Medical Imaging and Graphics 117 (2024) 102433

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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



Towards explainable oral cancer recognition: Screening on imperfect images via Informed Deep Learning and Case-Based Reasoning

Marco Parola [a,*], Federico A. Galatolo [a], Gaetano La Mantia [b,c,d], Mario G.C.A. Cimino [a], Giuseppina Campisi [b], Olga Di Fede [b,c]

[a] Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, Pisa, 56122, Italy
[b] Department Di.Chir.On.S, 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 Biomedical and Dental Sciences and Morphofunctional Imaging, University of Messina, Messina, Italy

ARTICLE INFO

Keywords:
Oral cancer
Oncology
Medical imaging
Case-based reasoning
Informed deep learning
Explainable artificial intelligence

ABSTRACT

Oral squamous cell carcinoma recognition presents a challenge due to late diagnosis and costly data acquisition. A cost-efficient, computerized screening system is crucial for early disease detection, minimizing the need for expert intervention and expensive analysis. Besides, transparency is essential to align these systems with critical sector applications. Explainable Artificial Intelligence (XAI) provides techniques for understanding models. However, current XAI is mostly data-driven and focused on addressing developers' requirements of improving models rather than clinical users' demands for expressing relevant insights. Among different XAI strategies, we propose a solution composed of Case-Based Reasoning paradigm to provide visual output explanations and Informed Deep Learning (IDL) to integrate medical knowledge within the system. A key aspect of our solution lies in its capability to handle data imperfections, including labeling inaccuracies and artifacts, thanks to an ensemble architecture on top of the deep learning (DL) workflow. We conducted several experimental benchmarks on a dataset collected in collaboration with medical centers. Our findings reveal that employing the IDL approach yields an accuracy of 85%, surpassing the 77% accuracy achieved by DL alone. Furthermore, we measured the human-centered explainability of the two approaches and IDL generates explanations more congruent with the clinical user demands.

1. Introduction

Oral squamous cell carcinoma (OSCC) represents a challenge in oncology, with a significant global burden of morbidity and mortality. As one of the most prevalent malignancies affecting the oral cavity, OSCC not only poses a substantial threat to public health but also underscores the critical need for improved early detection strategies (Kim and Kim, 2020). The etiological factors contributing to OSCC are multifaceted, encompassing tobacco use, alcohol consumption, viral infections, and genetic predisposition (Bugshan and Farooq, 2020).

Currently, the main adjuvant treatments for OSCC are chemotherapy and radiotherapy, with surgical resection serving as the primary therapy. Nevertheless, with such multimodality treatments, which result in severe mutilation and a parallel deterioration in life quality, the 5-year overall survival rate stays at 60% (Zhou et al., 2022). Early identification of OSCC is critical for successful intervention and improved patient outcomes. In this context, screening emerges as a crucial tool in the proactive management of oral cancer (Bramati et al., 2021).

Screening plays a key role in the early diagnosis of OSCC, as it is a proactive approach to identify individuals at risk or in the early stages of the disease. Effective screening must be rapid, applicable on a mass scale, and not depend on expensive or invasive instrumentation. The need for time-efficient analysis further underscores the importance of employing screening tools using conventional instruments, albeit with the trade-off of potentially less accurate data quality. The challenges of conventional screening require a paradigm shift in approach. To address these limitations, integrating artificial intelligence (AI) into healthcare systems holds great promise (Rajpurkar et al., 2022). Integrating AI-based technology into healthcare systems offers a transformative approach to oral cancer screening (Lee et al., 2021). By leveraging advanced image analysis and AI, specifically Deep Learning (DL), hospitals can support and augment the work of their staff,

* Corresponding author.
E-mail address: marco.parola@ing.unipi.it (M. Parola).
URL: https://mlpi.ing.unipi.it/m.parola/,github.com/MarcoParola (M. Parola).

https://doi.org/10.1016/j.compmedimag.2024.102433
Received 7 January 2024; Received in revised form 20 June 2024; Accepted 31 August 2024
Available online 11 September 2024

improving the accuracy and speed of diagnostic processes. Although integrating DL-based technology into oral cancer screening is a promising frontier, it also presents some drawbacks. Two main challenges deserve attention: limitations in data collection and explainability problems.

Regarding data collection, choosing appropriate devices for data collection is a critical consideration in oral cancer screening. Traditional devices, such as smartphones, are inexpensive and accessible, making them suitable for mass screening efforts. However, these tools introduce a significant amount of noise into the data, resulting from factors such as reflections. This noise can potentially compromise the accuracy of the screening process. On the other hand, expensive and accurate machines, such as specialized scanners, offer high-quality information that is less operator-dependent. Although these machines are adept at minimizing noise, their high cost and limited accessibility make them less practical for widespread screening initiatives. Moreover, the dependence on labeled data for training DL models poses another challenge. The manual task of labeling data is susceptible to imperfections and subjectivity. Large datasets with accurate annotations are essential for the effective training of AI algorithms, which is a practical obstacle to the implementation of these systems.

Regarding the DL explainability issue, these models are referred to as "black boxes" as they do not provide details on how they reach a particular conclusion or prediction (Kim and Kim, 2020). This opacity can be problematic in healthcare, where understanding the reasoning behind a decision is as important as the decision itself. To address the lack of explainability and transparency of NNs, a solution lies in hybrid approaches integrating DL with case-based reasoning (CBR) called DL-CBR systems[1] (Keane and Kenny, 2019). CBR is an AI paradigm that solves new problems by referring to similar past cases. Its main advantages include its ability to enable reasoning from a limited number of examples and match past cases to facilitate problem-solving with structured solutions, leading to another advantage: CBR is interpretable; presenting retrieved cases is effective for justifying CBR model decisions. Finally, because it includes retrieval, similarity concepts, and case adaption information, CBR offers ways to incorporate preexisting knowledge into the reasoning process (Leake et al., 2023).

However, the potential of CBR to integrate human knowledge into these systems is often untapped or underutilized. The current trend predominantly emphasizes leveraging DL for feature extraction, frequently overlooking the opportunity to enhance interpretability through the incorporation of domain-specific information or presenting an inconsistency between the case base and the knowledge base (Leake et al., 2023; Weber et al., 2018; Chourib et al., 2020; Tjoa and Guan, 2021); despite, several works have underscored the critical importance of human knowledge integration within these systems (Raj, 2023; Lieber et al., 2018).

To overcome these additional difficulties, Information Deep Learning (IDL) is a viable solution (Von Rueden et al., 2021) as they use DL techniques to provide NNs with better performance by incorporating preexisting information into the architecture, resulting in better generalization of the model and more aligned with human reasoning.

Faced with (i) the limitations of data quality collection, (ii) the DL explainability issue, and (iii) CBR's underexploited potential of including human knowledge, we propose a unified framework based on IDL and CBR able to solve lesion detection and classification problems and provide visual explanations by integrating medical knowledge in the form of similarity. The strength of this framework lies primarily in the well-known DL-CBR hybrid system, which articulates its decisions through visual examples, combining the robustness of DL with case-based reasoning. This example-based learning approach facilitates interpretability by demonstrating the logic behind the model results (Bouzar-Benlabiod et al., 2023). When faced with a diagnostic task, the system refers to visually rich cases that best encapsulate the pathology, similar to how a medical professional might draw on experience. Then, we further enhance this system with IDL, marking a novel advancement, as IDL has never been used to infuse similarity information between cases into the CBR before. Guiding the NN learning process with established medical knowledge enables visual explanations to converge with the insight and understanding of medical professionals (Oberste and Heinzl, 2023).

The main contributions of this work are: (i) Designing a screening system that provides human-centered explanations, departing from traditional developer-centered applications to enhance understanding in the medical domain. By adopting IDL in combination with CBR, we introduce a novel IDL-CBR framework, addressing the previously untapped integration of human knowledge in CBR systems. Such framework demonstrates the capability to incorporate physician-driven examples, contributing to bridging the gap between AI and human expertise in medical reasoning. (ii) Ensuring the robustness of our model against different imperfections, due to a no-standardized acquisition process or labeling error. By effectively addressing challenges inherent in noisy images and artifacts, the solution demonstrates commendable performance, particularly in scenarios involving conventional instruments, thus eliminating the need for expensive and human-dependent scanners. (iii) Conducting extensive benchmark experiments on detection and classification tasks by adopting state-of-the-art DL architectures on our dataset. Finally, (iv) promoting research collaboration, as we have collected and labeled a dataset that we release publicly, addressing the scarcity of comprehensive public datasets in oral cancer and promoting an open data approach.

The paper is organized as follows. Section 2 covers the literature review and resumes some core concepts for a better work understanding. The workflow is detailed in Section 3 by defining the different tasks it can solve and how the different DL architectures are trained and organized in the final screening system, while the case study and experiment results are discussed in Section 4 and Section 5, respectively. Section 6 discusses the work by analyzing the research's strengths and weaknesses. Finally, Section 7 concludes and outlines future research possibilities.

2. Literature review and background

In this subsection, we provide an overview of some main concepts for a comprehensive understanding of the research presented in this work. Such concepts serve as the foundation on which our study is based, facilitating a deeper understanding of the methodologies, results, and implications discussed throughout the article.

2.1. Case-based reasoning

Case-based reasoning (CBR) is a well-known versatile problem-solving paradigm composed of 4 phases, as shown in Fig. 1: (i) Retrieval, (ii) Reuse, (iii) Revise, and (iv) Retain.

The first phase is *Retrieval* (*Ret*), in which the CBR system searches in the base of n cases the subset of k ones more relevant to the current problem. This relevance is determined by a similarity function $sim(c_i, c_j)$, such as cosine or euclidean distance, which measures the similarity between the features of the new problem c_i and those of stored cases c_j.

$$Ret(c_i) = \{c_j \mid \arg\_\max\ sim(c_i, c_j), \quad j = 1, \ldots, n\} \tag{1}$$

The second stage is *Reuse*, in which the subset of selected cases is examined to identify potential strategies that can be applied to the new problem. The system evaluates which components of the retrieved cases are applicable to the current problem and can be reused directly or with minor adaptations. The next stage is *Revise*, which focuses on adapting the solutions of the selected cases to fit the current problem. This stage often requires understanding the differences and similarities between

[1] Also called ANN-CBR: artificial neural networks (ANNs) with case-based reasoning (CBR).

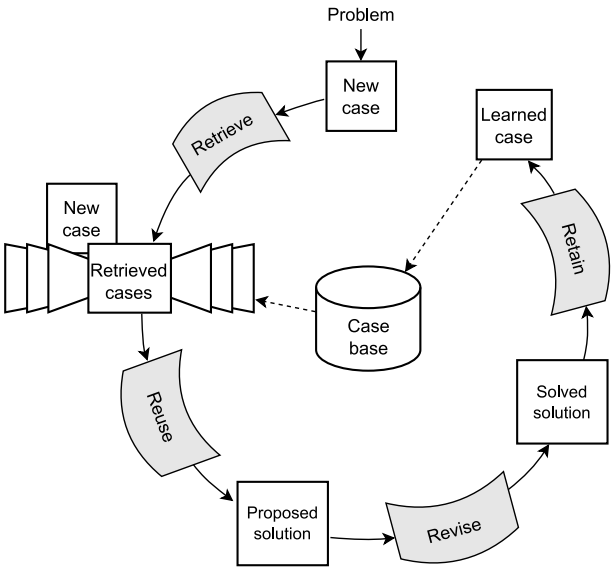


Fig. 1. Case-based reasoning cycle composed of four phases: Retrieval, Reuse, Revise, and Retain.

the retrieved cases and the new problem. The last stage is called *Retain* phase, during which the case base is updated with the modified solution to the present issue and any fresh information. This guarantees the system's knowledge repository's expansion and flexibility over time by constantly updating and improving it.

With encouraging results, the CBR paradigm has been widely used for medical imaging (Bouzar-Benlabiod et al., 2023; Marie et al., 2020; Gao and Gao, 2021). However, there is a paucity of research in academia exploiting this approach for medical image analysis in the oral cancer context. The existing literature contains a single work on CBR in the context of oral health care (Ehtesham et al., 2019), it works on tabular records, which require manual entry of injury-related descriptors by physicians to facilitate retrieval of similar cases. This approach differs from ours, as an automatic feature extraction from images through DL techniques. However, efforts have been invested in the use of CBR in the medical imaging domain, although not specifically in the oral cancer domain. Examples include studies related to cervical cancer detection using MRI (Neves et al., 2018), breast cancer research using tabular data (Gu et al., 2020), and mammographic imaging investigations in the context of breast cancer detection (Barnett et al., 2021).

2.2. Deep learning and imperfect data

Another approach we want to review is DL, which has proven to be the most effective solution for tackling complex computer vision tasks such as image classification, object detection, and segmentation. DL relies on neural networks (NN) to learn how to perform these tasks.

The success of DL models in image-related tasks largely depends on the quality and integrity of the input data. However, real-world image datasets are often subject to imperfections that can have an impact on the performance of DL models. These imperfections can occur in different forms, such as inaccuracies in labels and the presence of artifacts (Xu et al., 2023). A common source of imperfections in image datasets is related to inaccuracies in labeling. Human annotators may introduce errors due to inconsistencies in labeling guidelines or simple oversights (Karimi et al., 2020). Another common form of imperfection in image data is the presence of artifacts that can arise from various sources, including distortions introduced during acquisition using non-standard protocols or multiple capture devices (Varoquaux and Cheplygina, 2022).

According to Karimi et al. (2020) taxonomy, to mitigate the impact of imperfections on performance, three main strategies have been proposed:

- Methods focusing on robust model selection against imperfections.
- Methods aiming to reduce label noise in the dataset.
- Methods that perform model training and modeling of label noise in a unified framework.

In this work, we adopt an approach based on model robustness, introducing an ensemble architecture into the workflow to handle labeling artifacts and imperfections.

Additionally, we review some studies exploring this approach in the oral cavity domain between 2021 and 2023, encompassing both detection and classification tasks. A common thread in most of these studies is the use of in-house photographic image sets to conduct the experiments. Specifically, these datasets are collected by the clinical subgroup of the research team. The scarcity of research using publicly available datasets is partly attributed to the difficulty of finding such resources online. The resulting literature consists of works that have conducted experiments on datasets composed of 300–1500 images.

In Welikala et al. (2020) multiple classification tasks were addressed by adopting the pretrained ResNet-101 architecture, including (i) binary classification problems distinguishing *lesion* from *non-lesion* with an F1 of 87.07; (ii) *referral* from *non-referral* with an F1 of 78.30, as well as (iii) a multiclass classification problem combining the previous problem labels, achieving an overall F1 of 50.57. Additionally, the study employed Faster R-CNN for lesion detection, achieving an F1 of 41.18. The dataset used in this research was collected from the MeMoSA project, comprising 1433 images.

Another study, Warin et al. (2022), focused on the binary classification of "potentially malignant oral disorders" and "normal oral mucosa". DenseNet-121 and ResNet-50 models were employed, obtaining an F1 of 95 for classification, while detection tasks were handled by Faster R-CNN and YOLOv4, obtaining F1 scores of 80.31 and 52.38, respectively. The authors conducted the experiments on a dataset comprising 600 images.

Another interesting study is Bansal et al. (2022), one of the few cases in which public photographic datasets were adopted. They faced a binary classification problem, distinguishing between *Cancer* and *Non-cancer* labels. To address this challenge, several DL models were compared, including ResNet50, MobileNetV2, VGG19, VGG16, and DenseNet. DenseNet was the best-performing model, with an accuracy of 99.38%. It is worth noting that the dataset used in this research was obtained from the public platform Kaggle[2] and although not large (Barot, 2020). The dataset included 131 images, of which 44 were associated with *non-cancer* cases and 87 with *cancer* cases.

2.3. DL-CBR system and Informed Deep Learning

Despite their remarkable performance, DL models sometimes are not considered suitable for the healthcare application domain as they are black boxes, and they suffer from the explainability issue. The XAI focuses on developing techniques and methods that shed light on the inner workings of black-box models, enabling users to understand the motivations behind model predictions.

Mark T. Keane and Eoin M. Kenny made significant contributions to XAI by conducting extensive research in the integration of CBR and DL as a framework for providing explanations in AI known as "ANN-CBR twins" (Keane and Kenny, 2019). This framework provides a coherent method for post-hoc explanation-by-example and/or counterfactual explanations, allowing for transparent insight into the decision-making

[2] https://www.kaggle.com

processes of complex NNs applied to different domain applications (Dai et al., 2022; Kenny et al., 2023; Delaney et al., 2021).

Numerous studies on DL and CBR, also referred to as Content-Based Image Retrieval (CBIR), have been proposed in the literature, with a major emphasis on all aspects of refining the DL training process. Barata and Santiago (2021) proposed a DL model for skin cancer diagnosis that provides explainability through CBIR. They explored several state-of-the-art approaches to improve the feature space learned by the NN based on contrastive, distillation, and triplet losses. In Allegretti et al. (2021), the authors proposed a framework in which ResNet is originally trained to classify dermoscopic images; then, the feature extraction module is decoupled for generating image embeddings which allows the images to be checked for similarity. In Bouzar-Benlabiod et al. (2023), the authors use a DL-CBR system for classification from a mammographic image. Since the accuracy of CBR depends on the quality of the extracted features, an image enhancement process is proposed to improve the quality of the extracted features and provide a better final diagnosis via a segmentation method, based on the U-Net architecture.

Although all these studies have made significant progress in combining DL and CBR and found important user benefits (Patrício et al., 2023), they mainly focus on DL methods to create a suitable feature space for CBR. Therefore, the human knowledge and feedback in these processes have not been fully addressed and exploited. This was a major starting point; from here we investigated the possibility of introducing human knowledge within these decision support systems.

Under this premise, Informed Machine Learning (IML) represents a paradigm that integrates domain expertise into intelligent systems, thereby enhancing the XAI scenario (Oberste et al., 2023). As highlighted in the survey (Von Rueden et al., 2021), *"[...], the essence of informed machine learning is that this prior knowledge is explicitly integrated into the machine learning pipeline, ideally via clear interfaces defined by the knowledge representations"*. Indeed, IML serves as a bridge between the vast knowledge gained by domain experts and the capabilities of data-driven machine learning models. The authors aimed to explore the integration of prior knowledge into the machine learning pipeline, articulating the research into three key aspects:

- what is the integrated source of knowledge;
- how is the knowledge expressed;
- where is it embedded in the learning process.

3. Methodology

In this section, we delve into the design of the proposed oral cavity screening health system relying on IDL and CBR. The methodology is organized around a comprehensive workflow consisting of two distinct phases, as shown in Fig. 2: an offline phase and an online one.

The first part, referred to as the offline phase, focuses on training several DL architectures, each of which has been tailored to address specific tasks relevant to oral cave screening: lesion detection and relevant features extraction are addressed with a pure DL approach. The pathology classification and the corresponding visual explanation through examples is the innovative aspect addressed with an IDL strategy. The experiments for training and evaluating the architectures addressing these tasks are described in detail in Sections 3.2, 3.3 and 3.4, respectively. Further insights into the specific data augmentation techniques employed for each experiment will be discussed in Section 5.

In the second part of our workflow, the online phase, we orchestrate the integration of previously trained DL models into a unified pipeline, thus building the healthcare system designed to assist healthcare operators and physicians. This phase transforms the knowledge gained during the offline phase into practical, real-time decision support for clinicians engaged in oral cavity screening.

3.1. CBR system design

The CBR system development by medical experts starts with the collection of a data set and continues with the construction of the system itself. The process begins with the selection of an initial image that is added to the reference cases, beginning the reference case base of the CBR system. After the initial case is added, clinicians undertake an iterative loop to refine the CBR system in which the entire dataset is scanned image by image. The process consists of the following steps: (i) a new image is selected from the dataset and compared with the reference cases. (ii) Comparison between cases: the objective is to assess the similarity between the target image and the reference cases. The comparison is made according to the diagnostic criteria and characteristics of the image. (iii) Problem resolution: If the CBR system successfully supports identifying a match between the current image and the reference cases, it means that the current reference cases are descriptive and correctly model the data scanned to this stage. In this case, the system moves on to the next image in the dataset. (iv) Reconfiguration of reference cases: In cases where the CBR system is unable to produce a solution, the reference cases are reconfigured. This involves adding the target image to the base of the reference cases, expanding their knowledge. In addition, some redundant or less informative reference cases can be removed to maintain the relevance and efficiency of the case base.

The data collection phase will be described in detail in Section 4, however, it is important to point out that the labeling phase also involves providing meta-information describing the similarity between the images according to the base cases of the CBR system; since starting from this information we are able to implement our IDL-CBR system and, then, evaluate its explainability.

The dataset collection phase encompassed the acquisition of similarity information data. This was accomplished through a ranking approach, wherein we considered a set of n reference cases denoted as R and a set of m target cases denoted as T. For each target case $j \in T$, a distinct subset R'_j of l reference cases was selectively chosen from the available pool of R. In this process, the size of l was not necessarily uniform, varying from case to case. Subsequently, a ranking was assigned to each reference case within the selected subset, ranging from 1 to l depending on its degree of similarity to the current target case. This methodology yielded a sparse matrix of dimensions n by m, where each cell could either remain empty or contain a numeric value between 1 and l denoting the position of the reference case within the rank for the respective target case as shown in Eq. (2). This data structure can be represented as shown in Eq. (3). On the other hand, creating a full matrix instead of a sparse one could make it inconsistent with the real world. The manual task of defining the ranking introduces an element of subjectivity and may be subject to human error, as the similarity criteria may vary depending on the physician performing the task.

$$\mathrm{Rank}[i,j] = \begin{cases} rank(i,j) & \text{if reference case } i \in R'_j \subset R \\ null & otherwise \end{cases} \tag{2}$$

$$Rank = \begin{bmatrix} 2 & 3 & null & 1 & l\ldots & null \\ 1 & null & null & l-1 & \ldots & 3 \\ null & null & 2 & 3 & \ldots & l \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ l-1 & 1 & 3 & null & \ldots & null \end{bmatrix} \tag{3}$$

3.2. Lesion detection

In the initial phase of the research, we undertook a comprehensive exploration of state-of-the-art DL object detection models to identify the most effective approach for our oral cavity lesion detection task.

Identifying which part of the oral cavity is affected by a lesion can be approached as a supervised learning task, specifically, object detection. Let us consider a set of labeled images, referred to as D, in

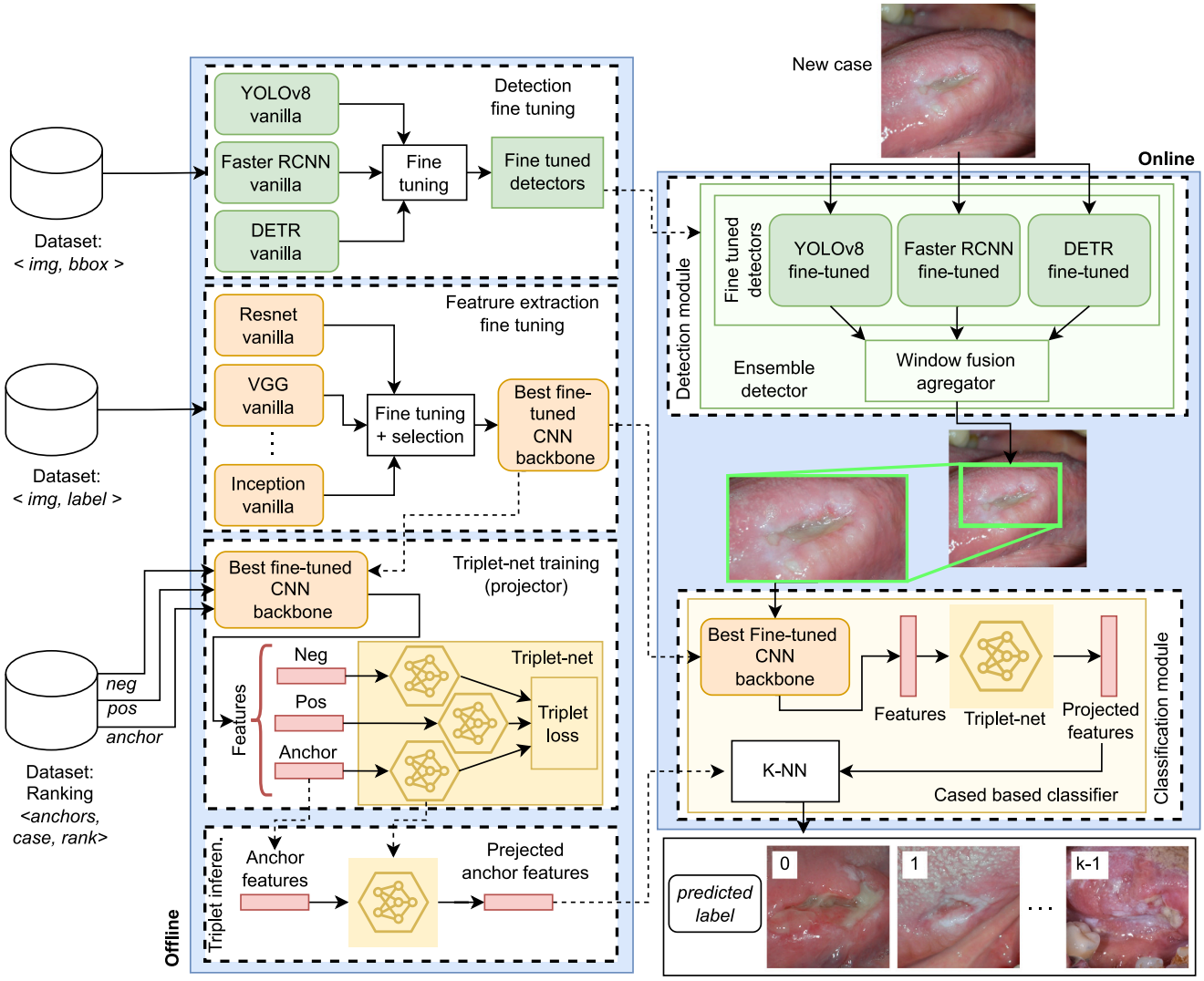


Fig. 2. Overview of the DL architecture training process and model deployment for implementing the oral cancer screening system.

which each image I is paired with a set of bounding boxes denoted as B_I and their corresponding class labels C_I. These annotations were provided by a domain expert. Each bounding box B_i is characterized by four parameters: (x, y, w, h), where (x, y) signifies the top-left corner's coordinates, while (w, h) indicates the width and height of the bounding box, respectively. Additionally, the class label c_i indicates the bounded object class.

The goal is to define a mapping function $f_\theta(I)$ to make accurate predictions for both bounding boxes and class labels of objects in test images. θ identifies trainable parameters. The learning process trains the model parameters θ by optimizing the loss function $\mathcal{L}$ over the dataset $\mathcal{D}$ as shown in Eq. (4). Specifically, two distinct weighted contributions compose $\mathcal{L}$: (i) the localization loss $\mathcal{L}_{loc}$ evaluating the bounding box identification, weighted by λ_{loc}; and (ii) the classification loss $\mathcal{L}_{\text{cls}}$ measuring the classification performance weighted by λ_{cls}. The exact formulations of $\mathcal{L}_{loc}$ and $\mathcal{L}_{cls}$ depend on the experiments' design choices. Typical implementations involve *meansquarederror* loss for localization as well as using *cross entropy* loss for classification. The set $\mathcal{B}_{\text{pos}}$ regards the collection of indices corresponding to positive samples. Additionally, B_i and c_i refer to the forecasted bounding box and class label for an object, while B_i* and c_i* denote the actual bounding box and actual class label for the same object.

$$\mathcal{L}(f_\theta(I), B_I, L_I) = \lambda_{\text{loc}} \sum_{i \in \mathcal{B}_{\text{pos}}} \mathcal{L}_{\text{loc}}(B_I, B_I*) + \lambda_{\text{cls}} \sum_{i \in \mathcal{B}_{\text{pos}}} \mathcal{L}_{\text{cls}}(c_I, c_I*) \tag{4}$$

Following our previous pilot study (Parola et al., 2023), the initial DL experiment involved a tuning phase of three well-known models, You Only Look Once version 8 (YOLOv8), Detection Transformer (DETR), and Faster R-CNN.

Next, we introduce an ensemble architecture at the beginning of the DL workflow as ensemble strategy can be an effective solution to handle artifacts and annotation errors (Karimi et al., 2020). By placing this architecture at the beginning of the workflow, we ensure that data imperfection problems are handled at the early stage.

The ensemble model combines the strengths of the three individual models relying on the window fusion submodule to aggregate predictions done by single models. Specifically, let B_{ij} denote the jth bounding box predicted by model i and let $\mathcal{B} = \{B_{Y1}, \ldots, B_{Yn}, B_{F1}, \ldots, B_{Fm}, B_{D1}, \ldots, B_{Dl}\}$ be the set of bounding boxes detected by three distinct models for a given input image. The ensemble model generates a set of predicted bounding boxes, denoted as $\mathcal{B}_E = \{B_{E1}, \ldots, B_{Ep}\}$, by

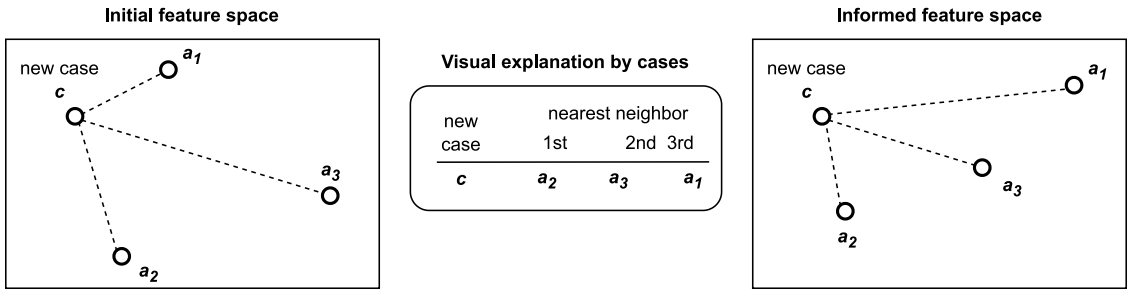


Fig. 3. On the left, four data points in the initial feature space, comprised of one target case c and three reference cases a_i. On the right, their positions in the new informed feature space after shifting according to the nearest-neighbor relationships provided by the experts.

assessing the overlaps among these distinct boxes.[3] More precisely, each bounding box $B_{E\alpha}$ is derived from a subset $\hat{\mathcal{B}} \subseteq \mathcal{B}$ through the IoU operation, as shown in Eq. (5); here, $\hat{\mathcal{B}}$ encompasses all elements satisfying $\forall B_\beta, B_\gamma \in \mathcal{B}, IoU(B_\beta, B_\gamma) > th, \gamma \neq \beta$.

Furthermore, to establish a robust ensemble model, we impose the constraint $\exists\ B_{ij}, B_{hk} \mid i \neq h; i, h \in M$, indicating that the ensemble bounding box is computed based on at least two bounding boxes predicted by distinct models. Consequently, if a lesion is identified solely by one model, the ensemble categorizes it as a false positive.

$$B_{E\alpha} = \cap_i B_i \quad \forall B_i \in \hat{\mathcal{B}};\ \alpha = 1, \ldots, p \tag{5}$$

The metrics below were introduced to compare various object detection models (Padilla et al., 2020). The Jaccard coefficient-based measure known as Intersection over Union (IoU) calculates the amount of overlap between the predicted and actual bounding boxes. It is determined by dividing the intersection's area by the union's area between the two bounding boxes.

To evaluate the object detection predictions, for each detected object in an image, the IoU is computed with each ground truth bounding box. A detection is considered a true positive (TP) if the IoU with any ground truth box archives a threshold th; otherwise, it is a false positive (FP). Ground truth boxes not matched by any detection are considered false negatives (FN); allowing Precision (Prec) to be computed as TP/TP+FN and Recall to be computed as TP/TP+FN. Thus, Average Precision (AP) can be calculated as the area under the Prec against Recall curve at different confidence thresholds IoU. Finally, the mean Average Precision (mAP) can be employed to assess an object detection task's performance by computing the average AP among the different classes at a th value of 50% or by averaging th values between 50% and 95%, as shown in Eqs. (6) and (7), respectively.

$$mAP@50 = \frac{1}{N}\sum_{i=1}^{N} AP@50_i \tag{6}$$

$$mAP@95-50 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{9} AP@(50+5j)_i \tag{7}$$

3.3. Feature extraction fine tuning

The second experimental phase was devoted to fine-tuning the pre-trained convolutional models on our dataset. This effort sought to exploit the potential of these models by adapting the feature domain to our dataset. To ensure a complete and unbiased evaluation, we maintained uniformity by employing an identical dense NN structure for all convolutional architectures evaluated. This approach allowed us to isolate performance differences arising solely from variations in convolutional feature extraction patterns. As part of this comparative analysis, we examined the effectiveness and adaptability of four distinct families of pre-trained convolutional architectures: RegNet, ResNet, ShuffleNet, and GoogLeNet.

The performance of the models is evaluated both with the global accuracy metric over the entire dataset to get an overview, and by evaluating the accuracy per class of these architectures by obtaining a finer granularity of performance on the classification task. This evaluation aims to measure the ability of each model to distinguish between specific classes of lesions, thus providing insight into their effectiveness in identifying lesions with different characteristics, considering that lesions can have different severities.

3.4. Explainable classification

Finally, we present the details of the third and final experiment conducted as part of our research to enhance the Classification by the CBR system. This experiment focuses on addressing the crucial need for explainable diagnosis, ensuring that the system not only classifies new cases accurately but also provides an intelligible human-centered explanation.

This explanation is expressed as a rank of closeness (similarity) among the still-active cases in the case base. It is inspired by the idea that when presented with a new medical case, doctors often rely on their experience and knowledge to identify previously encountered cases that are most similar to the current one. By ranking these similar cases in terms of their relevance and similarity, the system provides an explanation that aligns with how a human expert might approach the problem.

According to Von Rueden et al. (2021) taxonomy, our IDL solution to integrate human case reasoning knowledge within the system is based on: *Expert Knowledge* as the source of information to be integrated into the system; a *Kwoledge Graph* approach as the representation; and *Training Data* as the method of knowledge integration. The knowledge graph is a similarity graph built by the medical staff of our team. In this representation, each node corresponds to an oral lesion, while edges capture the similarity relationships between these lesions. The edges are weighted to quantify the degree of similarity between cases.

The final goal of the IDL module is to manipulate the feature space in a way that respects the rank of the most similar cases provided by the domain expert. This process falls under the *Training Data* of taxonomy as it has been implemented via a triple loss function. Triplet loss has been extensively employed in both DL and DL-CBR systems. However, its use has been mainly limited to a data-driven similarity learning approach (Tang et al., 2023; Schuler et al., 2023; Tjoa and Guan, 2021), meaning processing the data to create a good representation in feature space for a final downstream task. In contrast, our proposed work pioneers the usage of triplet loss for integrating human similarity information into a CBR system, bridging the gap of the untapped potential DL-CBR of incorporating domain knowledge of the current methodologies (see Figs. 3 and 4).

Triplet loss is a loss function stimulating the model to learn embeddings (representations) such that the distance between an anchor sample and a positive sample is less than the distance between the

[3] $Y = YOLO, F = FasterRCNN, D = DETR, E = Ensemble.$

anchor sample and a negative sample. Eq. (8) presents the triplet loss $\mathcal{L}(A, P, N)$ formula:

$$\mathcal{L}(A, P, N) = \max(0, d(A, P) - d(A, N) + \alpha) \tag{8}$$

where A represents the anchor sample; P represents the *positive* sample (i.e., a sample similar to the anchor), while N represents the *negative* sample (i.e., a sample dissimilar to the anchor); $d(.,.)$ is a distance metric. Finally, α is the margin, a hyperparameter ensuring the positive sample is closer to the anchor than the negative sample by at least this margin.

In triplet net training experiments with triplet loss, triplet samples are derived from the sparse rank matrix described in Section 3.1, and the approach considers all possible triplets within this matrix. Given a matrix row, as many triplets are generated by setting the target case relative to the current row as A and for each possible pair of the `non-null` elements, the two elements result P and N according to their relative position in the rank.

This approach provides a visual explanation by presenting sets of reference cases sorted by similarity with respect to the target case. We adopt the k nearest neighbor (kNN) algorithm to retrieve them. To measure the effectiveness of this approach, we propose an explainability performance evaluation focused on measuring the alignment between the similarity ranking produced by our classification module and that assigned by expert clinicians, which serves as ground truth (Kumar and Vassilvitskii, 2010). For this purpose, we rely on established a metric from the literature for measuring similarity between rankings to quantitatively assess the fidelity of our visual explainability solution: Spearman Footrule and Kendall Tau distances. The metric values range in the interval $[0, 1]$, where 0 denotes identical ranks and 1 signifies opposed ranks.

The Spearman Footrule distance ϕ is a metric used to quantify the dissimilarity between two rankings ρ_1 and ρ_2, by computing the sum of the absolute differences between the positions of each item in the two rankings. Eq. (9) show the Spearman Footrule distance ϕ formula, where $\rho_j(i)$ represents the position of item i in ranking j, $j = 1, 2$ and n is the total number of items being ranked.

Kendall Tau distance τ quantifies the number of pairwise disagreements between item orderings in two rankings as shown in Eq. (10); where C is the number of concordant pairs, which are pairs of items that have the same order in both ρ_1 and ρ_2; while D is the number of discordant pairs.

$$\phi(\rho_1, \rho_2) = \sum_{i=1}^{n} |\rho_1(i) - \rho_2(i)| \tag{9}$$

$$\tau(\rho_1, \rho_2) = 2\frac{C - D}{n(n-1)} \tag{10}$$

4. Oral case study

Images of the oral cavity were collected between 2021 and 2023 from people during medical consultations at the P. Giaccone University Hospital in Palermo, Italy, which houses the Oral Medicine Unit. Dental hygienists, consultants, and trainees practicing oral medicine took the images using both smartphone cameras or regular cameras to save on expensive and advanced imaging equipment. We collected images of three distinct pathologies. Table 1 shows the occurrences of collected images by class. After collecting the images, the CBR base-case construction process described in Section 3.1 was performed, at the end of which we obtained a base-case reference consisting of 30 cases.

However, images acquired without capture process standardization and standard instruments, as in our case study, often introduce noise due to inherent variations in their characteristics. These variations occur as disparities in illumination conditions, spatial resolutions, color spaces, and sensor characteristics.

An additional critical aspect contributing to the dataset imperfections was the image annotation phase. We have utilized the COCO `Annotator` tool to annotate the photographs. A trained dental team manually annotated the lesions in the images. Each lesion was annotated with a boundary segment and a corresponding label; the boundary rectangle was generated by the tool from the segment, by selecting among all the rectangles containing the segment the one with the minimum area.

Table 1
Data set occurrences.

Class labels	Short	Occurences
All		567
Aphthous	ap	198
Neoplastic	ne	170
Traumatic	tr	199

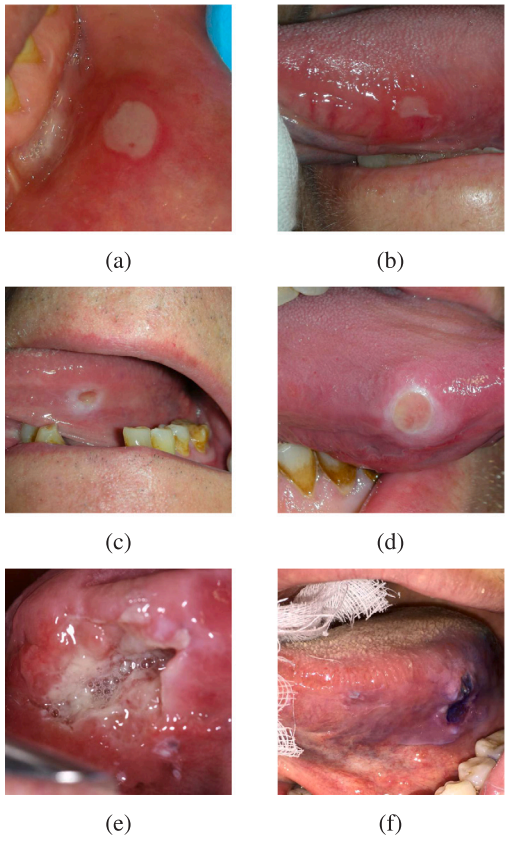


Fig. 4. Six photographic images examples of our dataset: (a) and (b) aphthous; (c) and (d) traumatic; (e) and (f) neoplastic.

However, as shown in Fig. 5, analysis of bounding box distribution among the different classes reveals non-uniformity. Specifically, the dimensions (height and width) of bounding boxes belonging to the "neoplastic" class differ from those of the other two classes. This discrepancy in bounding box size suggests a distinct level of meticulousness in the photo-taking process that covers much of the image for the "neoplastic" class and from which the larger bounding boxes are derived. The background portion of an image can affect performance (Cha et al., 2021).

As anticipated, the data collection phase also included a CBR construction phase, at the end of which a case base of 31 reference cases was created as follows: 9 aphthous, 12 neoplastic, and 10 traumatic injuries. Next, 375 rankings were generated to quantify the similarity between 375 instances of oral lesions and the 31 reference cases. Such rankings were determined by health experts, who evaluated between 20 and 10 cases per instance, rather than all 31, resulting in a sparse matrix. Finally, this matrix served the derivation of 30284 triplets,

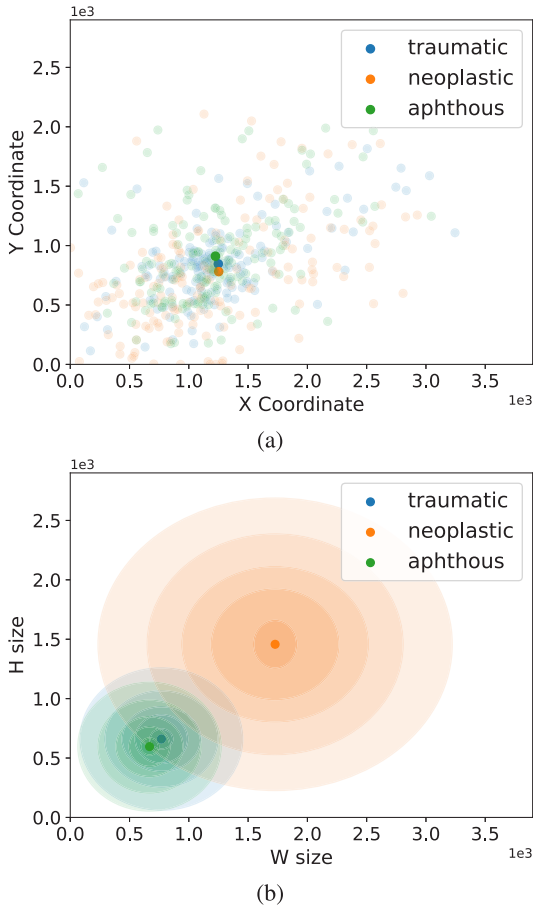


Fig. 5. Distribution of bounding box encoding (X,Y,W,H). In (a) the scatter plot of left corner bbox (X,Y) per class. In (b) the distribution of bounding box dimensions (W,H) using concentric ellipses; one ellipse per percentile [5, 25, 50, 75, 90].

Table 2
Comparison of map metrics by classes among the YOLOv8, Faster R-CNN, and DETR models.

Model	Class	MAP@50	MAP@50-95
YOLOv8	ne	.498	.203
	ap	.345	.152
	tr	.613	.217
	ALL	.469	.191
FasterRCNN	ne	.725	.289
	ap	.279	.107
	tr	.394	.155
	ALL	.466	.183
DETR	ne	.639	.299
	ap	.306	.114
	tr	.472	.173
	ALL	.473	.195

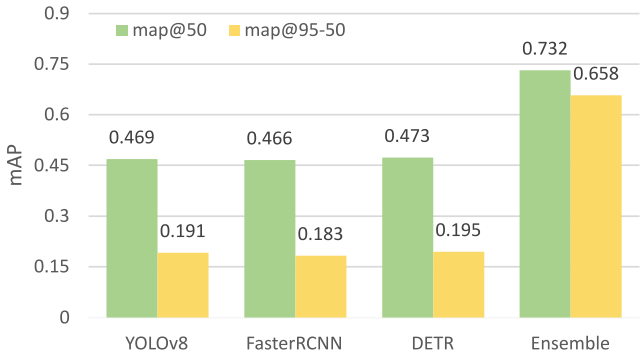


Fig. 6. Comparison of mAP@50 and mAP@95-50 metrics between YOLOv8, DETR, FasterRCNN, and ensemble on the test set.

according to Eqs. (2) and (3), in order to train a projection network using the loss of triplets.

5. Experiments and results

The experiments in this study were carried out using the `PyTorch Lightning` framework on a Linux-based machine equipped with an RTX GPU. To foster transparency of our findings and facilitate experiment replication, we have made the source code publicly available at Parola (2023).

5.1. Lesion detection

In this section we present experiments related to injury detection tasks; in particular, we present a benchmark comparison between the three detection architectures described in the previous section.

In the data augmentation phase, we apply a series of transformations, by employing the `imgaug` library: (i) Brightness and Contrast Adjustment: We randomly adjust the brightness (95%–105% of the original) and contrast (95%–105% of the original) to simulate varying lighting conditions. (ii) Hue and Saturation Modification: Random changes in hue and saturation (−10 to 10) are introduced to add color variability to the images. (iii) Horizontal Flipping: For horizontal variance, we apply a 50% chance of horizontal flipping to the images, effectively doubling the dataset. (iv) Affine Transformations: To introduce spatial variability, we perform affine transformations. This includes random translations within a range of −10% to 10% in both the x and y directions, random rotations (−10 to 10 degrees), and non-uniform scaling (0.9 to 1.1) in the x and y dimensions while maintaining the aspect ratio. The filling method adopted after combining transformations is *nearest neighbor*.

The training of each model was run for 100 epochs, assuming an early stopping condition monitoring the loss function with a patience value of 10 epochs. For each architecture, we tested different learning rates (lr) from the set: $[5e^{-4}, 1e^{-5}, 5e^{-5}, 5e^{-6}, 1e^{-6}]$. Training, validation, and testing datasets were randomly generated from the entire dataset. To ensure consistency of results, we performed multiple experiments using different partition generations. The final split is the one publicly released and described in Data availability section. YOLO's best results were achieved with the early stop condition trigger at epoch 61 and a lr value of $1e^{-5}$. FasterRCNN performed best with a lr value of $1e^{-5}$ at epoch 55. Finally, at epoch 89, the early stop condition was met with a lr value of $1e^{-6}$, giving the best results for DETR.

Table 2 shows the map@50 and map@95-50 results obtained by each model in the survey task, both on the entire dataset and for individual classes. Fig. 6 shows the performance between the single architectures and the ensemble model.

Fig. 7 illustrates various scenarios encountered during the lesion detection task, highlighting both the successes and challenges associated with predictive accuracy. In the first case (a), all four models correctly predicted the presence of the lesion, illustrating a desirable outcome. However, it is essential to acknowledge that such success is not always guaranteed, as we will demonstrate. In case (b), while two out of three models detected the lesion, the ensemble model accurately assigned the correct label due to average voting weights, emphasizing the strength of combining multiple models. In contrast, case (c) introduces a common issue of artifacts, where a flash reflection was misinterpreted as an additional aphthous lesion by the DETR model. In case (d), a case of human-introduced noise in the annotation stage can be observed, as

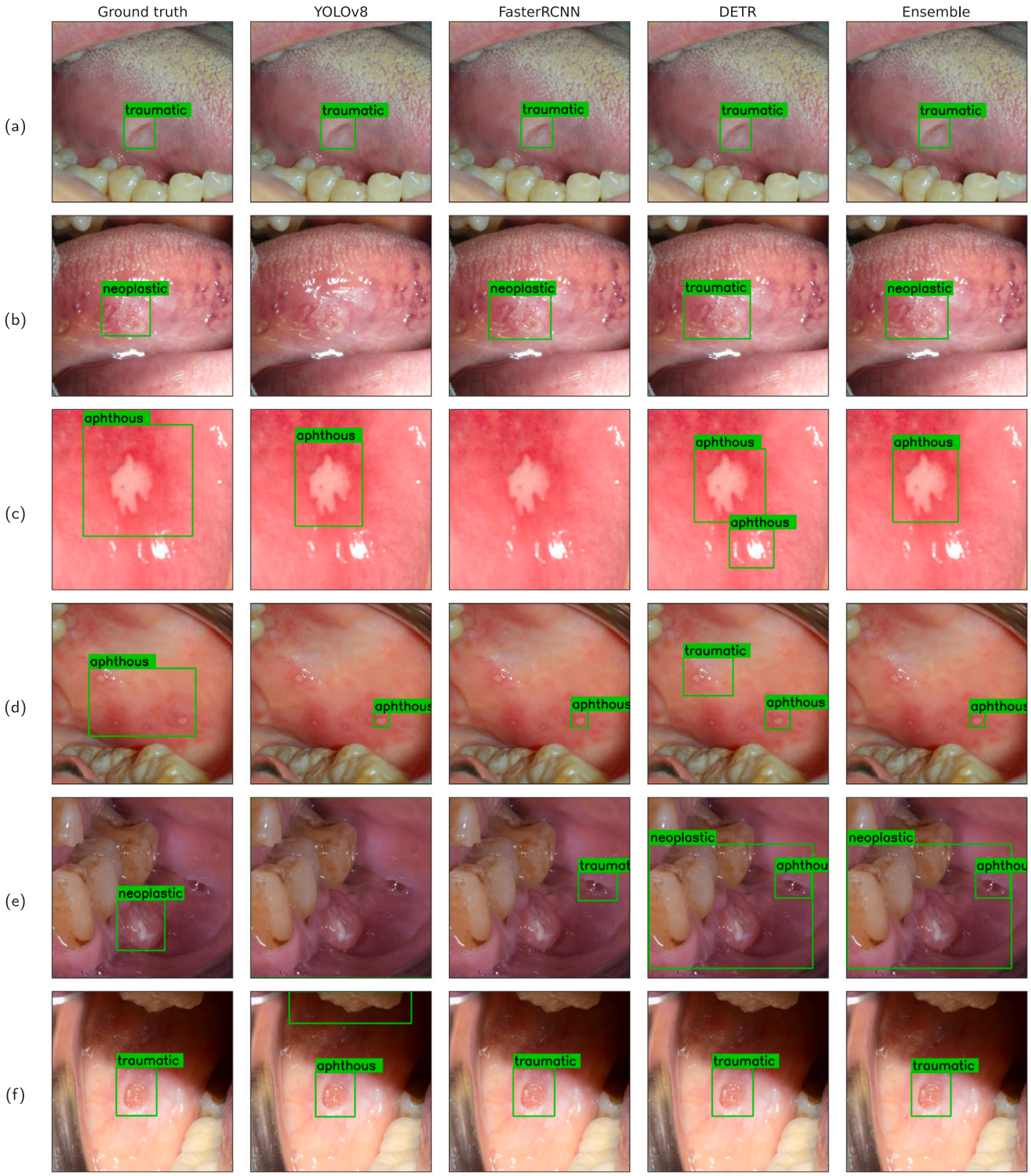


Fig. 7. Six examples of bounding box prediction for individual architectures and the ensemble model versus the ground truth.

a multi-aphthous lesion case was encapsulated in a single bounding box, however, the models still demonstrated lesion detection capability. Case (e) exemplifies the impact of noise, with saliva bubbles leading to misclassifications by two models (FasterRCNN and DETR), resulting in the prediction of the ensemble model as well. Lastly, case (f) brings out the non-standardization of image acquisition processes, as a gauze was mistaken for a lesion by the YOLO model.

5.2. Feature extraction fine tuning

In this section, we present a comparison of several pre-trained convolutional architectures available within the `PyTorch` library, which have been fine-tuned to adjust them to the oral cavity context.

To demonstrate the effectiveness of ensemble model detection in mitigating bounding box size annotation bias, we conducted image classification experiments under two distinct scenarios: using whole

Table 3
Comparison of pre-trained convolutional models fine-tuned on an oral cancer dataset. The table presents performance metrics for two scenarios: one ✓ in which input images are cropped to bounding boxes and another ✗ in which whole images are fed to the models.

Model			Performance			
Family	vers	crop	Acc	Acc_{ne}	Acc_{ap}	Acc_{tr}
RegNet	x-1–6	✓	**.80**	.96	.81	.62
	y-8		.45	.20	.50	.62
	y-1–6		.64	.92	.35	.67
	x-1–6	✗	.76	.95	.76	.60
	y-8		.40	.22	.41	.56
	y-1–6		.61	.88	.35	.61
ResNet	50	✓	.68	.96	.46	.62
	34		**.72**	.92	.81	.42
	18		.64	.86	.65	.39
	50	✗	.66	.95	.43	.60
	34		.71	.92	.79	.41
	18		.63	.86	.62	.38
ShufNet	x0–5	✓	.60	.90	.71	.18
	x1–0		.63	.91	.80	.13
	x2–0		.71	.96	.58	.58
	x0–5	✗	.58	.89	.65	.19
	x1–0		.64	.90	.78	.15
	x2–0		.67	.91	.54	.59
GoogleNet	–	✓	.65	.96	.77	.21
	–	✗	.67	.98	.76	.25

images and using images cropped to the bounding boxes, resulting in zoomed images on the lesion and a consequent reduction of the annotation bias.

For the sake of fairness, the fine-tuning process was performed by concatenating the same fully connected dense NN at the end of each pre-trained model. The fully connected NN performing classification consists of (i) a hidden layer of 64 neurons with `ReLu` activation function, and (ii) an output layer of three neurons. Moreover, a dropout layer with a rate of 0.5 is interposed between them. The training, validation, and test sets align with the experimental design detailed in the preceding section. Additionally, an early stopping criterion was employed to prevent model overfitting. The overall and per-class accuracy performances between the various models are shown in Table 3.

The two pre-trained convolutional architectures, both trained in the cropped scenario, that achieved the highest accuracy performance are `ResNet34` and `RegNet-x-1-6`. Therefore, these models will be adopted in the last experiment shown in Section 5.3 to implement the overall system.

5.3. Projection training

In the last experiment, we focus on a comparative analysis of the overall workflows implemented through both DL and IDL approaches to understand how effective the IDL approach is in projecting human knowledge-based features by measuring the similarity between the ranks provided by the clinicians and the system output ranks.

The pure DL workflow consists of a fine-tuned convolutional model as feature extractor on top of which the kNN algorithm is run to retrieve the most similar cases. The workflow implemented using IDL consists of a fine-tuned convolutional model, a projector network trained using the triplet-loss, and the kNN algorithm.

In Table 4 we present the comparison between these two approaches. For each approach, we considered the two convolutional networks that achieved the highest performance in the first experiment. The triple network introduced in the IDL approach consists of: a linear layer with 64 neurons, Gelu, Dropout layer, linear layer with 64 neurons. The training process of this network was significantly faster compared to the previous two experiments due to the lower complexity

Table 4
Comparison of the DL and IDL implementations of the screening system.

Method	CNN	Classification	Similarity	
		Acc.	ϕ	τ
DL	Resnet34	.701	.491	.618
	Regnet-X16GF	.777	.480	.604
IDL	Resnet34	.812	.399	.483
	Regnet-X16GF	**.854**	**.386**	**.470**

of the data, which consisted of 64 neurons' feature vectors. The lowest loss value was achieved after 16 epochs with a lr value of $1e^{-6}$. The parameter k for the kNN algorithm was set to 5, following an evaluation of various values between $[3, 5, 7, 9]$.

`RegNet-x-1-6` serves as the foundation for the two workflow implementations that produced the highest accuracy results both for DL and IDL strategies.

Concerning the similarity between the cases retrieved by the two different strategies and those provided by the doctor, Resnet is the model that best models this similarity when implementing the IDL strategy, achieving a value of $\phi = 38,6\%$ and a value of $\tau = 47,0\%$.

6. Discussion

To develop an oral cancer screening system working on photographic images, we proposed a solution relying on the CBR paradigm, which provides visual output explanations, and Informed Deep Learning IDL, which integrates medical domain expert prior knowledge into the system. The entire procedure was divided into three main experiments.

Concerning the object detection task, our ensemble model achieved a mAP@50 value of 0.732 and a mAP@95-50 value of 0.658, emphasizing the ensemble's ability to maintain high accuracy even at different thresholds. In particular, the strength of our ensemble lies in its ability to combine the different forecasts of several models by proposing the intersection of these and discarding forecasts that do not match other bounding boxes. This mitigates the impact of false positives, resulting in the removal of artifacts in the system pipeline. Furthermore, the model performs successfully despite the bias in the bounding box annotations made by the physicians. This underlines the importance of exploiting aggregation methods to strengthen the robustness of the object detection module.

During the benchmark phase aimed at selecting the best pre-trained convolutional architecture, we set up experiments under two different scenarios – using both whole images and images cropped to the bounding boxes – as part of a fine-tuning process. Our experimental results indicate that performing image classification on cropped images generally provides comparable or better performance than using whole images. This improvement can be observed for all the architectures under investigation (except GoogleNet, where we observed a worsening in accuracy) demonstrating the effectiveness of reducing annotation bias through the use of zoomed-in images on the lesion.

At the end of this phase, our evaluation process led us to select `RegNet-x-1-6` and `ResNet34` trained on the cropped setting as the most effective, capable of achieving an accuracy metric equal to 80% and 72%, respectively.

In addition, observing the accuracy values obtained on the individual classes, it is interesting to note for all models that the class that obtained the highest values was *neoplastic* associated with cancer. This is a very important result as it is precisely this class that one wants to recognize with the highest accuracy.

In the third experiment, we compared a pure DL strategy with one based on IDL to implement an oral cancer screening system. The comparison was conducted considering both accuracy in solving a classification problem and validating the visual explanation by introducing two metrics to assess rank similarity: ϕ and τ.

We observed that the accuracy performance of DL architectures in the first experiment is similar to the kNN one performed on the features generated without the human knowledge projection. Furthermore, we demonstrated that both the accuracy and human reasoning similarity of the IDL strategy incorporating a NN to feed human knowledge is more effective than the pure DL-based implementation. Indeed, IDL achieved an accuracy value of 85.4% compared to 77.7% for the DL strategy (using RegNet_X_1_6GF as the backbone). With a ϕ value of 38.6% and a τ value of 47.0%, the IDL also performed better than the DL approach in terms of visual explainability. The DL obtained values of 48.0% and 60.4%, respectively.

It is crucial to emphasize how validating the visual explainability of the system requires human experience in the form of rankings provided by domain experts. This poses two significant challenges: firstly, assembling this information requires considerable effort; secondly, without validation by multiple domain experts, the data may be susceptible to errors and subjective judgments, thus compromising the accuracy of performance measurements.

7. Conclusions

Our research has addressed the need for efficient and cost-effective computerized screening systems in the medical domain, emphasizing the importance of transparency in their design. We recognize that while existing XAI primarily caters to developers' needs for model improvement, our IDL approach integrates medical domain experts' knowledge, providing relevant insights for clinical consumers.

The integration of IDL and CBR in our healthcare screening framework represents a significant step forward in generating human-centered predictions that align with clinical users' requests and are valuable for medical decision-making. To support diagnosis without requiring expensive scanners or high-quality images, our solution shows the potential of DL to work with noisy images containing artifacts from low-cost equipment by introducing an ensemble model for object detection at the beginning of the workflow eliminating false predictions.

The contribution of this work lies in two main areas. First, we designed a screening system that prioritizes human-centered explanations, moving away from traditional developer-centered applications. This change improves understanding in the medical domain, as our DL architectures assimilate physician-driven examples, bridging the gap between artificial intelligence and human experience in medical reasoning. Secondly, the robustness of our model against various imperfections, resulting from non-standardized acquisition processes or labeling errors, has been demonstrated. By addressing the challenges related to noisy images and artifacts, our solution achieved commendable results, especially in scenarios involving conventional instruments, thus eliminating the reliance on expensive and human-dependent scanners.

CRediT authorship contribution statement

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

Data availability

Publicly available datasets are provided in this study. Oral-AI dataset available at Dii and University of Pisa (2023).

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

During the preparation of this work, the authors used OpenAI ChatGPT 3.5 service in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

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).

References

Allegretti, S., Bolelli, F., Pollastri, F., Longhitano, S., Pellacani, G., Grana, C., 2021. Supporting skin lesion diagnosis with content-based image retrieval. In: 2020 25th International Conference on Pattern Recognition. ICPR, pp. 8053–8060. http://dx.doi.org/10.1109/ICPR48806.2021.9412419.

Bansal, K., Bathla, R., Kumar, Y., 2022. Deep transfer learning techniques with hybrid optimization in early prediction and diagnosis of different types of oral cancer. Soft Comput. 26 (21), 11153–11184.

Barata, C., Santiago, C., 2021. Improving the explainability of skin cancer diagnosis using CBIR. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2021, Springer International Publishing, Cham, pp. 550–559.

Barnett, A.J., et al., 2021. Interpretable mammographic image classification using case-based reasoning and deep learning. arXiv preprint arXiv:2107.05605.

Barot, S., 2020. Oral cancer (lips and tongue) images, kaggle.com. www.kaggle.com/datasets/shivam17299/oral-cancer-lips-and-tongue-images.

Bouzar-Benlabiod, L., Harrar, K., Yamoun, L., Khodja, M.Y., Akhloufi, M.A., 2023. A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification. Comput. Biol. Med. 163, 107133.

Bramati, C., Abati, S., Bondi, S., Lissoni, A., Arrigoni, G., Filipello, F., Trimarchi, M., 2021. Early diagnosis of oral squamous cell carcinoma may ensure better prognosis: A case series. Clin. Case Rep. 9 (10).

Bugshan, A., Farooq, I., 2020. Oral squamous cell carcinoma: metastasis, potentially associated malignant disorders, etiology and recent advancements in diagnosis. F1000Research 9.

Cha, K., Woo, H.-K., Park, D., Chang, D.K., Kang, M., et al., 2021. Effects of background colors, flashes, and exposure values on the accuracy of a smartphone-based pill recognition system using a deep convolutional neural network: Deep learning and experimental approach. JMIR Med. Inf. 9 (7), e26000.

Chourib, I., Guillard, G., Mestiri, M., Solaiman, B., Farah, I.R., 2020. Case-based reasoning: Problems and importance of similarity measure. In: 2020 5th International Conference on Advanced Technologies for Signal and Image Processing. ATSIP, pp. 1–6. http://dx.doi.org/10.1109/ATSIP49331.2020.9231755.

Dai, X., Keane, M.T., Shalloo, L., Ruelle, E., Byrne, R.M., 2022. Counterfactual explanations for prediction and diagnosis in XAI. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. pp. 215–226.

Delaney, E., Greene, D., Keane, M.T., 2021. Instance-based counterfactual explanations for time series classification. In: International Conference on Case-Based Reasoning. Springer, pp. 32–47.

Dii, University of Pisa, 2023. Oral-AI dataset. URL: https://131.114.50.176/owncloud/s/B42Mhg5UtjxKuuQ/download.

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 Dis. 25 (6), 1555–1563.

Gao, X.W., Gao, A., 2021. COVID-CBR: a deep learning architecture featuring case-based reasoning for classification of COVID-19 from chest X-ray images. In: 2021 20th IEEE International Conference on Machine Learning and Applications. ICMLA, IEEE, pp. 1319–1324.

Gu, D., Su, K., Zhao, H., 2020. A case-based ensemble learning system for explainable breast cancer recurrence prediction. Artif. Intell. Med. 107, 101858.

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Med. Image Anal. 65, 101759.

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: Case-Based Reasoning Research and Development: 27th International Conference, ICCBR 2019, Otzenhausen, Germany, September 8–12, 2019, Proceedings 27. Springer, pp. 155–171.

Kenny, E.M., Delaney, E., Keane, M.T., 2023. Advancing post-hoc case-based explanation with feature highlighting. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. IJCAI-23.

Kim, Y.-J., Kim, J.H., 2020. Increasing incidence and improving survival of oral tongue squamous cell carcinoma. Sci. Rep. 10 (1), 7877.

Kumar, R., Vassilvitskii, S., 2010. Generalized distances between rankings. In: Proceedings of the 19th International Conference on World Wide Web. WWW '10, Association for Computing Machinery, New York, NY, USA, pp. 571–580.

Leake, D., Wilkerson, Z., Crandall, D.J., 2023. Combining case-based reasoning with deep learning: Context and ongoing case feature learning research. In: Neuro-Symbolic Learning and Reasoning in the Era of Large Language Models.

Lee, C.H., Zhang, Z., Zhao, X., 2021. A survey of smart healthcare for the elderly based on user requirements and supply accessibility. In: 5th International Conference on Crowd Science and Engineering. pp. 108–112.

Lieber, J., Nauer, E., Prade, H., Richard, G., 2018. Making the best of cases by approximation, interpolation and extrapolation. In: Case-Based Reasoning Research and Development: 26th International Conference, ICCBR 2018, Stockholm, Sweden, July 9–12, 2018, Proceedings 26. Springer, pp. 580–596.

Marie, F., Henriet, J., Lapayre, J.-C., 2020. A new adaptation phase for thresholds in a CBR system associated to a region growing algorithm to segment tumoral kidneys. In: Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28. Springer, pp. 97–111.

Neves, J., Vicente, H., Ferraz, F., Leite, A.C., Rodrigues, A.R., Cruz, M., Machado, J., Neves, J., Sampaio, L., 2018. A deep learning approach to case based reasoning to the evaluation and diagnosis of cervical carcinoma. In: Modern Approaches for Intelligent Information and Database Systems. Springer, pp. 185–197.

Oberste, L., Heinzl, A., 2023. User-centric explainability in healthcare: A knowledge-level perspective of informed machine learning. IEEE Trans. Artif. Intell. 4 (4), 840–857.

Oberste, L., Rüffer, F., Aydingül, O., Rink, J., Heinzl, A., 2023. Designing user-centric explanations for medical imaging with informed machine learning. In: International Conference on Design Science Research in Information Systems and Technology. Springer, pp. 470–484.

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, IEEE, pp. 237–242.

Parola, M., 2023. GitHub oral case based reasoning code repository. URL: https://github.com/MarcoParola/oral2.

Parola, M., Mantia, G.L., Galatolo, F., Cimino, M.G., Campisi, G., Di Fede, O., 2023. Image-based screening of oral cancer via deep ensemble architecture. In: 2023 IEEE Symposium Series on Computational Intelligence. SSCI, pp. 1572–1578.

Patrício, C., Neves, J.C., Teixeira, L.F., 2023. Explainable deep learning methods in medical image classification: A survey. ACM Comput. Surv. 56 (4).

Raj, K., 2023. A neuro-symbolic approach to enhance interpretability of graph neural network through the integration of external knowledge. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. CIKM '23, Association for Computing Machinery, New York, NY, USA, pp. 5177–5180.

Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J., 2022. AI in health and medicine. Nat. Med. 28 (1), 31–38.

Schuler, N., Hoffmann, M., Beise, H.-P., Bergmann, R., 2023. Semi-supervised similarity learning in process-oriented case-based reasoning. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence. Springer, pp. 159–173.

Tang, Z., Sun, Z.-H., Wu, E.Q., Wei, C.-F., Ming, D., Chen, S.-D., 2023. MRCG: A MRI retrieval framework with convolutional and graph neural networks for secure and private IoMT. IEEE J. Biomed. Health Inf. 27 (2), 814–822.

Tjoa, E., Guan, C., 2021. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans. Neural Netw. Learn. Syst. 32 (11), 4793–4813.

Varoquaux, G., Cheplygina, V., 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit. Med. 5 (1), 48.

Von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., et al., 2021. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans. Knowl. Data Eng. 35 (1), 614–633.

Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S., Jantana, P., 2022. 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), 699–704.

Weber, R.O., Johs, A.J., Li, J., Huang, K., 2018. Investigating textual case-based XAI. In: Case-Based Reasoning Research and Development: 26th International Conference, ICCBR 2018, Stockholm, Sweden, July 9–12, 2018, Proceedings 26. Springer, pp. 431–447.

Welikala, R.A., Remagnino, P., Lim, J.H., Chan, C.S., Rajendran, S., Kallarakkal, T.G., Zain, R.B., Jayasinghe, R.D., Rimal, J., Kerr, A.R., Amtha, R., Patil, K., Tilakaratne, W.M., Gibson, J., Cheong, S.C., 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.

Xu, M., Kim, H., Yang, J., Fuentes, A., Meng, Y., Yoon, S., Kim, T., Park, D.S., 2023. Embracing limited and imperfect data: A review on plant stress recognition using deep learning. arXiv preprint arXiv:2305.11533.

Zhou, J.-Y., Wang, W.-J., Zhang, C.-Y., Ling, Y.-Y., Hong, X.-J., Su, Q., Li, W.-G., Mao, Z.-W., Cheng, B., Tan, C.-P., et al., 2022. Ru (II)-modified TiO2 nanoparticles for hypoxia-adaptive photo-immunotherapy of oral squamous cell carcinoma. Biomaterials 289, 121757.