Interpretable Machine Learning for Oral Lesion Diagnosis Through Prototypical Instances Identification

Alessio Cascione[1], Mattia Setzu[1], Federico A. Galatolo[1], Mario G.C.A. Cimino[1], and Riccardo Guidotti[1,2](✉)

[1] University of Pisa, Largo Bruno Pontecorvo 3, Pisa, PI 56127, Italy
a.cascione@studenti.unipi.it,
{mattia.setzu,federico.galatolo,mario.cimino,riccardo.guidotti}@unipi.it
[2] KDD Lab, ISTI-CNR, Via G. Moruzzi 1, Pisa, PI 56124, Italy
riccardo.guidotti@isti.cnr.it

Abstract. Decision-making processes in healthcare can be highly complex and challenging. Machine Learning tools offer significant potential to assist in these processes. However, many current methodologies rely on complex models that are not easily interpretable by experts. This underscores the need to develop interpretable models that can provide meaningful support in clinical decision-making. When approaching such tasks, humans typically compare the situation at hand to a few key examples and representative cases imprinted in their memory. Using an approach which selects such exemplary cases and grounds its predictions on them could contribute to obtaining high-performing interpretable solutions to such problems. To this end, we evaluate PivotTree, an interpretable prototype selection model, on an oral lesion detection problem. We demonstrate the efficacy of using such method in terms of performance and offer a qualitative and quantitative comparison between exemplary cases and ground-truth prototypes selected by experts.

Keywords: Interpretable Machine Learning · Explainable AI · Instance-based Approach · Pivotal Instances · Transparent Model · Dental Health AI · Oral Disease Prediction

1 Introduction

One of the sectors that has significantly benefited from the application of Machine Learning (ML) tools is healthcare [8,21]. However, although the models employed to solve diagnostic tasks are powerful in terms of predictive capability, their reliance on complex architectures often makes it difficult for experts and users to understand their reasoning. Moreover, the "cognitive process" employed by these models is frequently not comparable to how humans reason to solve the same tasks [49]. Given the pivotal role of these tools as decision-support systems for practitioners in healthcare, explaining and interpreting their predictions has become crucial and is the focus of active research in Explainable AI (XAI) [2].

D. Pedreschi et al. (Eds.): DS 2024, LNAI 15244, pp. 316–331, 2025.
https://doi.org/10.1007/978-3-031-78980-9_20

As humans, our cognitive processes and mental models frequently depend on case-based reasoning [37], where past exemplary cases are stored in memory and retrieved to solve specific tasks. This type of reasoning is so deeply embedded in us that even young children can recognize and interact with unfamiliar objects they have never encountered before, provided these objects resemble something they already know [42]. Moreover, this ability extends across various modalities: we identify authors by their writing style, recognize relatives by shared facial features, and classify music genres based on similarities to familiar tracks [23]. In the healthcare sector, practitioners frequently diagnose or identify new conditions by referencing past case reports [20,38]. Additionally, the experiments detailed in [11] demonstrated that pattern recognition, grounded in examples gained through experience, is the diagnostic strategy with the highest likelihood of success.

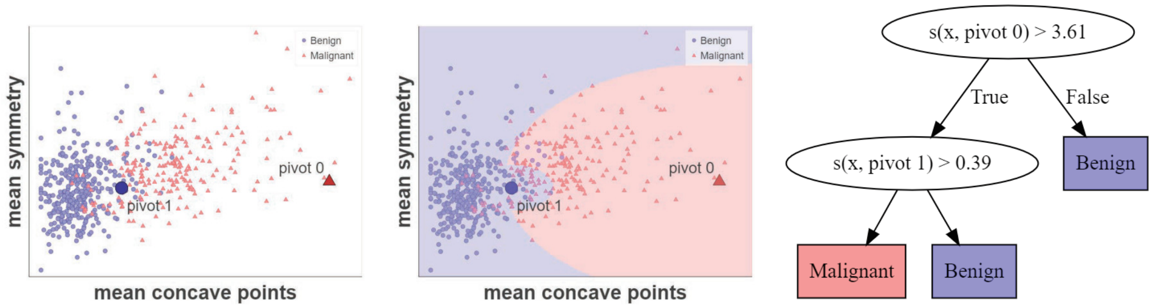
Given these premises, a promising approach to designing inherently interpretable ML models for the healthcare sector is to explore the intuitive notion of similarity between *discriminative* and *descriptive* instances. The underlying assumption is that grounding a model's predictions on the similarity between test instances and exemplar cases would yield a naturally interpretable and trustworthy tool for medical experts and end-users alike. In this paper, we present a case study with an interpretable similarity-based model for decision-making applied to a specific medical context, i.e., for an oral lesion prediction task.

In particular, we study PivotTree [7], a hierarchical and interpretable case-based model inspired by Decision Tree (DT) [6]. By design, PivotTree can be used both as a *prediction* and *selection* model. As a selection model, PivotTree identifies a set of training exemplary cases named *pivots*; as a predictive model, PivotTree leverages the identified pivots to build a similarity-based DT, routing instances through its structure and yielding a prediction, and an associated explanation. Unlike traditional DTs, the resulting explanation is not a set of rules having features as conditions, but rules using a set of pivots to which the instance to predict is compared. Like distance-based models, PivotTree allows to select exemplary instances in order to encode the data in a similarity space that enables case-based reasoning. Finally, PivotTree is a *data-agnostic* model, which can be applied to different data modalities, jointly solving both pivot selection and prediction tasks. Given its modality agnosticism, PivotTree represents an advancement over traditional DTs. As shown in [7], the case-based model learned by PivotTree offers interpretability even in domains like images, text, and time series, where conventional interpretable models often underperform and lack clarity. Furthermore, unlike conventional distance-based predictive models such as k-Nearest Neighbors (kNN) [15], PivotTree introduces a hierarchical structure to guide similarity-based predictions.

Figure 1 provides an example of PivotTree on the `breast cancer` dataset[1], wherein cell nuclei are classified according to their characteristics computed from a digitized image of a fine needle aspirate of a breast mass. Starting from a dataset of instances, PivotTree identifies a set of two pivots (Fig. 1 *(a)*) in this

[1] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.

case belonging to the two distinct classes *Benign* and *Malignant*. Said *pivots* are used to learn a case-based model wherein novel instances are represented in terms of their similarity to the induced pivots (Fig. 1 *(b)*). Building on pivot selection, PivotTree then learns a hierarchy of pivots wherein instances are classified. This hierarchy takes the form of a Decision Tree (Fig. 1 *(c)*): novel instances navigate the tree, percolating towards pivots to which they are more similar or dissimilar, and landing into a classification leaf. In the example, given a test instance x: if its similarity to *pivot 0* is lower than 3.61 (following the right branch), then x is classified as a *Benign*, i.e., x is far away from the *Malignant pivot 0* (see Fig. 1 *(b)*). Instead, following the left branch, if x's similarity to *pivot 1* is higher than 0.39 (left branch), then x is still classified as *Benign* as it is very similar to the *Benign pivot 1*, otherwise x is classified as *Malignant* as it is sufficiently similar to the *Malignant pivot 0*. In contrast, a traditional Decision Tree would model the decision boundary with feature-based rules, e.g., "if *mean concave points* $<$ 2.4 then *Benign* else if *mean symmetry* $<$ 1.7 then *Malignant*". However, traditional DTs *(i)* can only model axis-parallel splits, and *(ii)* cannot be employed on data types with features without clear semantics such as medical images. Hence, improving on traditional DTs, the case-based model learned by PivotTree can provide interpretability even in domains such as images, text, and time series, by exploiting a suitable data transformation.



(a) Select (b) Predict (c) Explain

Fig. 1. PivotTree as *(a)* selector, *(b)* interpretable model, *(c)* Decision Tree.

In this paper we demonstrate that PivotTree represents a promisingly effective approach for *interpretability of oral lesion detection*, and we compare its selected pivots with instances identified as representative by domain experts. After an initial review of the literature concerning XAI in the healthcare sector, and prototype-based approach for explainability in Sect. 2, in Sect. 3 we summarize the PivotTree method. Then, in Sect. 4 we report the experimental results on the oral lesion diagnostic problem. Finally, Sect. 5 completes our contribution and discusses future research directions.

2 Related Work

The wide use of explainability techniques for the medical field has been extensively reviewed in previous work [3,16]. ML [12], and specifically case-based reasoning, already finds application in the medical domain, where interpretable and uninterpretable models [5,10] already tackle a variety of tasks, including breast cancer prediction [28,41,48], melanoma detection [17,31–33] and Covid-19 detection [39]. The latter, in particular, introduces two-level interpretations: prototypes are also defined *contrastively*, i.e., both highly similar and highly dissimilar prototypes are provided, and they are also accompanied by heatmaps indicating regions of higher importance. These approaches integrate the discovery of prototypes directly into the model, which often uses similarity-based scoring function to perform predictions. Other examples include [44], which combines knowledge distillation with heterogeneous prototype selection for mammograms, building on [9], and [24] which leverage prototype learning for Autism spectrum disorder detection from fMRI images. In [25] besides prototypes criticism are also identified, i.e., instances representatives of some parts of the input space where prototypical examples do not provide good explanations.

Focusing on oral cancer detection, a relevant example is [46], which proposes an end-to-end, two-stage model for oral lesion detection and classification. This model leverages YOLOv5l [22] for detection and EfficientNet-B4 [43] for classification, making it suitable for deployment as a mobile application. In [27], the authors fine-tune a Single Shot Multibox Detector (SSD) [29] to identify the presence and location of oral disease. Finally, in [50] a self-supervised pretraining strategy is defined, followed by a semi-supervised learning approach on epithelial regions for carcinoma detection. A case-based approach specifically for oral lesion is offered in [13], which works with tabular descriptors by physicians. More at large, and aside from case-based interpretations, interpretability in the medical sector has been gaining attention for quite some years [34]. However, all the aforementioned works offer black-box models for oral lesion detection and classification. On the other hand, in terms of interpretability tools for oral cancer detection, only a handful of proposals are currently in place. In [1] an ensemble approach for oral cancer prediction using tabular data, which by design relays on SHAP values [4] for explainability, is discussed. In [14] an approach using gradient-weighted class activation mapping is presented and [40] provides visual explanations leveraging attention mechanisms also adding expert knowledge by incorporating manually edited attention maps in order to update classification results. Differently from the literature presented so far, to the best of our knowledge, our study is the first inquiring on explainability through prototypes for the oral lesion detection problem using a data-agnostic model.

3 Pivot Tree in a Nutshell

In this section, we present the main characteristics of PivotTree. For more detailed information and extensive benchmarking, we refer readers to [7].

Given a set of n instances represented as real-valued m-dimensional feature vectors[2] in $\mathbb{R}^m$, and a set of class labels $C = \{1, \ldots, c\}$, in case-based reasoning, the objective is to learn a function $f : \mathbb{R}^m \rightarrow C$ approximating the underlying classification function, with f being defined as a function of k exemplary cases named *pivots*. Similarity-based case-based models define f on a similarity space $\mathcal{S}$, often inversely denoted as "distance space", induced by a similarity function $s : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ quantifying the similarity of instances [36]. Given a training set $\langle X, Y \rangle$, and a similarity function s, our objective is to learn a function $\pi : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{k \times m}$ that selects a set $P \subseteq X$ of k pivots maximizing the performance of f. The instances in X are mapped into Z through $\mathcal{S}$, wherein they are represented in terms of their similarity to the pivots P. f is then trained on $\langle Z, Y \rangle$, and at inference time, instances are first mapped through the similarity space, before being fed to f.

Aiming for transparency of the case-based predictive model f, our objective is to employ as an interpretable model f Decision Tree classifiers (DT) or k-Nearest Neighbors approaches [18] (kNN). When f is implemented with a DT, split conditions will be of the form $s(x, p_i) \geq \beta$, i.e., "if the similarity between instance x and pivot p_i is greater or equal then β, then ...", allowing to easily understand the logic condition by inspecting x and p_i for every condition in the rule. On the other hand, when f is implemented as a kNN, every decision will be based on the similarity with a few neighbors derived from the pivot set P. A human user just needs to inspect x and the similarities with the pivots P and the instances in the neighborhood. When the number of pivots is kept small, the interpretability of both methods increases, limiting the expressiveness. Vice versa, using a selection model π that returns a large number k of pivots can increase the performance at the cost of interpretability. Our proposal aims to balance these two aspects by allowing the selection of a small number of pivots that still guarantee comparable performance to interpretable predictive models.

PivotTree implements the selection function π, and leverages existing interpretable models to implement f. Much like Decision Tree induction algorithms [6], PivotTree greedily learns a hierarchy of nodes wherein pivots lie. Node splits are selected so that the downstream performance of f is maximized, i.e., the split is chosen to maximize the information gain of the node. Notably, PivotTree does not operate directly on the data, but rather on the induced similarities, thus the split is chosen among a set of candidates defining lower or higher similarities to a set of candidate pivots: the traditional "$x_i \leq \alpha$" split is replaced by a similarity rule of the form $s(x_i, p_i) \leq \alpha$ thresholding the similarity of instances to pivots. The training data is then routed according to the split, and the operation repeats recursively. Pivots come in two families: *discriminative*, which guide instance routing, and *descriptive*, which instead describe the node. The former are selected to maximize the performance, while the latter are selected to maximize similarity to the other instances percolating the node.

[2] For the sake of simplicity, we consistently treat data instances as real-valued vectors. Any data transformation employed in the experimental section to maintain coherence with this assumption will be specified when needed.

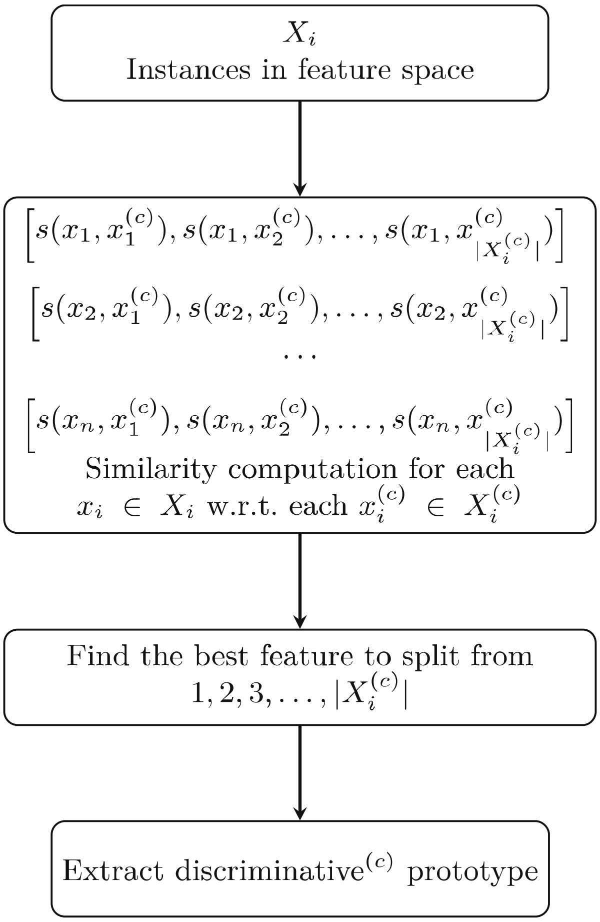

X_i
Instances in feature space
$\left[s(x_1, x_1^{(c)}), s(x_1, x_2^{(c)}), \ldots, s(x_1, x_{|X_i^{(c)}|}^{(c)})\right]$
$\left[s(x_2, x_1^{(c)}), s(x_2, x_2^{(c)}), \ldots, s(x_2, x_{|X_i^{(c)}|}^{(c)})\right]$
$\cdots$
$\left[s(x_n, x_1^{(c)}), s(x_n, x_2^{(c)}), \ldots, s(x_n, x_{|X_i^{(c)}|}^{(c)})\right]$
Similarity computation for each
$x_i \in X_i$ w.r.t. each $x_i^{(c)} \in X_i^{(c)}$
Find the best feature to split from
$1, 2, 3, \ldots, |X_i^{(c)}|$
Extract discriminative$^{(c)}$ prototype

Fig. 2. PivotTree workflow for selection of discriminative pivots of class c for a non-terminal node with X_i training instances. With $x^{(c)}$ we indicate an instance of class c and in analogous way with *discriminative*$^{(c)}$ the fact we are referring to a class c pivot.

Figure 2 displays the selection process for the *discriminative* pivots of class c in a node: choosing the best splitting feature in the similarity space implies finding the c-instance which best separates the current training data when instance similarity is taken into account, i.e., the *discriminative* pivot of such class.

In a sense, the *descriptive* and *discriminative* pivots extracted by PivotTree can be associated with the prototypical examples and criticisms identified by [25]. However, their usage is markedly different. By design, PivotTree is a data-agnostic model that leverages the concept of similarity to conduct both selection and prediction tasks simultaneously. While some data types, e.g., relational data, are more amenable than others, e.g., images or text, to similarity computation, with our contribution, we aim to address all data types as one. By decoupling similarity computation and object representation, PivotTree can be applied to any data type supporting a mapping to $\mathbb{R}^m$, i.e., text through language model embedding, images through vision models, graphs through graph representation models, etc. In the healthcare sector, this approach can be highly beneficial due to the heterogeneous nature of the data types involved in diagnostic processes: sequential data like EEG/ECG signals, text-based clinical reports, and medical images of lesions can all be processed using a unified PivotTree framework by transforming each data type into an appropriate vector representation. This integrated approach can improve diagnostic explainability by allowing for a comprehensive analysis of multimodal healthcare data. In the following experiments, we focus specifically on images, particularly on oral lesion images, using embeddings provided by a pre-trained deep learning model.

4 Experiments

In this section, we evaluate the performance of PivotTree[3] (PTC) on the DoctOral-AI dataset[4]. Our objective is to demonstrate that PivotTree is an accurate predictor and selector tool for the task and show how comparable the learned pivots are to ground-truth cases deemed prototypical by expert doctors.

Classification Models. We refer to PivotTree used as Classification model with PTC. We use P to denote the set of pivots identified by PivotTree, and O to denote the set of ground-truth prototypes. DT_P and kNN_P refer to DT and kNN models, respectively, trained in the similarity space obtained by computing the similarity between each instance and every pivot in P. Similarly, DT_O and kNN_O are trained in the similarity space derived from the ground-truth prototypes in O. As further baselines, we compare PivotTree with kNN and DT directly trained on feature space. Finally, as deep learning (DL) model we rely on the Detectron2 (D2) model [47] fine-tuned on the DoctOral-AI dataset. We report the performance of D2 to observe the loss in accuracy at the cost of interpretability. A comparison on DoctOral-AI w.r.t other DL architectures is also offered in [35].

Experimental Setting. We evaluated the predictive performance of the aforementioned models by measuring Balanced Accuracy and F1-score, Precision and Recall by computing the metric for each label and reporting the unweighted mean. In line with [7], for PivotTree hyperparameter selection[5], both as a predictor and a selector, we aim to maintain a low number of pivots and an interpretable classifier structure. Therefore, the optimal *maxdepth* is searched within the interval $[2, 4]$. When using PivotTree as a selector, we assess the performance of using different pivot types – *discriminative*, *descriptive*, both, and using only those considered as splitting pivots – to identify which combination achieves the best selection performance when paired with DT or kNN. The best performance for kNN_P are obtained with *maxdepth* = 3, while for PTC and DT_P with *maxdepth* = 4. Leveraging both *discriminative* and *descriptive* pivots consistently yields better results. Finally, for the baseline DT and kNN the best performance is achieved with *maxdepth* = 4 and $k = 5$, respectively, both in

[3] A Python implementation, along with experimental details, is available at https://github.com/acascione/PivotTree_DoctOral.

[4] https://mlpi.ing.unipi.it/doctoralai/.

[5] For every tree, we set 3 as min nbr. of instances a node must have to be considered leaf, and 5 as the min nbr. of instances a node must have to perform a split.

the original space and in the similarity feature space. As distance function, we always adopt the Euclidean distance.

Dataset and Embedding Model. The DoctOral-AI dataset comprises 535 images of varying sizes, which define a multiclassification oral lesion detection task with classes *neoplastic* (31.58%), *aphthous* (32.52%), and *traumatic* (35.88%). Neoplastic ulcers typically exhibit the loss of epithelial layers, with raised, poorly defined margins. The base of these ulcers is often grayish, yellowish, or whitish, presenting a crater-like or raised appearance, generally composed of necrotic tissue with a granular texture. Aphthous ulcers, on the other hand, are characterized by the loss of epithelial layers and have flat, erythematous (red) margins with a grayish-yellow base, surrounded by red mucosa. Traumatic ulcers can feature raised or flat margins, bordered by a whitish or reddish rim, with a crater-like base in shades of white, gray, and yellow. Over time, the edges of chronic traumatic ulcers may harden and thicken. This detailed categorization is crucial for accurate diagnosis and treatment in clinical settings. The dataset is divided into 70% development and 30% testing, the former further divided on a 80%/20% split for training and validation. We embed images with a Detectron2 (D2) [47] CNN architecture fine-tuned on the dataset[6]. We resized each image into an 800 × 800 format. Then relevant feature maps are selected from the D2's backbone output and passed to the D2's region of interest pooling layer. Finally, a pooling layer and a flattening layer map the feature maps to a 256-dimensional embedding. We also report the performance of D2 to observe the loss in accuracy at the cost of interpretability.

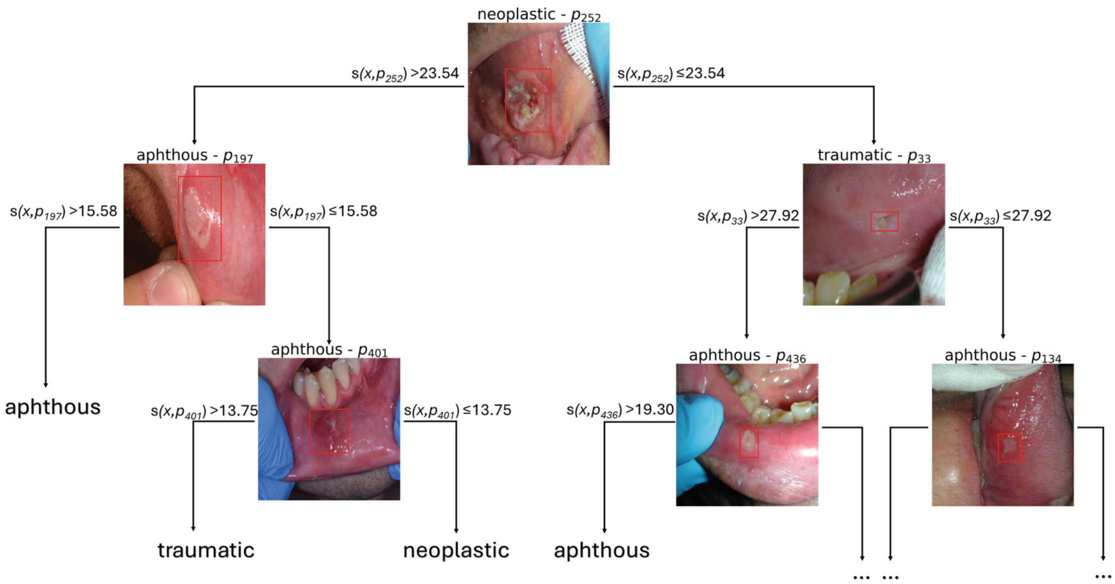


Fig. 3. Partial visual depiction of best PTC configuration on the test set. Branches are labeled with similarity threshold values used for prediction.

[6] We offer details regarding the training process in https://github.com/galatolofederico/oral-lesions-detection.

Qualitative Results. Fig. 3 depicts a visual representation of PTC decision rules and splitting pivots associated with the initial nodes[7]. Given a hypothetical instance x to predict, the predictive reasoning employed by the trained model proceeds as follows: x is first compared to p_{252}, a *neoplastic* instance. If the similarity between x and p_{252} is sufficiently high, then x traverses the left branch and is compared to the *aphthous* pivot p_{197}. If x is sufficiently similar to p_{197}, the model concludes the prediction and assigns x to the *aphthous* class. Otherwise, an additional comparison with p_{401} is performed, leading to a final classification as either *neoplastic* or *traumatic*. We underline that the path leading to *traumatic* decision lacks pivots belonging to such class. This suggests that the model can effectively perform comparisons with pivots belonging to other classes to exclude their possibility for x, thereby assigning x to the remaining class by exclusion[8]. On the other hand, if the initial comparison identifies x as dissimilar from the *neoplastic* p_{252}, the model then compares it to the *aphthous* p_{33} and applies analogous reasoning for subsequent comparisons.

Table 1. Mean predictive performance and number of pivots. Best performer in **bold**, second best performer in *italic*, third best performed underlined.

Model	Bal. Acc.	F1-score	Precision	Recall	Nbr. Pivots
D2	**0.859**	**0.854**	**0.854**	**0.858**	-
PTC	*0.834*	*0.832*	*0.839*	*0.834*	*9*
DT$_P$	0.833	0.830	0.830	0.833	47
kNN$_P$	0.811	0.807	0.810	0.811	**5**
DT$_O$	0.739	0.734	0.742	0.740	*9*
kNN$_O$	0.801	0.795	0.798	0.801	*9*
DT	0.770	0.766	0.772	0.770	-
kNN	0.809	0.808	0.811	0.810	-

Quantitative Results. Table 1 reports the mean predictive performance, and the number of pivots of the various predictive models[9]. D2 has the highest performance, at the cost of being not interpretable. However, a not markedly inferior performance is achieved by PivotTree predictor, i.e., PTC, that only requires 9 pivots (6 of which are shown in Fig. 3). The third best performer is PivotTree used as selector for a DT, i.e., DT$_P$. Unfortunately, such performance is accompanied by high complexity, as DT$_P$ requires 47 pivots. Finally,

[7] The actual trained tree has a *maxdepth* of 4. For visualization purposes, we limit the visualization to the initial nodes.

[8] We intend to fix this (possible) issue by extending PivotTree with Proximity Trees [30] to compare the test x against two pivots instead of only one.

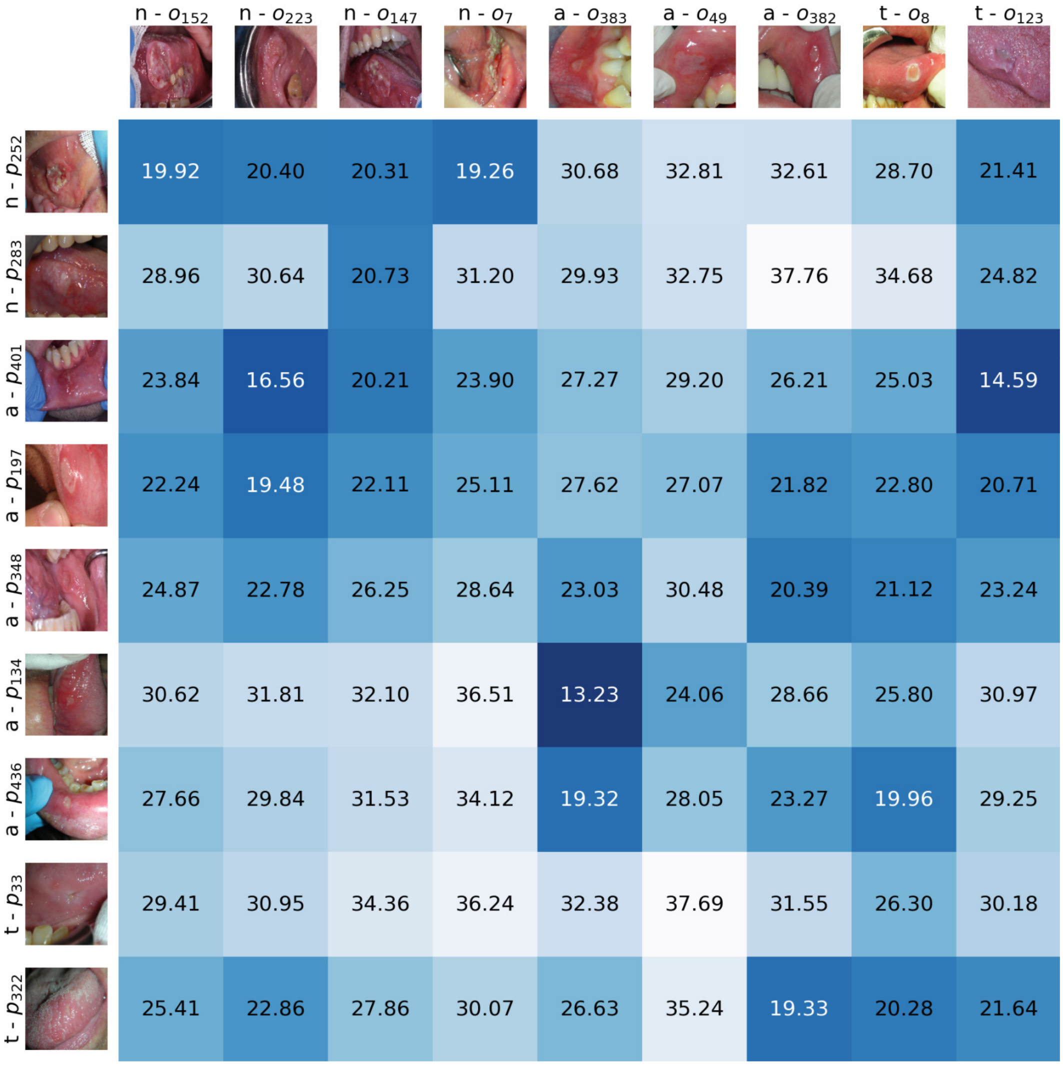
[9] For DT and PivotTree selector/predictor models we trained each best configuration with 50 different random states. Since the standard deviation of the values resulted to be negligible, we report only the average result.

kNN$_P$, i.e., PivotTree used as selector for a kNN is the predictor requiring the smallest number of pivots. Overall, PivotTree both employed as selector and predictor leads to competitive results compared to D2. We underline how PTC has the best trade-off between accuracy and complexity, showing competitive results with respect to the fine-tuned D2 but providing an interpretable predictor through its pivot structure, and the low number of pivots adopted.

Remarkably, selecting the set of pivots P through PivotTree leads to a kNN and a DT which are better than those resulting using the ground-truth prototypes, especially for the DT case, underlying that those instances which for humans are clear examples, perhaps didactic examples, of certain cases, are not necessarily the best ones to discriminate through an automatic AI system.

Finally, we remark that the performance of any PivotTree-based model is better than those of the kNN and DT classifiers directly trained on embeddings.

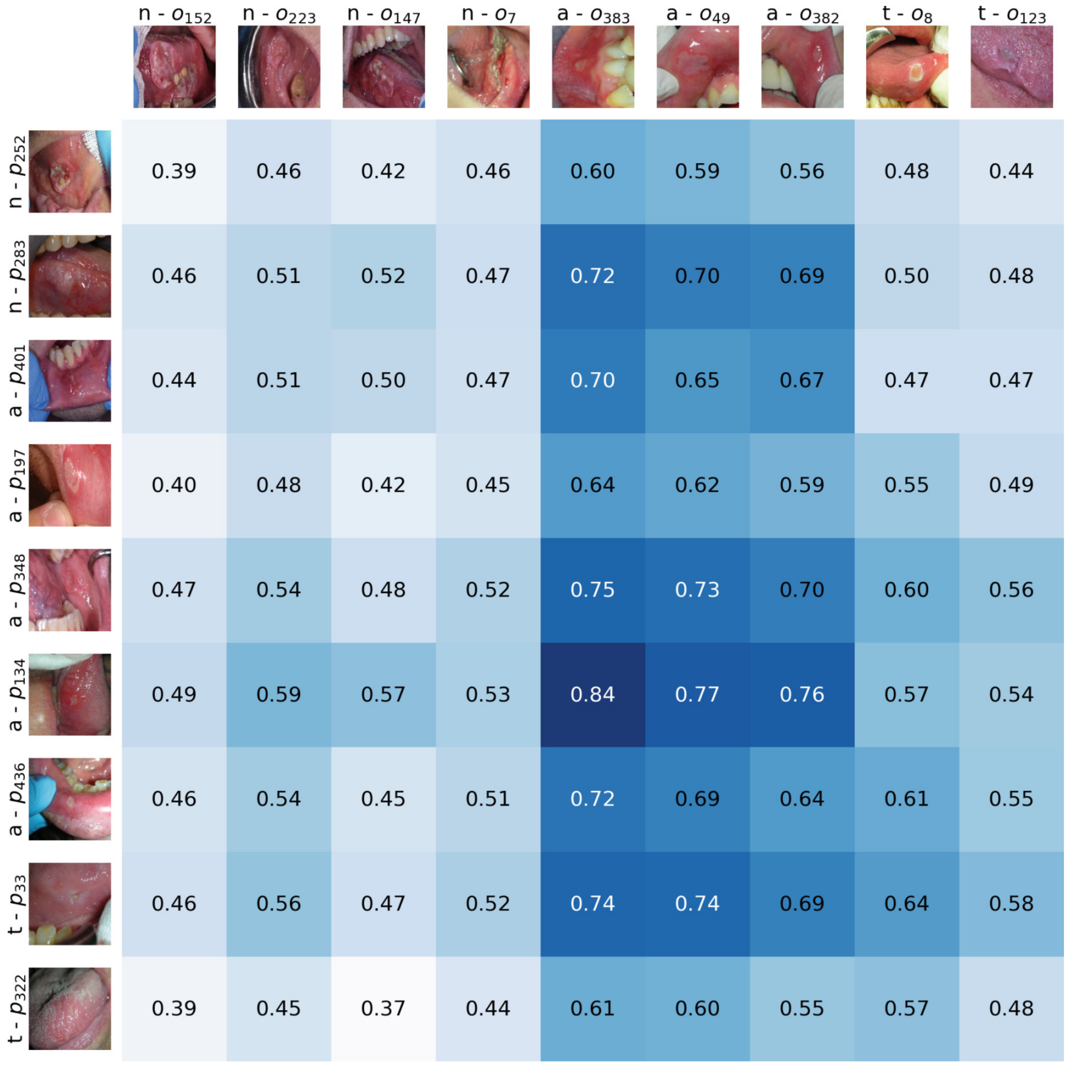
Pivot-Prototypes Comparison. We provide here a quantitative comparison in terms of similarities between the pivots selected through PTC P with the ground-truth prototypes O. In particular, we consider as similarity measures the Euclidean distance on the D2 embeddings, and the Structural Similarity (SSIM) [45] on the original images. For the latter, we first resize the images regions of interest to 300×300 pixels. SSIM identifies changes in structural information by capturing the inter-dependencies among similar pixels, especially when they are spatially close. In Figs. 4 and 5 we report two heatmaps highlighting the similarities between the PivotTree pivots (rows) and ground-truth prototypes (columns), on Euclidean and SSIM similarity, respectively. Darker colors indicate higher similarity. For the similarity comparison through Euclidean distance, we specify that the average distance between each pair of instances in the DoctOral-AI training set is 26.90 ± 6.48. When examining the average distance between pivot and ground-truth pairs w.r.t. each class in the heatmap, we find the following values: 23.93 for *neoplastic*, 24.65 for *aphthous*, and 24.60 for *traumatic*. This shows how the mean pairwise distances within individual classes are generally close to the overall mean pairwise distance. Pivots and ground-truth prototypes tend to not present robust similarities. Furthermore, we notice how for pivots p_{403} and p_{238}, both members of *aphthous* class, the most similar ground-truth prototypes belong to a different class. On the other hand, for the other pivots, the closest ground-truth counterpart is consistently one of the same class, sometimes with a very high similarity: some examples are p_{134} with o_{382} and p_{403} and o_{223}. A different tendency can be observed in Fig. 5 when using SSIM: the average SSIM with respect to each class is 0.46 for *neoplastic*, 0.70 for *aphthous*, and 0.57 for *traumatic*, with a mean similarity in the overall training set of 0.58 ± 0.10. This highlights a notably high internal similarity for the *aphthous* class. As evident from Fig. 5, the highest similarity is always observed when comparing pivots with the *aphthous* ground-truth prototypes, differently from Fig. 4 which shows higher variability across classes more oriented towards the right matching. This comparison corroborates the idea of relying on the Euclidean distance on the D2 embedding space for PivotTree.



	n - o_{152}	n - o_{223}	n - o_{147}	n - o_{7}	a - o_{383}	a - o_{49}	a - o_{382}	t - o_{8}	t - o_{123}
n - p_{252}	19.92	20.40	20.31	19.26	30.68	32.81	32.61	28.70	21.41
n - p_{283}	28.96	30.64	20.73	31.20	29.93	32.75	37.76	34.68	24.82
a - p_{401}	23.84	16.56	20.21	23.90	27.27	29.20	26.21	25.03	14.59
a - p_{197}	22.24	19.48	22.11	25.11	27.62	27.07	21.82	22.80	20.71
a - p_{348}	24.87	22.78	26.25	28.64	23.03	30.48	20.39	21.12	23.24
a - p_{134}	30.62	31.81	32.10	36.51	13.23	24.06	28.66	25.80	30.97
a - p_{436}	27.66	29.84	31.53	34.12	19.32	28.05	23.27	19.96	29.25
t - p_{33}	29.41	30.95	34.36	36.24	32.38	37.69	31.55	26.30	30.18
t - p_{322}	25.41	22.86	27.86	30.07	26.63	35.24	19.33	20.28	21.64

Fig. 4. PivotTree pivots (rows) and ground-truth prototypes (columns) comparison as Euclidean distances on D2 embedding. The darker the color the more similar are a pivot and a ground truth prototype. The first letter identifies the class of the instances: *n*eoplastic, *a*phthous, and *t*raumatic.

Furthermore, we evaluate how pivots extracted using PTC group instances together compared to ground-truth prototypes. We partition dataset instances with Voronoi partition, each instance associated to the prototype closest to it in the D2 embedding space. In Fig. 6, we compare group size and entropy calculated w.r.t. the target variable Y. In PTC groups, we highlight how *aphthous* pivots tend to aggregate the majority of instances, alongside the *traumatic* pivot p_{322}. *Aphthous* pivots exhibit higher entropy levels and form less pure groups, except for the smaller, entirely pure group centered around p_{134}. Conversely, the *neoplastic* instance p_{252} significantly captures a substantial percentage of its class, paralleled by p_{33} which similarly captures *traumatic* instances effectively. On the other hand, regarding ground-truth prototypes, most instances group around o_8, o_{382}, and o_{233}, representing *traumatic*, *aphthous*, and *neoplastic* classes, respectively. Many *neoplastic* instances are well-represented by o_{152},



	n - o_{152}	n - o_{223}	n - o_{147}	n - o_{7}	a - o_{383}	a - o_{49}	a - o_{382}	t - o_{8}	t - o_{123}
n - p_{252}	0.39	0.46	0.42	0.46	0.60	0.59	0.56	0.48	0.44
n - p_{283}	0.46	0.51	0.52	0.47	0.72	0.70	0.69	0.50	0.48
a - p_{401}	0.44	0.51	0.50	0.47	0.70	0.65	0.67	0.47	0.47
a - p_{197}	0.40	0.48	0.42	0.45	0.64	0.62	0.59	0.55	0.49
a - p_{348}	0.47	0.54	0.48	0.52	0.75	0.73	0.70	0.60	0.56
a - p_{134}	0.49	0.59	0.57	0.53	0.84	0.77	0.76	0.57	0.54
a - p_{436}	0.46	0.54	0.45	0.51	0.72	0.69	0.64	0.61	0.55
t - p_{33}	0.46	0.56	0.47	0.52	0.74	0.74	0.69	0.64	0.58
t - p_{322}	0.39	0.45	0.37	0.44	0.61	0.60	0.55	0.57	0.48

Fig. 5. PivotTree pivots (rows) and ground-truth prototypes (columns) comparison as SSIM on raw regions of interest. Same rules from Fig. 4 apply.

and similarly for *aphthous* instances with o_{383}. Instances of pure or almost pure groups are observed for *neoplastic* and *aphthous* classes, whereas a highly pure group for *traumatic* instances is lacking in this scenario. Only a single entirely pure group for *neoplastic* instances is found for ground-truth prototypes, whereas PTC pivots are able to isolate two entirely pure groups around p_{33} and p_{134}.

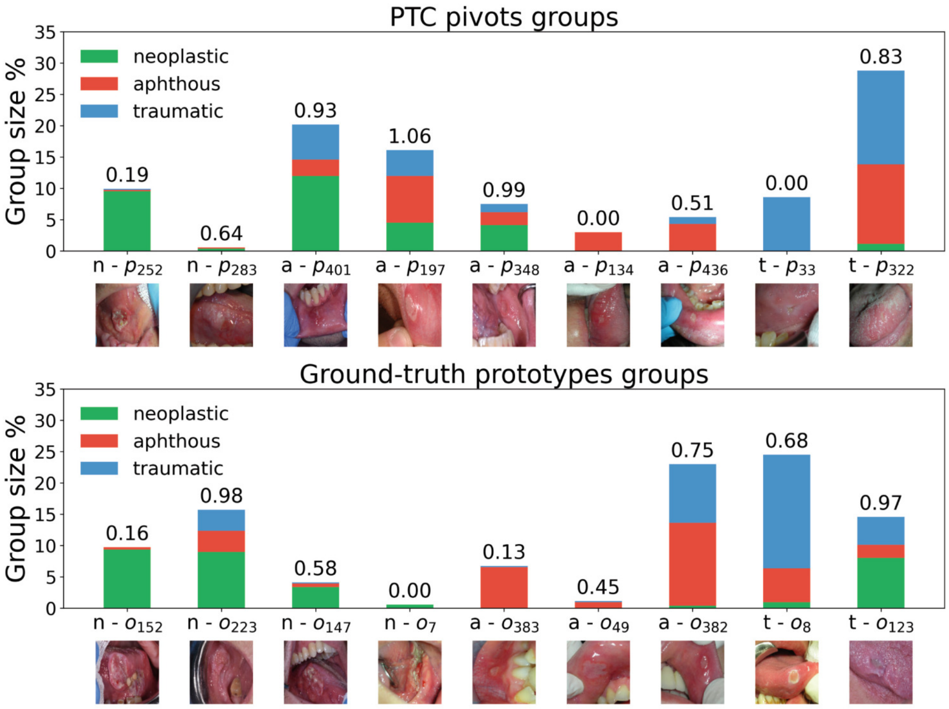


Fig. 6. Group sizes in percentage using PTC pivots and ground-truth prototypes as centers, respectively. Entropy values of each group with respect to the target variable are indicated by values above each bar.

5 Conclusion

We have discussed PivotTree application in the case of oral lesion prediction, showing its superiority as a predictor with respect to other simple interpretable models and as selector when paired with such simple models trained on the similarity space induced by the selected pivots. Furthermore, we have compared expert-selected prototypes with PivotTree-selected pivots, highlighting how a strong similarity can be observed in some of the pairs. Given its flexibility, PivotTree lends itself to be applied for several other diagnostic task in the healthcare sector. Future investigations include testing PivotTree on medical data of different modalities (time-series, text reports, tabular data) in order to assess its performance, comparing it against neural prototype-based approaches for medical data as explored in [26,39] and evaluating the interpretability of identified pivots through human subjects. Additional analysis could investigate the trade-off between performance and explainability by evaluating how PivotTree compares to competing post-hoc explainers. Moreover, future research could focus on developing specialized interpretability metrics for PivotTree and other case-based models, as this study primarily relied on depth and the number of pivots to assess interpretability and complexity. Furthermore, other splitting strategies could be analyzed, one being a direct comparison between pairs of pivots as shown in ProximityTree models [30] or attempting to generate instead of select the PivotTree model [19].

Acknowledgments. This work has been partially supported by the European Community Horizon 2020 programme under the funding schemes ERC-2018-ADG G.A. 834756 “XAI: Science and technology for the eXplanation of AI decision making”, “INFRAIA-01-2018-2019 - Integrating Activities for Advanced Communities”, G.A. 871042, “SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics”, by the European Commission under the NextGeneration EU programme - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) - Project: “SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics” - Prot. IR0000013 - Avviso n. 3264 del 28/12/2021, and M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI”, 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”, by the Italian Project Fondo Italiano per la Scienza FIS00001966 MIMOSA, by the “Reasoning” project, PRIN 2020 LS Programme, Project number 2493 04-11-2021, by the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence), by the European Union, Next Generation EU, within the PRIN 2022 framework project PIANO (Personalized Interventions Against Online Toxicity) under CUP B53D23013290006.

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

References

1. Adeoye, J., et al.: Explainable ensemble learning model improves identification of candidates for oral cancer screening. Oral Oncol. **136**, 106278 (2023)
2. Ali, S., et al.: The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. Comput. Biol. Medicine **166**, 107555 (2023)
3. Band, S.S., et al.: Application of explainable artificial intelligence in medical health: a systematic review of interpretability methods. Inf. Med. Unlocked **40**, 101286 (2023)
4. Baptista, M.L., et al.: Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. Artif. Intell. **306**, 103667 (2022)
5. Bichindaritz, I., Marling, C.: Case-based reasoning in the health sciences: what’s next? Artif. Intell. Med. **36**(2), 127–135 (2006)
6. Breiman, L., et al.: Classification and Regression Trees. Wadsworth (1984)
7. Cascione, A., et al.: Data-agnostic pivotal instances selection for decision-making models. In: Bifet, A., Davis, J., Krilavičius, T., Kull, M., Ntoutsi, E., Žliobaitė, I. (eds.) ECML/PKDD, vol. 14941, pp. 367–386. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-70341-6_22
8. Celard, P., et al.: A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. Neural Comput. Appl. **35**(3), 2291–2323 (2023)
9. Chen, C., et al.: This looks like that: deep learning for interpretable image recognition. In: NeurIPS, pp. 8928–8939 (2019)
10. Choudhury, N., Begum, S.A.: A survey on case-based reasoning in medicine. IJACSA **7**(8), 136–144 (2016)

11. Coderre, S., et al.: Diagnostic reasoning strategies and diagnostic success. Med. Educ. **37**(8), 695–703 (2003)
12. Dixit, S., et al.: A current review of machine learning and deep learning models in oral cancer diagnosis: recent technologies, open challenges, and future research directions. Diagnostics **13**(7), 1353 (2023)
13. Ehtesham, H., et al.: Developing a new intelligent system for the diagnosis of oral medicine with case-based reasoning approach. Oral Dis. **25**(6), 1555–1563 (2019)
14. Figueroa, K.C., et al.: Interpretable deep learning approach for oral cancer classification using guided attention inference network. JBO **27**(1), 015001 (2022)
15. Fix, E.: Discriminatory analysis: nonparametric discrimination, consistency properties, vol. 1. USAF school of Aviation Medicine (1985)
16. Frasca, M., et al.: Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. Discov. Artif. Intell. **4**(1) (2024). https://doi.org/10.1007/s44163-024-00114-7
17. Grignaffini, F., et al.: Machine learning approaches for skin cancer classification from dermoscopic images: a systematic review. Algorithms **15**(11), 438 (2022)
18. Guidotti, R., et al.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2019)
19. Guidotti, R., et al.: Generative model for decision trees. In: AAAI, pp. 21116–21124. AAAI Press (2024)
20. Harasym, P.H., et al.: Current trends in developing medical students' critical thinking abilities. KJMS **24**(7), 341–355 (2008)
21. Javaid, M., et al.: Significance of machine learning in healthcare: features, pillars and applications. Int. J. Intell. Networks **3**, 58–73 (2022)
22. Jocher, G.: YOLOv5 by Ultralytics. https://github.com/ultralytics/yolov5
23. Johnson-Laird, P.N.: Mental models and human reasoning. Proc. Natl. Acad. Sci. **107**(43), 18243–18250 (2010)
24. Kang, E., et al.: Prototype learning of inter-network connectivity for ASD diagnosis and personalized analysis. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI. LNCS, vol. 13433, pp. 334–343. Springer (2022). https://doi.org/10.1007/978-3-031-16437-8_32
25. Kim, B., et al.: Examples are not enough, learn to criticize! criticism for interpretability. In: NIPS, pp. 2280–2288 (2016)
26. Kim, E., et al.: XProtoNet: diagnosis in chest radiography with global and local explanations. In: CVPR, pp. 15719–15728. Computer Vision Foundation / IEEE (2021)
27. Kouketsu, A., et al.: Detection of oral cancer and oral potentially malignant disorders using artificial intelligence-based image analysis. Head Neck **46**, 2253–2260 (2024)
28. Lamy, J., et al.: Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. Artif. Intell. Medicine **94**, 42–53 (2019)
29. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
30. Lucas, B., et al.: Proximity forest: an effective and scalable distance-based classifier for time series. Data Min. Knowl. Discov. **33**(3), 607–635 (2019)
31. Metta, C., et al.: Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling. In: ISCC, pp. 1–7. IEEE (2021)
32. Metta, C., et al.: Improving trust and confidence in medical skin lesion diagnosis through explainable deep learning. JDSA, 1–13 (2023). https://doi.org/10.1007/s41060-023-00401-z

33. Metta, C., et al.: Advancing dermatological diagnostics: interpretable AI for enhanced skin lesion classification. Diagnostics **14**(7), 753 (2024)
34. Panigutti, C., et al.: Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In: FAT*, pp. 629–639. ACM (2020)
35. Schank, R.C., Abelson, R.P.: Knowledge and Memory: The Real Story, pp. 1–85. Psychology Press (2014)
36. Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition - Foundations and Applications, Series in Machine Perception and Artificial Intelligence, vol. 64. WorldScientific (2005)
37. Schank, R.C., Abelson, R.P.: Knowledge and Memory: The Real Story, pp. 1–85. Psychology Press (2014)
38. Shin, H.S.: Reasoning processes in clinical reasoning: from the perspective of cognitive psychology. KJME **31**(4), 299 (2019)
39. Singh, G., Yow, K.C.: An interpretable deep learning model for Covid-19 detection with chest x-ray images. IEEE Access **9**, 85198–85208 (2021)
40. Song, B., et al.: Interpretable and reliable oral cancer classifier with attention mechanism and expert knowledge embedding via attention map. Cancers **15**(5), 1421 (2023)
41. Song, B., et al.: Classification of mobile-based oral cancer images using the vision transformer and the SWIN transformer. Cancers **16**(5), 987 (2024)
42. Spelke, E.S.: What babies know: Core Knowledge and Composition Volume 1, vol. 1. Oxford University Press (2022)
43. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (2019)
44. Wang, C., et al.: Knowledge distillation to ensemble global and interpretable prototype-based mammogram classification models. n: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI. LNVCS, vol. 13433, pp. 14–24. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_2
45. Wang, Z., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
46. Welikala, R.A., et al.: Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. IEEE Access **8**, 132677–132693 (2020)
47. Wu, Y., et al.: Detectron2 (2019). https://github.com/facebookresearch/detectron2
48. Yagin, B., et al.: Cancer metastasis prediction and genomic biomarker identification through machine learning and explainable artificial intelligence in breast cancer research. Diagnostics **13**(21), 3314 (2023)
49. Yang, G., et al.: Unbox the black-box for the medical explainable AI via multimodal and multi-centre data fusion: a mini-review, two showcases and beyond. Inf. Fusion **77**, 29–52 (2022)
50. Zhou, J., et al.: A pathology-based diagnosis and prognosis intelligent system for oral squamous cell carcinoma using semi-supervised learning. ESWA **254**, 124242 (2024)