



ELSEVIER

Contents lists available at ScienceDirect

Information Sciences

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

REPROT: Explaining the predictions of complex deep learning architectures for object detection through reducts of an image

Marilyn Bello ^a, Gonzalo Nápoles ^{b,*}, Leonardo Concepción ^{d,c}, Rafael Bello ^d, Pablo Mesejo ^a, Óscar Cerdón ^a

^a Universidad de Granada, Granada, Spain

^b Tilburg University, Tilburg, the Netherlands

^c Hasselt University, Hasselt, Belgium

^d Universidad Central de Las Villas, Santa Clara, Cuba

ARTICLE INFO

Keywords:

Deep learning
Visual explanation
Rough set theory
Reduct
Prototype image

ABSTRACT

Although deep learning models can solve complex prediction problems, they have been criticized for being ‘black boxes’. This implies that their decisions are difficult, if not impossible, to explain by simply inspecting their internal knowledge structures. Explainable Artificial Intelligence has attempted to open the black-box through model-specific and agnostic post-hoc methods that generate visualizations or derive associations between the problem features and the model predictions. This paper proposes a new method, termed REPROT, that explains the decisions of complex deep learning architectures based on local reducts of an image. A ‘reduct’ is a set of sufficiently descriptive features that can fully characterize the acquired knowledge. The created reducts are used to build a ‘prototype image’ that visually explains the inference obtained by a black-box model for an image. We focus on deep learning architectures whose complexity and internal particularities demand adapting existing model-specific explanation methods, making the explanation process more difficult. Experimental results show that the black-box model can detect an object using the prototype image generated from the reduct. Hence, the explanations will be given by “the minimum set of features sufficient for the neural model to detect an object”. The confidence scores obtained by architectures such as Inception, Yolo, and Mask R-CNN are higher for prototype images built from the reduct than those built from the most important superpixels according to the LIME method. Moreover, the target object is not detected on several occasions through the LIME output, thus supporting the superiority of the proposed explanation method.

1. Introduction

Although machine learning and deep learning models often produce accurate predictions, sometimes we need to know what motivates a certain decision. For instance, suppose that a model issues an alert on the condition of patients suffering from a specific disease in a hospital. The alert will be triggered when a patient requires immediate intervention by healthcare personnel. It may happen that, *a priori*, the medical doctor does not know why the patient’s condition has worsened and must analyze all the patient’s

* Corresponding author.

E-mail address: g.r.napoles@tilburguniversity.edu (G. Nápoles).

<https://doi.org/10.1016/j.ins.2023.119851>

Received 25 July 2023; Received in revised form 23 October 2023; Accepted 31 October 2023

Available online 7 November 2023

0020-0255/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

vital signs over the last few hours to isolate possible causes. However, using an interpretable model, the medical doctor could focus on those variables related to the warning, saving valuable time and effort.

Explainable Artificial Intelligence (XAI) [3,5] refers to procedures used to explain decisions generated by AI-based systems to humans. Explicability can be *ante-hoc* (directly interpretable models [24,8]) or *post-hoc* (techniques to explain a previously trained model or prediction [2,37,14,34,9]). Post-hoc techniques have been widely used for their usefulness in explaining the predictions of black boxes [52,49,40,4]. Post-hoc techniques include model-agnostic methods (which can be seamlessly applied to any machine learning model regardless of its internal processing or internal representations [48,37,26,38,33]) and model-specific methods (tailored or specifically designed to explain particular machine learning models [2,29,41,42]). Following the taxonomy in [3], model-agnostic methods may rely on model simplification [37], feature relevance estimation [26], and visualization techniques [10].

Although most of these techniques are based on local explanations (i.e., explanations for a specific instance [37]), others rely on constructing global explanations (i.e., explanations focused on the global decisions of the prediction model [12]). In particular, Local Interpretable Model-Agnostic Explanations (LIME [37]) is a model simplification method for local explanations. The intuition behind LIME is that it is much easier to approximate a black-box model by a simple model locally rather than by trying to approximate a model globally. LIME focuses on fitting a local surrogate model to explain why specific predictions are made. Surrogate models are interpretable models (e.g., a linear regression model) that are learned over the projections of the original black-box model. LIME tests what happens to the model predictions when data changes are introduced. It generates a new dataset consisting of perturbed samples (i.e., perturbations made to the instance to be explained) and the associated black-box model predictions. It then trains an interpretable model weighted by the proximity of the sample to the instance of interest. The linear model coefficients are finally interpreted as the importance of the features either for (positive) or against (negative) the predicted class.

This paper proposes a model-agnostic method that provides visual explanations for the predictions of black-box classifiers dedicated to computer vision tasks [21,15]. The proposed explanation method, named REPROT, creates a ‘prototype image’ instead of training a local surrogate model to provide a visual explanation of the output of a black-box model for a given image. When building a prototype, we rely on the concept of Region of Interest (ROI) [7] and extend the concept of ‘reduct’ from the Rough Set Theory (RST) [31,32] to the universe of discourse of images. This theory provides a solid mathematical tool for feature selection, inconsistency analysis, and knowledge discovery for structured decision systems. However, we must first extend the information system concept in the RST formalism for image classification tasks to apply this theory to unstructured data. Similarly, we must redefine the inseparability relation to obtain an ROI-based equivalence relation for pairs of images. Therefore, a reduct can be understood as the minimal feature set that allows correctly classifying images and inducing minimal length decision rules describing the problem domain. However, reduct computation is computationally expensive, especially for image datasets having complex structures. To circumvent this issue, the proposed method relies on the local reducts of a given image. These information granules are obtained from the neighborhood of the image to be explained and the model’s predictions for that neighborhood. After that, the prototype image is built from the set of features included in the generated reduct. To validate the explanations produced by our method, the classification process is performed under the hypothesis that the model should be able to produce the output being explained from the prototype image only.

Overall, the paper brings two contributions within the data analysis and XAI fields, respectively. The first contribution concerns a theoretical formalism based on RST that extends the notions of reducts and regions of interest in universes of discourse where objects are images instead of instances described by well-defined features. To the best of our knowledge, the proposed RST formalism can be considered the first study in which this mathematical theory is applied to unstructured data where decision tables do not exist. The second contribution concerns the agnostic explanation method that generates high-confidence explanations in computer vision settings. When compared with existing explanation approaches reported in the literature, our method includes the following advantageous features:

- Firstly, in contrast to model-specific post-hoc explanation approaches such as Layer-wise Relevance Propagation (LRP [2]) and DeepTaylor [29], the proposed explanation method can be applied to almost any black-box model dedicated to computer vision tasks due to its agnostic nature. Notice that the internal particularities of complex neural architectures prevent these methods from being used directly, often requiring algorithmic adjustments. Examples of these networks include Yolo [51] and Mask R-CNN [15].
- Secondly, we empirically demonstrate that the precision scores obtained by the neural models when describing the images with the local reducts are higher than those built from the most important superpixels discovered by other methods of an agnostic nature, such as LIME or ANCHORS [38] methods. More importantly, we reported several cases where using LIME or ANCHORS superpixels to describe the images caused misclassifications. The high confidence and accurate predictions using local RST reducts provide evidence of the superior descriptive power of local reducts used to generate the explanations.
- Thirdly, the proposed RST extension to universes of discourses comprised of images and the REPROT method pave the road toward producing more complex explanations. For example, the foundations of both contributions can be coupled with symbolic reasoning mechanisms as the module proposed by [30] to generate high-confidence counterfactual explanations.

The rest of the paper is organized as follows. Section 2 presents the model-agnostic approaches prevalent in the literature. Section 3 introduces some basic concepts of RST adapted to image-based information systems describing a universe of discourse. Section 4 presents the proposed local explanation approach based on an image prototype. Section 5 conducts numerical simulations and pertinent discussions. Finally, Section 6 concludes the paper and provides future research directions.

2. Related work

In the field of XAI, several post-hoc local explainability methods characterized as model-agnostic methods have been developed, such as [37,26,38,33]. These techniques generate explanations on the inference of a learned model from a specific input without depending on the internal architecture of the model, operating only from its input and output [48,19]. Some of the existing agnostic approaches include:

- Local Interpretable Model-Agnostic Explanations (LIME) [37]: This method creates a surrogate model from solutions inferred by the neural model in the neighborhood of the input being explained. LIME learns an interpretable linear model in the vicinity of an input instance. The utility of the learned linear model is that it calculates the importance of the input features of the instance to be explained concerning the output predicted by the neural model. This importance represents the explanation that is finally provided to the user. The neighborhood is constructed by perturbing the instance to be explained and inferring the solutions of these new instances generated using the neural model being explained. Within the image domain, the method is classified as a Saliency Map, i.e., an image in which a pixel's brightness represents the pixel's relevance. LIME presents its explanation from superpixels, i.e., sets of pixels whose coefficients in the linear model have higher positive values.
- Shapley Additive Explanations (SHAP) [26]: This method is based on Shapley values, a concept from cooperative game theory. Shapley values assign a value to each player in a cooperative game based on their contribution to the game's outcome. In machine learning, the "players" are the input features, and the "game" is the model's prediction. SHAP assigns a Shapley value to each feature for a specific prediction made by a model. These values represent the contribution of each feature to the prediction, considering all possible combinations of features. SHAP ensures that the sum of the Shapley values of all features equals the difference between the model's prediction for a particular case and the average prediction of all instances, ensuring balance in feature attribution. However, calculating Shapley values for all features or pixels may require a high computational cost.
- High-Precision Model-Agnostic Explanations (ANCHORS) [38]: This method can be considered an extension of LIME, which provides explanations in the form of decision rules (if-then rules) called anchors. An anchor consists of a set of features needed to locally explain the response of a black-box model, so that the value of any feature not included in the anchor can be changed and the prediction is not affected. Since the number of possible anchors is exponential, a search method is used to construct a subset of them. When building the surrogate model, the rules are computed by progressively adding equality conditions on the premise (the anchor is initialized with an empty rule) until the rule reaches a set accuracy threshold to perform the prediction.
- Randomized Input Sampling for Explanation (RISE) [33]: This method explains deep learning models' predictions for image classification. RISE generates a set of randomized versions of the input image by applying a specific perturbation method. These perturbed images are created by randomly zeroing a portion of the image pixels, which the authors define as masking an image. RISE calculates an importance score for each pixel of the original image based on the variability of the model predictions when different parts of the image are masked. Pixels that, when masked, cause the most variation in model predictions are considered the most important for the model decision. The importance scores are used to generate a heatmap overlaid on top of the original image.

Given the current applications of these approaches in explaining various deep neural models [39,22,6], particularly those dedicated to image classification, they will be included in the comparative study alongside our proposal.

3. RST-based formalism for an image universe

RST [31,32] is defined by two main components: an information system and an inseparability relation [53]. Since we are dealing with images and objects to be detected in these, we need to modify the classical composition of the information system. Therefore, we introduce an RST-based formalism for the domain of image processing.

In this domain, an information system can be defined as a 3-tuple $(\mathcal{U}, \mathcal{S}, \mathcal{O})$, where \mathcal{U} is a non-empty and finite set of images, and \mathcal{S} is a non-empty, finite family of sets of superpixels (i.e., sets of pixels generated from a segmentation algorithm [50]), whereas $\mathcal{O} = \{o_1, o_2, \dots, o_D\}$ is the set of possible objects to be detected in a given image. Specifically, $\mathcal{U} = \{I_1, I_2, \dots, I_N\}$ and $\mathcal{S} = \{S_{I_1}, S_{I_2}, \dots, S_{I_N}\}$, where N is the total number of images and there is a one-to-one correspondence between \mathcal{U} and \mathcal{S} . Also, it should be noticed that the segmentation algorithm generates the same number of superpixels for every image and returns an ordered set. Fig. 1 summarizes the relationship between the components of an information system in the RST formalism.

Let us establish that $K = |S_{I_j}|$ such that $|\cdot|$ denotes the set cardinality operator. Since $|S_{I_j}|$ is constant $\forall j$, $S_{I_j} = (S_{I_j}^1, S_{I_j}^2, \dots, S_{I_j}^K) \forall j$ represents the ordered set with the superpixels. Given a set containing indexes from 1 to K , the ordered subset $S'_{I_j} \subseteq S_{I_j}$ contains all superpixels in S_{I_j} (for all j) with superscripts corresponding to these indexes and keeping its original order. For example, the set of indexes $\{1, 2, 5\}$ would induce $S'_{I_j} = (S_{I_j}^1, S_{I_j}^2, S_{I_j}^5) \forall j$. Moreover, we say that $S'_{I_x} = S'_{I_y}$ if the images I_x and I_y are deemed identical for the corresponding superpixels. Equation (1) formalizes the inseparability relation that enforces this behavior,

$$IND_1 = \{(I_x, I_y) \in \mathcal{U} \times \mathcal{U} : S'_{I_x} = S'_{I_y}\}. \quad (1)$$

In short, Equation (1) establishes the relationships between two or more images in the universe for a given set containing indexes to reference the superpixels generated by a segmentation algorithm. Since the condition in Equation (1) is excessively strict in most situations, a relaxed equivalence relation between two images must be defined. Notice that this type of relation must satisfy three



Fig. 1. Relationship between the components of the 3-tuple $(\mathcal{U}, \mathcal{S}, \mathcal{O})$ in the RST formalism.

properties: reflexivity, symmetry, and transitivity. According to Equation (2), two images are considered equivalent if the object $o_i \in \mathcal{O}$ is detected in both of them or none,

$$R_1^{o_i} = \{(I_x, I_y) \in \mathcal{U} \times \mathcal{U} : o_i \in \mathcal{L}(I_x) \wedge o_i \in \mathcal{L}(I_y) \vee o_i \notin \mathcal{L}(I_x) \wedge o_i \notin \mathcal{L}(I_y)\} \quad (2)$$

where $\mathcal{L}(I_x)$ and $\mathcal{L}(I_y)$ are the sets of annotated objects in the images I_x and I_y , respectively. Consequently, $[I_x]_{R_1^{o_i}} = \{I_y \in \mathcal{U} : I_y R_1^{o_i} I_x\}$ defines the equivalence class of I_x according to the $R_1^{o_i}$ relation, i.e., all images equivalent to I_x with respect to the i -th annotated object. In general, only two equivalence classes arise (one containing images where the object o_i is detected and the other with images where the object was not detected).

3.1. Adaptation for the regions of interest of an image

Next, we will adapt the previous RST-based formalism for the case of the ROI [7] of specific images in \mathcal{U} .

Definition 1. An ROI of an image can be defined using a binary mask. Mask pixel values of 1 indicate image pixels that belong to the ROI, while values of 0 indicate image pixels that are part of the background.

In this paper, ROIs are produced using a binary mask for the superpixels generated from a segmentation algorithm. Aiming to build an ROI from a given set of superpixels (S_{I_j}), we first need to define a subset $S'_{I_j} \subseteq S_{I_j}$. Subsequently, we activate every pixel contained in S'_{I_j} (i.e., pixels that are part of the superpixels in S'_{I_j}) and deactivate every other pixel. In simpler terms, superpixels are turned on and off.

For every image $I_j \in \mathcal{U}$, the proposed information system can be defined as a 3-tuple $(P_{I_j}, S_{I_j}, \mathcal{O})$, where P_{I_j} is a non-empty and finite set of all possible ROIs for I_j , and $S_{I_j} \in \mathcal{S}$. It should be noticed that the ROIs in P_{I_j} are built from S_{I_j} . Specifically, $P_{I_j} = \{ROI_1, ROI_2, \dots, ROI_M\}$, where M is the total number of ROIs for the image I_j . It is worth mentioning that $M = 2^K - 1$ (K is the number of superpixels produced after the segmentation), considering that every ROI has at least one superpixel turned on. Fig. 2 summarizes the composition of an information system in adapting RST to image ROIs.

The intuition behind Equation (1) is still valid if we consider that I_x and I_y represent the same image. Therefore, the case when $S'_{I_x} = S'_{I_y}$ is analogous to $ROI_v = ROI_w$ where ROI_v and ROI_w are generated from S'_{I_x} and S'_{I_y} , respectively. Aiming at simplifying and improving clarity, Equation (3) formalizes an inseparability relation that establishes the relationships between two or more ROIs in the universe (P_{I_j}),

$$IND_2 = \{(ROI_v, ROI_w) \in P_{I_j} \times P_{I_j} : ROI_v = ROI_w\}. \quad (3)$$

Equation (1) involves a strict condition that only holds if both ROIs are the same. However, the equivalence relation between two ROIs is defined by Equation (4), in which two ROIs are considered equivalent if the object $o_i \in \mathcal{O}$ is detected in both of them or none,

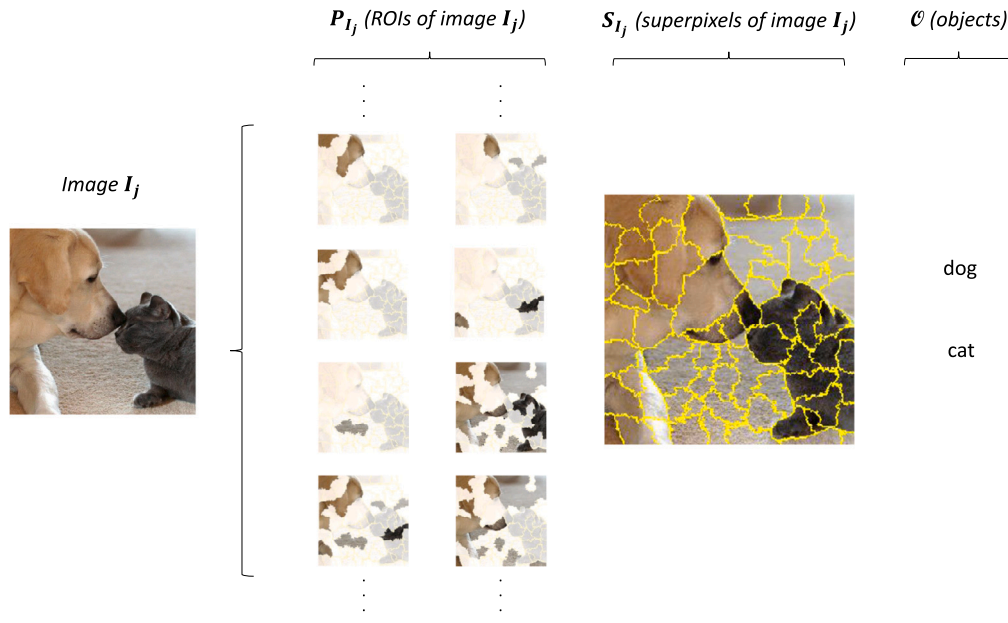


Fig. 2. Relationship between the components of the 3-tuple (P_{I_j}, S_{I_j}, O) in the adaptation of RST to image ROIs.

$$\begin{aligned}
 R_2^{o_i} = & \{(ROI_v, ROI_w) \in P_{I_j} \times P_{I_j} : \\
 & o_i \in \mathcal{L}(ROI_v) \wedge o_i \in \mathcal{L}(ROI_w) \vee \\
 & o_i \notin \mathcal{L}(ROI_v) \wedge o_i \notin \mathcal{L}(ROI_w)\}
 \end{aligned} \tag{4}$$

where $\mathcal{L}(ROI_v)$ and $\mathcal{L}(ROI_w)$ represent the sets of annotated objects in the regions ROI_v and ROI_w , respectively. Consequently, $[ROI_v]_{R_2^{o_i}} = \{ROI_w \in P_{I_j} : ROI_w R_2^{o_i} ROI_v\}$ defines the equivalence class of ROI_v according to the $R_2^{o_i}$ relation. Such information granule contains all ROIs equivalent to ROI_v with respect to the i -th annotated object.

4. REPROT: visual explanations using prototype images

Our approach builds a ‘prototype image’ that serves as a visual explanation for the output of a black-box model for a given image. A prototype represents each object detected by the model and consists of the essential superpixels when detecting the object. Before building the prototype image, we need to introduce the concept of local reduct in this domain. For simplicity, it is called ‘reduct’ throughout the paper. In addition, the concept of multi-reduct is also given below.

Definition 2. A local reduct $\mathcal{R}_{I_x}(o_i^*)$ for the image I_x is an ROI with the minimum set of superpixels from S_{I_x} on required to detect the object o_i^* .

In other words, $\mathcal{R}_{I_x}(o_i^*)$ is the minimal set of features (superpixels in this context) obtained from I_x to explain o_i^* . In addition, notice that a reduct might not be unique, and we could end up with a set of reducts with the same number of activated superpixels, as explained below.

Definition 3. The multi-reduct $\mathcal{M}_{I_x}(o_i^*)$ is the set of local reducts for the image I_x where the object o_i^* is detected.

We can conclude from Definition 2 that every reduct in $\mathcal{M}_{I_x}(o_i^*)$ contains the same number of superpixels. An empty multi-reduct might exist if we use objects that are never detected for the image. Nevertheless, in our case, it does not occur because we try to explain the detection of an object by a black-box model. Consequently, inside the multi-reduct, at least one reduct explains the detected object. Based on a chosen reduct $\mathcal{R}_{I_x}^*(o_i^*)$ we build the prototype image as defined below.

Definition 4. A prototype $\mathcal{P}_{I_x}(o_i^*)$ is an image built from the reduct $\mathcal{R}_{I_x}^*(o_i^*)$.

Let us suppose that I_x is the image of interest (i.e., an image for which a black-box model \mathcal{B} provided an output that must be explained), $\mathcal{L}(I_x) \subseteq \mathcal{O}$ is the set of objects predicted by the \mathcal{B} model for I_x , where $o_i^* \in \mathcal{L}(I_x)$ is the object to be explained. The prototype image $\mathcal{P}_{I_x}(o_i^*)$ consists of the superpixels in $\mathcal{R}_{I_x}^*(o_i^*)$ that must be on for the object o_i^* to be detected by the \mathcal{B} model. Consequently, a prototype consists of the minimum set of superpixels sufficient to detect an object.

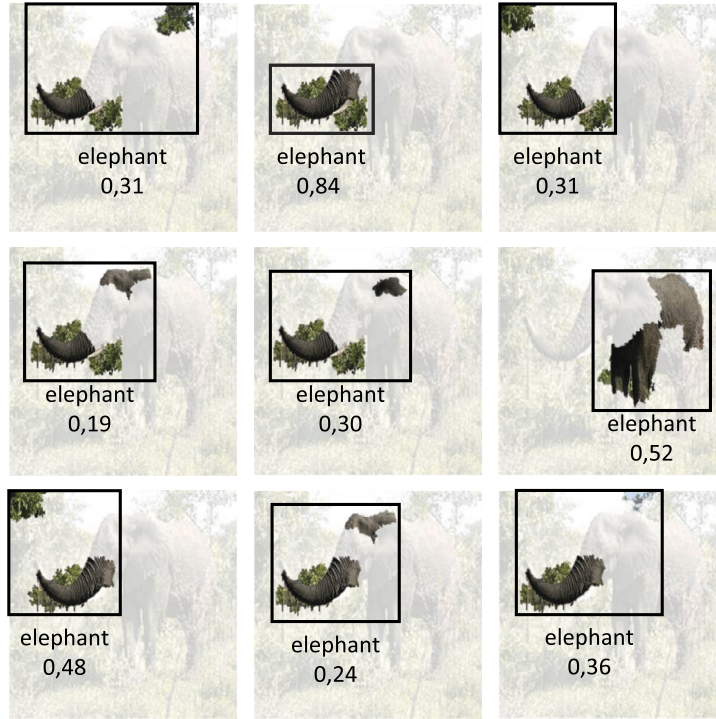


Fig. 3. Several reducts of an image classified as “African elephant” by the Inception-v3 model, with a precision of 0.67 (i.e., when the whole image is considered). Each subfigure includes the superpixels composing the reduct and the precision obtained by the neural model when inference is performed from the prototype image.

As an illustrative example, Fig. 3 shows several reducts of an image classified as “African elephant” by the Inception-v3 model [46]. It is interesting to see how the model can detect the elephant from the second reduct (i.e., the second subfigure at the top) more accurately than if the whole image is considered. In terms of explanations, the trunk and the front legs are the most relevant superpixels to classify the image as an “African elephant”, as long as a superpixel representing the jungle (e.g., tree branches) appears in the image being explained.

4.1. Method for building local reducts

The proposed explanation method builds a multi-reduct (Definition 4) using a heuristic approach to find a candidate reduct and uses it as a starting point for a brute force search. The reduct is determined from the neighborhood of the image being explained. This neighborhood is obtained from perturbed samples of the image (by turning on and off the superpixels of the image). The algorithm finds the equivalence class of an image according to the object to be explained and detects the most dissimilar sample. Subsequently, a candidate reduct is built from the most dissimilar sample image. Computing the most dissimilar sample image reduces the computational cost of obtaining the reduct since this image could represent the reduct or contain the candidate reduct that can be used as a starting point to find the multi-reduct. Having this, we turn the superpixels on and off until we detect the object itself. The obtained candidate establishes the greatest number of superpixels a reduct could have, so we try all possible combinations (for the original image) up to this number (starting from a superpixel). A detailed description of each step is shown below:

1. Choose the image of interest I_x and the object o_i^* detected by the \mathcal{B} model for which it is desired to have an explanation.
2. Create perturbations from the image to generate a set of ROIs for I_x . Such a set is called a neighborhood and is denoted by $H_{I_x} \subseteq P_{I_x}$.
 - (a) Superpixels S_{I_x} are generated using a segmentation algorithm (e.g., Quickshift [50]).
 - (b) A mask with random zeros (i.e., the superpixel is off) and ones (i.e., the superpixel is on) is generated. These shape a matrix with perturbations as rows and superpixels as columns. Since we want to find the minimum set of superpixels sufficient to detect an object, the probability of generating superpixels turned on should be close to zero. These masks generate H_{I_x} .
 - (c) Obtain the predictions for H_{I_x} as computed by the \mathcal{B} model.

We generate as many masks as the 10% of the total number of superpixels produced by the segmentation ($\lfloor \frac{2^n - 1}{10} \rfloor$).

3. Compute the equivalence class $[I_x]_{R_1^{o_i^*}}$ according to Equation (4), i.e., the ROIs in the neighborhood H_{I_x} where o_i^* is detected.
4. Calculate the similarity of I_x to all ROIs in $[I_x]_{R_1^{o_i^*}}$ using a similarity function, e.g., the cosine similarity [13].
5. Compute the most dissimilar ROI in $[I_x]_{R_1^{o_i^*}}$ to the image of interest I_x .

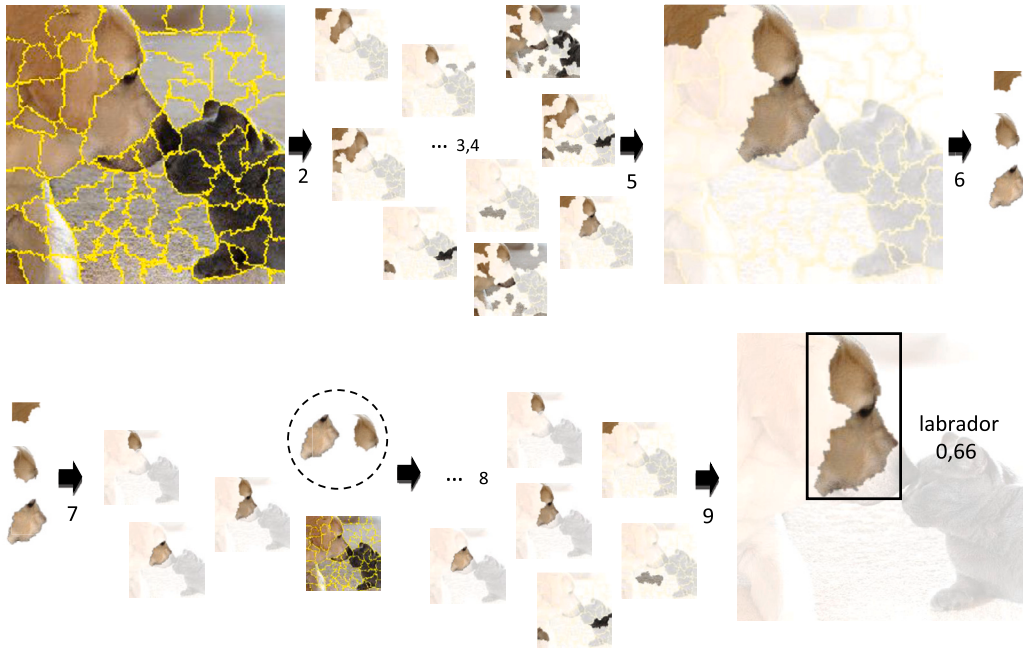


Fig. 4. Building an image prototype from the most accurate reduct (Steps 1-9 of the proposed algorithm) to answer the question: “Why is this good boy a Labrador?” In this example, the class to be explained is “Labrador” from the ImageNet dataset, given as the output of the Inception-v3 model.

6. Obtain the subset of superpixels ($S' \subseteq S_{I_j}$) for the most dissimilar ROI in $[I_x]_{R_1^{o_i}}$.
7. Build combinations of superpixels from S' (from minimum to maximum length) until a candidate reduct is obtained, i.e., until finding the minimum length combination in which the object o_i^* is detected.
8. Build combinations of superpixels from S_{I_j} (from minimum to maximum length) until all reducts are obtained, i.e., until finding the minimum length combination in which the object o_i^* is detected. The maximum length equals the number of superpixels in the candidate reduct. This set of reducts is the multi-reduct $\mathcal{M}_{I_x}(o_i^*)$.
9. Build the prototype image \mathcal{P}_{I_x} from one of the reducts in $\mathcal{M}_{I_x}(o_i^*)$.
 - (a) Select the reduct $\mathcal{R}_{I_x}(o_i^*)$ with the highest detection accuracy inside the multi-reduct $\mathcal{M}_{I_x}(o_i^*)$.
 - (b) Recreate an image \mathcal{P} equivalent to I_x where all superpixels in the image are on.
 - (c) \mathcal{P} is segmented using the same algorithm used in Step 2a.
 - (d) Turn off all superpixels that are not in the selected reduct $\mathcal{R}_{I_x}(o_i^*)$.

From the reducts, the complexity of the explanation is reduced substantially. This happens because the procedure uses superpixels (denoting information granules) instead of pixels when detecting the desired object. Moreover, starting with a candidate reduct from the most dissimilar sample image helps reduce the number of tried combinations among superpixels to find all the reducts.

In Step 2a, other segmentation algorithms can generate a different number of superpixels and regions with other sizes and shapes. With a small number of superpixels, REPROT runs faster, but it may produce a bigger reduct to explain the decision. Conversely, a high number of superpixels slows down the computations, but leads to a smaller reduct as an explanation. In any case, the reduct will always be found, but the amount of contained superpixels and the computational cost will vary.

In Steps 4 and 5, the final goal of the similarity function is finding the most dissimilar ROI in $[I_x]_{R_1^{o_i^*}}$, when compared to the image I_x . The number of superpixels in this ROI will be an upper bound for the number of superpixels in the reducts, which impacts the computational burden by an exponential factor. However, the similarity function will have no impact on the reducts found, and then, the composition of $\mathcal{M}_{I_x}(o_i^*)$ is independent of any similarity function used.

Finally, to better understand the procedure, Fig. 4 shows a general scheme that explains the inference of a black-box model from the most accurate reduct in detecting an object. In addition, the Algorithm 1 presents a pseudocode with more technical details.

5. Experimental study

An experimental study is carried out to analyze the effectiveness of the proposed explanation method based on three deep network architectures: Inception [46], Yolov5 [21], and Mask R-CNN [36,15]. These architectures are widely applied in three computer vision tasks: object classification [17,45], object detection [36,35], and instance segmentation [15].

Algorithm 1 REPROT.

Require: Image of interest I_x , Object o_i^* detected by model \mathcal{B} , Probability to turn on superpixels p

Ensure: Prototype image \mathcal{P}_{I_x}

STEP 2: {Create perturbations to generate a neighborhood H_{I_x} for I_x .}

```

 $H_{I_x} \leftarrow []$ 
 $S_{I_x} \leftarrow \text{Quickshift}(I_x)$ 
 $n \leftarrow \text{total\_superpixels}(S_{I_x})$ 
for  $i$  from 1 to  $\lfloor \frac{n-1}{10} \rfloor$  do
   $MASK \leftarrow \text{generate\_mask}(n, p)$ 
   $ROI \leftarrow \text{apply\_mask}(I_x, MASK)$ 
   $H_{I_x} \leftarrow H_{I_x} + [ROI]$ 
end for

```

STEP 2c): {Obtain the predictions for H_{I_x} as computed by the \mathcal{B} model.}

```

 $PREDICTIONS \leftarrow []$ 
for  $ROI \in H_{I_x}$  do
   $PREDICTIONS \leftarrow PREDICTIONS + [\mathcal{B}(ROI)]$ 
end for

```

STEP 3: {Compute the equivalence class $[I_x]_{R_i^*}$ according to Equation (4).}

```

 $[I_x]_{R_i^*} \leftarrow []$ 
for  $ROI \in H_{I_x}$  do
  if  $PREDICTIONS[ROI]$  is True then
     $[I_x]_{R_i^*} \leftarrow [I_x]_{R_i^*} + [ROI]$ 
  end if
end for

```

STEPS 4 and 5: {Find the most dissimilar ROI in $[I_x]_{R_i^*}$ to I_x using the cosine similarity.}

```

 $ROI_{odd} \leftarrow 1$ 
for  $ROI \in H_{I_x}$  do
   $SIMILARITY \leftarrow \text{cosine\_similarity}(I_x, ROI)$ 
  if  $SIMILARITY < ROI_{odd}$  then
     $ROI_{odd} \leftarrow ROI$ 
  end if
end for

```

STEP 6: {Obtain the subset of superpixels S' from S_{I_x} for the most dissimilar ROI.}

```

 $S' \leftarrow \text{get\_superpixels\_from}(ROI_{odd})$ 

```

STEP 7: {Build combinations of superpixels from S' until finding the minimum length combination in which the object o_i^* is detected.}

```

for  $i$  from 1 to  $|S'|$  do
  for  $SUBSET \in \text{subsets\_by\_cardinality}(S', i)$  do
     $ROI \leftarrow \text{get\_ROI}(SUBSET)$ 
    if  $\mathcal{B}(ROI)$  is True then
       $CANDIDATE \leftarrow SUBSET$ 
      break
    end if
  end for
  if  $\exists CANDIDATE$  then
    break
  end if
end for

```

STEP 8: {Build combinations of superpixels from S_{I_x} until all reducts are obtained. The maximum length is equal to the number of superpixels in the candidate reduct. This set of reducts is the multi-reduct $\mathcal{M}_{I_x}(o_i^*)$.}

```

 $\mathcal{M}_{I_x}(o_i^*) \leftarrow []$ 
for  $i$  from 1 to  $|CANDIDATE|$  do
  for  $SUBSET \in \text{subsets\_by\_cardinality}(S_{I_x}, i)$  do
     $ROI \leftarrow \text{get\_ROI}(SUBSET)$ 
    if  $\mathcal{B}(ROI)$  is True then
       $\mathcal{M}_{I_x}(o_i^*) \leftarrow []$ 
    end if
  end for
  if  $\mathcal{M}_{I_x}(o_i^*) \neq \emptyset$  then
    break
  end if
end for

```

STEP 9: {Build the prototype image \mathcal{P}_{I_x} from one of the reducts in $\mathcal{M}_{I_x}(o_i^*)$.}

```

 $\mathcal{R}_{I_x}(o_i^*) \leftarrow ROI : \text{accuracy}(\mathcal{B}(ROI))$  is the maximum for  $ROI \in \mathcal{M}_{I_x}(o_i^*)$ 
 $\mathcal{P}_{I_x} \leftarrow \text{apply\_mask}(I_x, \mathcal{R}_{I_x}(o_i^*))$ 
return  $\mathcal{P}_{I_x}$  as the prototype image.

```

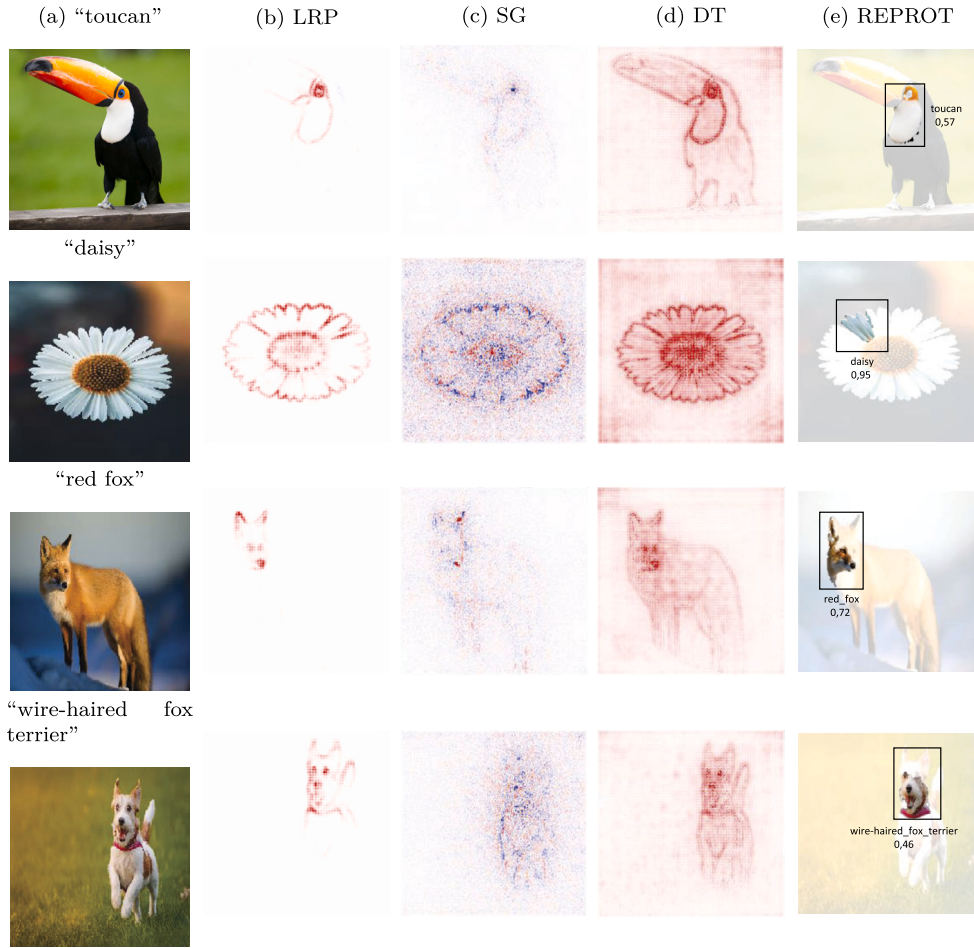


Fig. 5. Explanation of the Inception-v3 predictions by model-specific post-hoc methods: (a) Image of interest, (b) LRP output, (c) SmoothGrad (SG) output, (d) DeepTaylor (DT) output, and (e) REPROT output to explain “toucan”, “daisy”, “red fox”, and “wire-haired fox terrier” as classes detected by the Inception-v3 model. The Inception-v3 model can classify the prototype images as “toucan”, “daisy”, and “red fox” from a single superpixel and as “wire fox terrier” from two superpixels.

This study evaluates how well these deep networks can detect the object to be explained from a prototype image, i.e., how well does an explanation predict? Therefore, the intuition of our experimental framework is driven by the following hypothesis: “If $\mathcal{P}_{I_x}(o_i^*)$ is a prototype of I_x in terms of o_i^* , then o_i^* can be detected from $\mathcal{P}_{I_x}(o_i^*)$.”

5.1. Datasets and pre-trained models

In our experiments, the Inception-v3 model [46] (available in Keras) pre-trained on the ImageNet dataset [11] is used. The publicly available implementations at [21] and [1] are used for Yolov5 and Mask R-CNN, respectively. In both repositories, the MS-COCO dataset [25] is used. It should be noted that, although we have considered these neural architectures in our simulations, our proposal applies to any neural model with similar characteristics, such as the current Yolo variants [47,20].

5.2. Results and discussion

This subsection is divided into two main parts. In the first part, we compare the explanations generated by our proposal with those generated by existing state-of-the-art methods that can explain the inferences made by the Inception model. It is worth mentioning that the selected methods cover both model-specific and agnostic post-hoc methods. In the second part, we experiment with Yolov5 and Mask R-CNN using the explanation generated by our proposal and the LIME method. This design is motivated by the absence of similar state-of-the-art approaches explaining the outputs of these architectures moving beyond agnostic explanations. In addition, it is essential to note that not all existing model-agnostic methods are conducive to dealing with architectures characterized by multiple output layers.

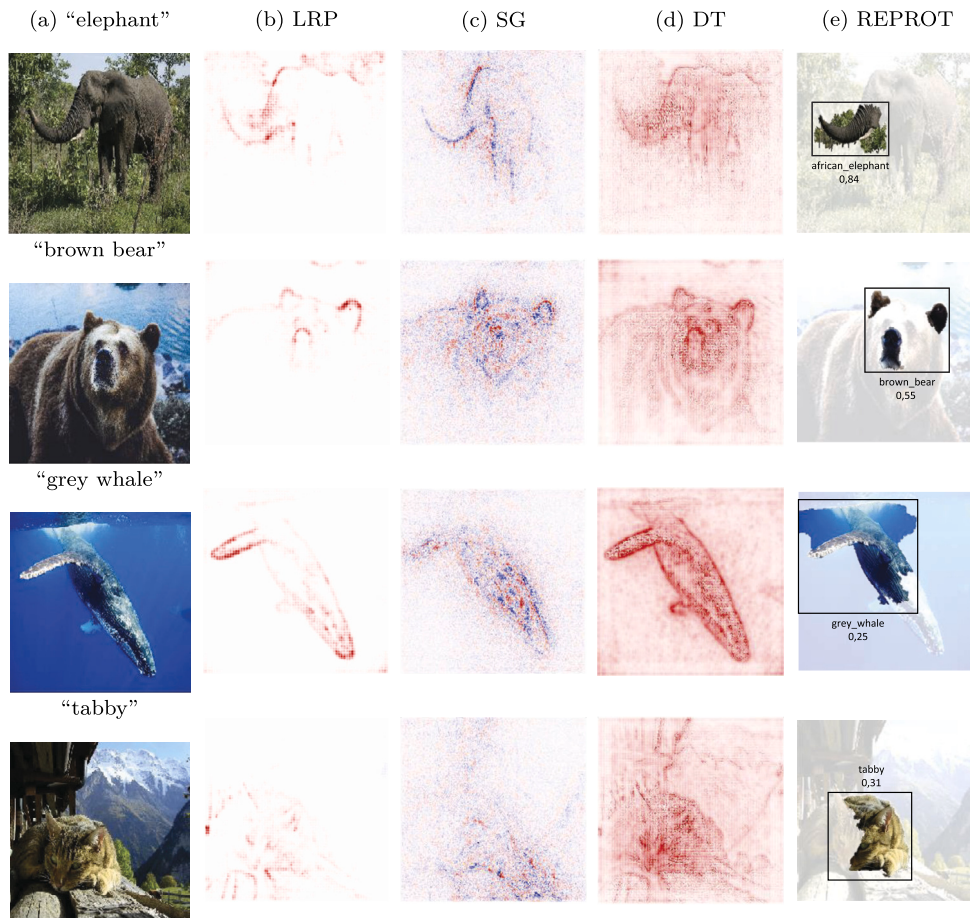


Fig. 6. Explanation of the Inception-v3 predictions by model-specific post-hoc methods: (a) Image of interest, (b) LRP output, (c) SmoothGrad (SG) output, (d) DeepTaylor (DT) output, and (e) REPROT output to explain “African elephant”, “brown bear”, “grey whale”, and “tabby” as classes detected by the Inception-v3 model. The Inception-v3 model can classify the prototype images as “African elephant” and “brown bear” from three superpixels, as “grey whale” from four superpixels, and as “tabby” from seven superpixels.

5.2.1. Explanation of the Inception-v3 predictions

Figs. 5, 6, 7 and 8 show the explanation results for eight images classified as “toucan”, “daisy”, “red fox”, “wire-haired fox terrier”, “African elephant”, “brown bear”, “grey whale”, and “tabby” by an Inception black-box model. In Figs. 5 and 6, columns (b), (c), and (d) represent the saliency maps obtained by the LRP [2], SmoothGrad [44], and DeepTaylor [29] methods, respectively. These are also post-hoc explainability methods, more specifically pixel attribution methods, developed to explain the inferences made by different black-box models such as VGG [43], ResNet [16], and Inception [46] from relevance maps. Their main disadvantage is that they depend on the internal particularities of the neural model, often requiring non-trivial algorithm modifications.

On the other hand, in Figs. 7 and 8, columns (b), (c), (d), and (e) represent the heatmap obtained by the RISE method and the top superpixels detected by ANCHORS, LIME, and the local reducts detected by our method, respectively. It should be stated that, in the case of LIME, the obtained reduct length is used as a visualization threshold, i.e., the first superpixels (as a function of reduct length) whose coefficients in the fitted linear regression model are the highest are shown. While LIME calculates a coefficient for every superpixel within an image, where the coefficient’s magnitude signifies the importance of each superpixel in the trained linear regression model, there exists no predefined threshold dictating the minimum number of superpixels required for generating a prediction.

After applying REPROT, a reduct of length equal to one is used to build the prototypes of the first three images classified as “toucan”, “daisy” and “red fox” by the Inception-v3 model. Longer reducts are used to build the remaining prototypes. The most relevant superpixels suggested by LIME for “brown bear” and “grey whale” are subpar since the Inception-v3 model fails to classify the prototypes from them. A similar result is obtained for the “African elephant” and “tabby” classes, where the Inception-v3 model correctly predicts their decision classes but reports smaller confidence values compared with the reduct approach. However, the outputs of these two methods agree when explaining the “toucan”, “daisy”, “red fox”, and “wire-haired fox terrier” classes, where REPROT and LIME agree on the most important superpixels to perform the inferences of these classes. ANCHORS and LIME have similar behavior when explaining images classified as a “red fox”, “wire-haired fox terrier”, “African elephant”, and “tabby”. However, it cannot correctly explain a “toucan”, “daisy”, or “brown bear”, outperforming LIME when explaining “grey whale”. In

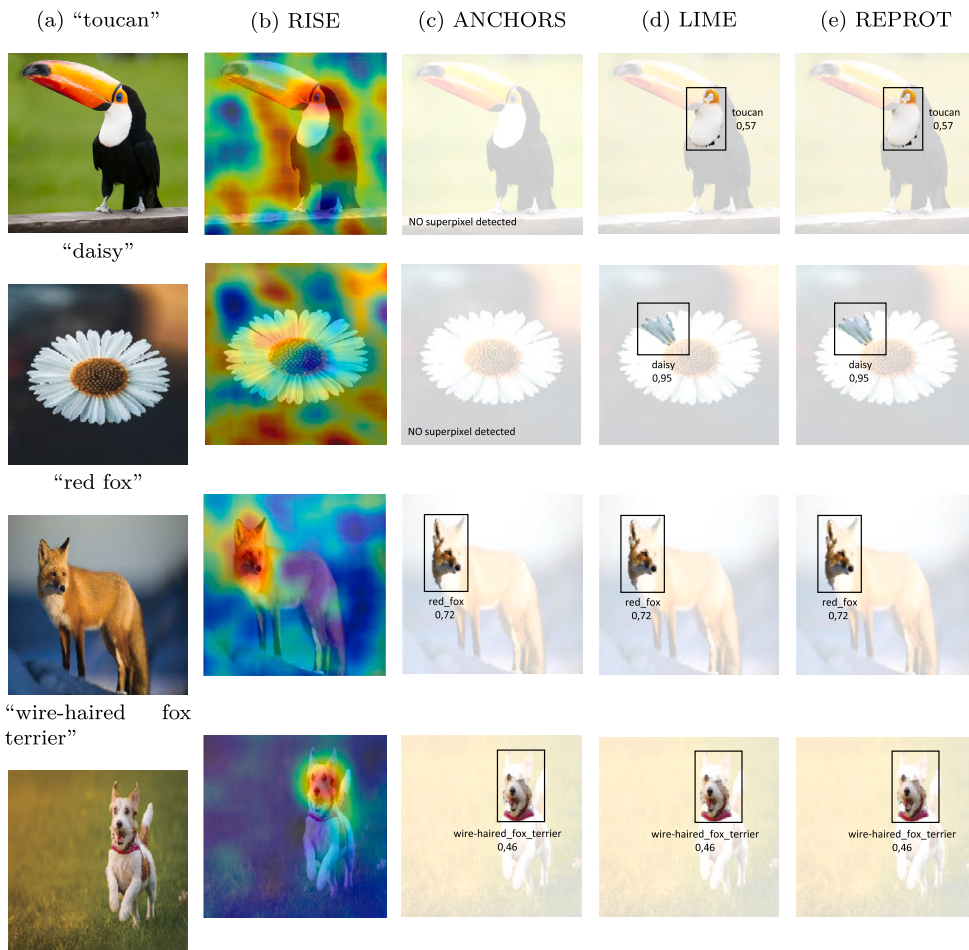


Fig. 7. Explanation of the Inception-v3 predictions by agnostic post-hoc methods: (a) Image of interest, (b) RISE output, (c) ANCHORS output, (d) LIME output, and (e) REPROT output to explain “toucan”, “daisy”, “red fox”, and “wire-haired fox terrier” as classes detected by the Inception-v3 model. The Inception-v3 model can classify the prototype images as “toucan”, “daisy”, and “red fox” from a single superpixel and as “wire fox terrier” from two superpixels.

addition, the relevance maps resulting from applying LRP, DeepTaylor, and RISE provide a similar explanation to that of the obtained reduct. This behavior can be obtained with SmoothGrad but to a lesser extent.

5.2.2. Explanation of the Yolov5 and Mask R-CNN predictions

Figs. 9 and 10 show the output of REPROT and LIME for four images where the objects detected are “person”, “horse”, and “tennis racket”. The figures include (at the top of each subfigure) the superpixels deemed sufficient for detecting objects by the Yolov5 and Mask R-CNN models. For the LIME method, only the number of superpixels matching the length of the corresponding reduct is shown. Superpixels are ordered according to their confidence values as determined by the linear regression model. Our method and LIME agree in their explanations when the “person”, “horse”, and “tennis racket” objects are detected by the Yolov5 model. The same applies to the “tennis racket” and “person” objects (in Fig. 10) recognized by the Mask R-CNN model. However, in the case of the “person” object (in Fig. 9), the LIME method reports the second highest coefficient superpixel, which is not strictly necessary to detect a person by the Mask R-CNN model. Note how the efficacy of the model is deteriorating. Moreover, the five highest coefficient superpixels given by LIME cannot explain the detection of the “horse” object since only one superpixel, i.e., the fourth superpixel, is common with the obtained reduct.

In addition, Table 1 reports the confidence scores (CS) [35] attached to Yolov5 (column 3) and Mask R-CNN (column 4). These scores denote the probability of an object appearing in the box and how well the predicted box fits the object. The rationale behind this experiment is to evaluate the performance of both models in detecting objects in prototype images, which are built from the superpixels given by LIME or the reducts generated by REPROT. In this table, the first column indicates the image of interest, and the second column reports the objects whose inference will be explained. The third and fourth columns report a) the number of superpixels composing the reduct, b) the CS of the inference performed for a prototype built from the reduct of an image, and c) whether the object is detected from a prototype built from the most important superpixels given by LIME. In the case of LIME, we

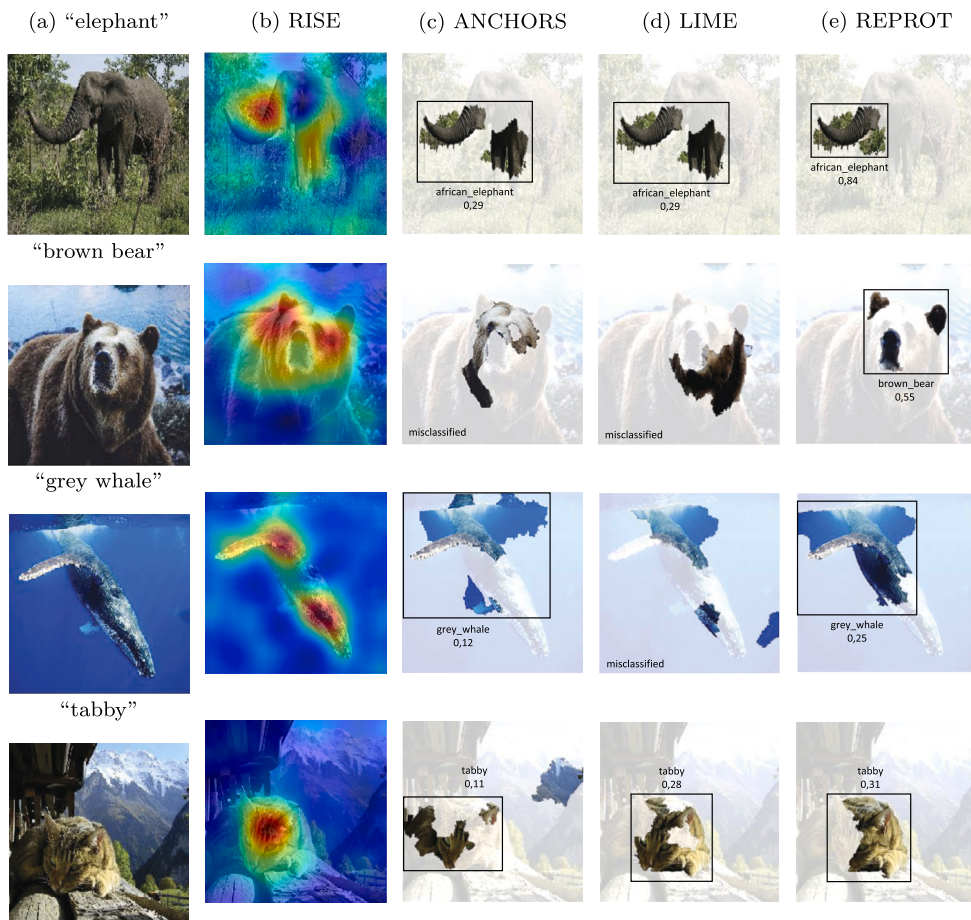


Fig. 8. Explanation of the Inception-v3 predictions by agnostic post-hoc methods: (a) Image of interest, (b) RISE output, (c) ANCHORS output, (d) LIME output, and (e) REPROT output to explain "African elephant", "brown bear", "grey whale", and "tabby" as classes detected by the Inception-v3 model. The Inception-v3 model can classify the prototype images as "African elephant" and "brown bear" from three superpixels, as "grey whale" from four superpixels, and as "tabby" from seven superpixels.

also report in parentheses the number of superpixels that, according to this method, are necessary for the neural model to detect the object.

The LIME and REPROT outputs agree that one or two superpixels are sufficient for the Yolov5 model to detect "person", "horse", "tennis racket", "zebra", and "traffic light" objects. On the contrary, in some cases, a larger number of superpixels are needed to detect them by the Mask R-CNN model (e.g., "person"). However, in the specific case of the "horse", the five superpixels with the highest coefficient detected by LIME are insufficient to detect the object. In addition, the "bed", "chair", "couch", "snowboard", "giraffe", "car", "motorcycle", "bus", and "dog" objects cannot be detected by the Yolov5 and Mask R-CNN models from the first superpixels (i.e., considering the reduct's length) resulting from applying the LIME method. In fact, this is also true for the "person" object in the seventh image (from top to bottom) and the "potted plant" object if the inference is made through the Yolov5 model. In other words, the first superpixels with larger coefficients given by LIME are not determinant enough to detect those objects by the neural model. This is further reaffirmed in those cases labeled by $(10 <)$, where even considering the first ten most important superpixels resulting from applying LIME, it is impossible to detect the object.

5.2.3. Evaluation of the quality of explanations

The accuracy of an explanation is a computational measure according to [27,28,18] designed to evaluate "how well does an explanation predict?". That is, high accuracy is especially important if the explanation is used for predictions in place of the machine learning model. Low accuracy can be acceptable if the accuracy of the machine learning model is also low and if the goal is to explain what the black-box model does. In this sense, we evaluate how well a black-box model can detect the object to be explained from a prototype image (i.e., REPROT output) compared to the output of other state-of-the-art methods.

Fig. 11 shows the mean accuracy obtained by each of the agnostic approaches on Inception, Yolo, and Mask R-CNN inferences from the COCO and ImageNet datasets. In particular, we rely on the output (i.e., superpixel-based) of those methods that by their agnostic nature can be applied to these types of architectures such as, REPROT, LIME, and ANCHORS.

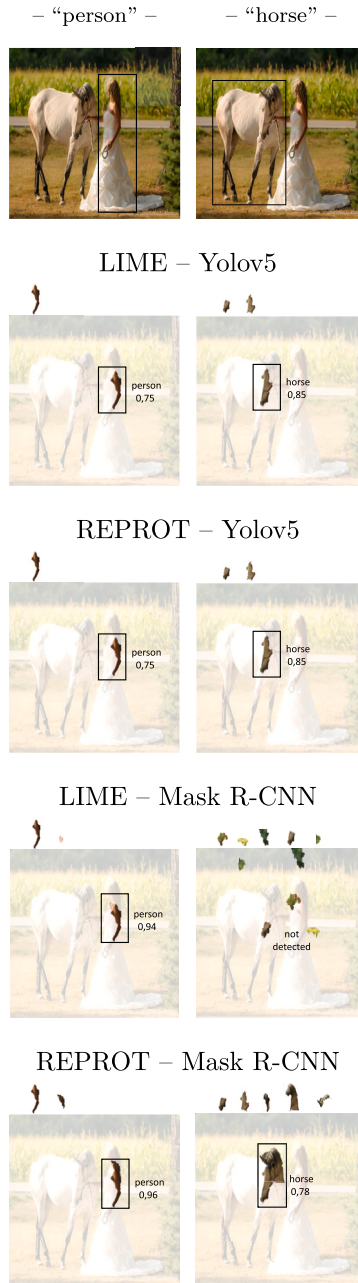


Fig. 9. Results of the explanation methods. The column represents the “person” and “horse” objects to be detected by the black-box models. The first row collects the image and the object of interest. The second and fourth rows include the LIME outputs to explain the inferences of the Yolov5 and Mask R-CNN models, respectively. The third and fifth rows include the prototypes to explain the inferences of the Yolov5 and Mask R-CNN models, respectively.

Observe how, the accuracy of our method is better than that of the other agnostic methods. REPROT not only outperforms LIME and ANCHORS, but also obtains similar results to when the inference process is performed with the full image. The latter is a highly desirable property in prototype-based classification approaches.

The *Kullback-Leibler divergence* [23], often referred to as relative entropy or I-divergence, and denoted $D_{KL}(P||Q)$ (see Equation (5)), serves as a statistical distance. It quantifies the dissimilarity between two probability distributions: the target distribution P and the reference distribution Q . In simple terms, D_{KL} measures the average additional surprise one can expect when employing Q as a model when the actual distribution is P .

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}. \tag{5}$$

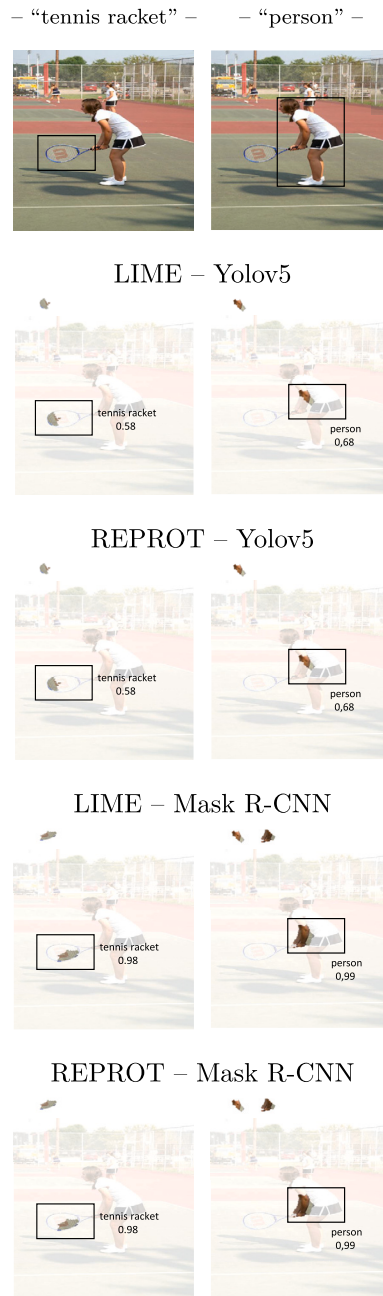


Fig. 10. Results of the explanation methods. The column represents the “tennis racket” and “person” objects to be detected by the black-box models. The first row collects the image and the object of interest. The second and fourth rows include the LIME outputs to explain the inferences of the Yolov5 and Mask R-CNN models, respectively. The third and fifth rows include the prototypes to explain the inferences of the Yolov5 and Mask R-CNN models, respectively.










D_{KL} allows us to assess the impact of the selected superpixels on classification accuracy. A high D_{KL} indicates that the chosen subset of superpixels significantly affects the classification results, while a low D_{KL} suggests minimal impact. In this sense, we aim to contrast the divergence values obtained for every model (Yolov5, Mask R-CNN, and Inception-v3) when comparing the distribution of accuracy values using the entire image and the superpixels detected by each method (REPROT, LIME, and ANCHORS). It must be noticed again that ANCHORS only applies to the Inception model. Additionally, we use the same images as in the previous experiments.

Fig. 12 shows that REPROT effectively presents the minimum D_{KL} value compared to the entire image. Therefore, the superpixels detected by our method produce the slightest difference in information or accuracy compared to those from LIME and ANCHORS. In other words, REPROT minimizes the information lost to approximate the model accuracy using whole images.

Before concluding our paper, it seems relevant to stress the advantages of the proposed explanation algorithm over LIME:

Table 1

Confidence scores of the Yolov5 and Mask R-CNN models derived from the inference of the prototypes built on the most important superpixels given by LIME and REPROT. In this table, the column associated with LIME also reports the number of superpixels that it estimates necessary to detect the object, where (10 <) means that the first 10 superpixels that LIME obtains as the most important are not determinant for detecting the object.

Image	Object	Yolov5			Mask R-CNN		
		REPROT #Superpixels	CS	LIME Detected or not?	REPROT #Superpixels	CS	LIME Detected or not?
	person	1	0.75	yes	2	0.96	yes
	horse	2	0.85	yes	5	0.78	no (10 <)
	person	1	0.68	yes	2	0.99	yes
	tennis racket	1	0.58	yes	1	0.98	yes
	potted plant	3	0.47	no (4)	2	0.9	yes
	bed	3	0.42	no (10)	6	0.77	no (10 <)
	chair	3	0.4	no (10 <)	3	0.93	no (10 <)
	couch	2	0.31	no (4)	3	0.84	no (10)
	person	1	0.55	yes	1	0.99	yes
	snowboard	1	0.52	no (3)	2	0.9	no (10 <)
	giraffe	2	0.46	no (10)	2	0.96	no (4)
	zebra	1	0.84	yes	1	0.91	yes
	car	1	0.71	no (3)	2	0.98	no (5)
	traffic light	1	0.67	yes	1	0.99	yes
	motorcycle	2	0.63	no (3)	2	0.84	no (3)
	person	2	0.63	no (5)	2	0.97	no (10 <)
	bus	3	0.62	no (8)	10	0.82	no (10 <)
	traffic light	1	0.74	yes	1	0.99	yes
	person	1	0.73	yes	1	0.99	yes
	dog	2	0.59	no (3)	2	0.99	no (3)

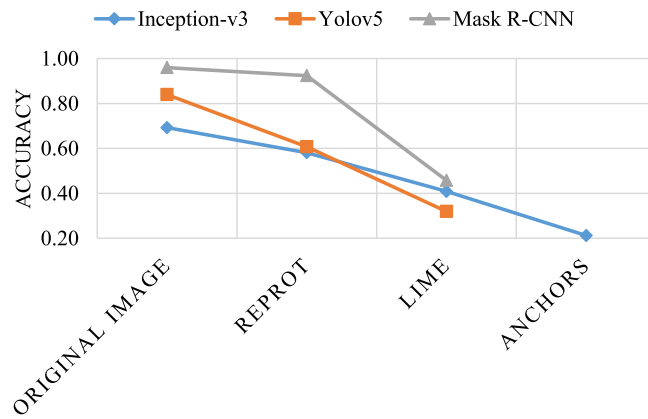


Fig. 11. Mean accuracy of Inception-v3, Yolov5, and Mask R-CNN models derived from the inference obtained from the prototypes built on the most important superpixels given by REPROT, LIME, and ANCHORS. Note that, the latter is not applicable to architectures such as Yolov5 and Mask R-CNN.

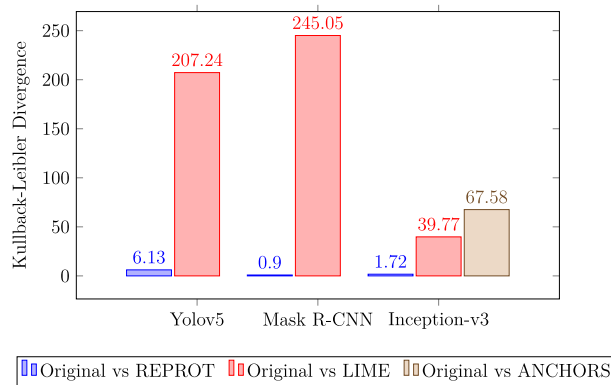


Fig. 12. Kullback-Leibler divergence for different models and distributions of accuracy values.

- No threshold importance is required. More explicitly, LIME produces linear regression coefficients for superpixels, which are a proxy to determine their importance in detecting the object. However, how to know which subset of the superpixel set is necessary to detect an object? In the numerical simulations, the length of the obtained reduct is assumed to be the importance threshold.
- The explanations do not rely on a local surrogate model, such as a linear regression model. The proposed explanation algorithm returns a prototype image with superpixels deemed sufficient to detect a given object using a black-box model.
- LIME does not always succeed in obtaining the most important superpixels for a black-box model to detect the object. However, from a reduct, it will always be possible to build a prototype image in which the object to be explained is detected.

6. Conclusions

This paper proposed REPROT, a method to explain the detection of an object in a given image by a complex black-box architecture. To support our findings in the image processing domain, extending the definitions of information systems and reducts in RST is crucial. The explanation is based on building a prototype image representing the minimum set of superpixels sufficient to detect a given object, i.e., a local reduct of the image. The prototype construction is done from the superpixels present in the most accurate local reduct from the multi-reduct computed for an image. The experiments performed on different complex deep learning architectures illustrated that the black-box model could detect the object with the generated prototype. Additional relevant remarks are given below:

- The local reduct of an image only gives information about the superpixels that must be present in this image so that the model can detect a given object. Moreover, it is worth mentioning there could be a combination of superpixels having a larger length (including the reduct) such that the object is also detected.
- According to the definition of multi-reduct, multiple local reducts can be obtained when there is more than one combination of superpixels with the same length from which the black-box model detects an object. This suggests that more than one prototype image could be built to explain the output of a model.

Experimental results show the superiority of our proposal over other post-hoc approaches existing in the literature. Its advantage over model-specific approaches lies in its agnostic nature, broadening its applicability spectrum. At the same time, it obtains comparable results. In some cases, it even outperforms other agnostic approaches, such as ANCHORS or LIME, in explaining single-layer output neural models, such as Inception, and multi-layer output models, such as Yolo and Mask R-CNN.

The future research efforts will focus on integrating the theoretical RST-based formalism presented in this paper with the symbolic explanation module proposed in [30] to generate counterfactual explanations on images. Such integration can be done in three steps. Firstly, we would need to create a symbolic dictionary of relevant local reducts and their prototypes that guide the classification process. Secondly, the Prolog-powered reasoning module would generate a counterfactual image containing altered reducts that lead to an alternative outcome. Finally, we would need to retrieve the image from the dataset that better resembles the counterfactual image generated by the symbolic reasoning module.

CRedit authorship contribution statement

Marilyn Bello: Conceptualization, Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Gonzalo Nápoles:** Methodology, Writing – review & editing. **Leonardo Concepción:** Formal analysis, Investigation. **Rafael Bello:** Conceptualization, Writing – review & editing. **Pablo Mesejo:** Project administration, Supervision. **Óscar Gordón:** Project administration, Supervision.

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

Data will be made available on request.

Acknowledgement

Research funded by MCIN/AEI/10.13039/501100011033/ and FEDER “Una manera de hacer Europa” under grant CONFIA (PID2021-122916NB-I00).

References

- [1] W. Abdulla, Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow, Code repository https://github.com/matterport/Mask_RCNN, 2017.
- [2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, *PLoS ONE* 10 (2015) e0130140.
- [3] A. Barredo, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion* 58 (2020) 82–115.
- [4] M. Bello, Y. Aguilera, G. Nápoles, M.M. García, R. Bello, K. Vanhoof, Layer-wise relevance propagation in multi-label neural networks to identify Covid-19 associated coinfections, in: *International Workshop on Artificial Intelligence and Pattern Recognition*, Springer, 2021, pp. 3–12.
- [5] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, *arXiv preprint, arXiv:2102.13076*, 2021.
- [6] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, *Data Min. Knowl. Discov.* (2023) 1–60.
- [7] R. Brinkmann, *The Art and Science of Digital Compositing: Techniques for Visual Effects, Animation and Motion Graphics*, Morgan Kaufmann, 2008.
- [8] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J.K. Su, This looks like that: deep learning for interpretable image recognition, *Adv. Neural Inf. Process. Syst.* 32 (2019) 8930–8941.
- [9] H. Chen, S. Lundberg, S.I. Lee, Explaining models by propagating Shapley values of local components, in: *Explainable AI in Healthcare and Medicine*, Springer, 2021, pp. 261–270.
- [10] P. Cortez, M.J. Embrechts, Using sensitivity analysis and visualization techniques to open black box data mining models, *Inf. Sci.* 225 (2013) 1–17.
- [11] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 248–255.
- [12] R. ElShawi, Y. Sherif, M. Al-Mallah, S. Sakr, ILIME: local and global interpretable model-agnostic explainer of black-box decision, in: *European Conference on Advances in Databases and Information Systems*, Springer, 2019, pp. 53–68.
- [13] W.H. Gomaa, A.A. Fahmy, et al., A survey of text similarity approaches, *Int. J. Comput. Appl.* 68 (2013) 13–18.
- [14] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, *Data Min. Knowl. Discov.* (2022) 1–55, <https://doi.org/10.1007/s10618-022-00831-6>.
- [15] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2961–2969.
- [16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [17] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: *International Conference on Machine Learning*, PMLR, 2015, pp. 448–456.
- [18] M.R. Islam, M.U. Ahmed, S. Barua, S. Begum, A systematic review of explainable artificial intelligence in terms of different application domains and tasks, *Appl. Sci.* 12 (2022) 1353.

- [19] M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: a survey, *Pattern Recognit. Lett.* 150 (2021) 228–234.
- [20] G. Jocher, A. Chaurasia, J. Qiu, Yolo by ultralytics, Code repository <https://github.com/ultralytics/ultralytics>, 2023.
- [21] G. Jocher, K. Nishimura, T. Mineeva, R. Vilarinho, yolov5, Code repository, <https://github.com/ultralytics/yolov5>, 2020.
- [22] I. Kakogeorgiou, K. Karantzas, Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing, *Int. J. Appl. Earth Obs. Geoinf.* 103 (2021) 102520.
- [23] S. Kullback, R.A. Leibler, On information and sufficiency, *Ann. Math. Stat.* 22 (1951) 79–86.
- [24] B. Letham, C. Rudin, T.H. McCormick, D. Madigan, Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model, *Ann. Appl. Stat.* 9 (2015) 1350–1371.
- [25] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: *European Conference on Computer Vision*, Springer, 2014, pp. 740–755.
- [26] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: *Proceedings of the International Conference on Neural Information Processing Systems*, ACM, 2017, pp. 4768–4777.
- [27] S. Mohseni, N. Zarei, E.D. Ragan, A survey of evaluation methods and measures for interpretable machine learning, arXiv preprint, arXiv:1811.11839, 2018.
- [28] S. Mohseni, N. Zarei, E.D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, *ACM Trans. Interact. Intell. Syst.* 11 (2021) 1–45.
- [29] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.R. Müller, Explaining nonlinear classification decisions with deep Taylor decomposition, *Pattern Recognit.* 65 (2017) 211–222.
- [30] G. Nápoles, F. Hoitsma, A. Knobens, A. Jastrzebska, M. Leon Espinosa, Prolog-based agnostic explanation module for structured pattern classification, *Inf. Sci.* 622 (2023) 1196–1227, <https://doi.org/10.1016/j.ins.2022.12.012>.
- [31] Z. Pawlak, Rough sets, *Int. J. Comput. Inf. Sci.* 11 (1982) 341–356.
- [32] Z. Pawlak, Rough sets and intelligent data analysis, *Inf. Sci.* 147 (2002) 1–12.
- [33] V. Petsiuk, A. Das, K. Saenko, RISE: randomized input sampling for explanation of black-box models, arXiv preprint, arXiv:1806.07421, 2018.
- [34] V. Pillai, H. Pirsiavash, Explainable models with consistent interpretations, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI, 2021, pp. 2431–2439.
- [35] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2016, pp. 779–788.
- [36] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: *Proceedings of the International Conference on Neural Information Processing Systems*, ACM, 2015, pp. 91–99.
- [37] M.T. Ribeiro, S. Singh, C. Guestrin, “Why should I trust you?” Explaining the predictions of any classifier, in: *International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 1135–1144.
- [38] M.T. Ribeiro, S. Singh, C. Guestrin, Anchors: high-precision model-agnostic explanations, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
- [39] S. Sattarzadeh, M. Sudhakar, A. Lem, S. Mehryar, K.N. Plataniotis, J. Jang, H. Kim, Y. Jeong, S. Lee, K. Bae, Explaining convolutional neural networks through attribution-based input sampling and block-wise feature aggregation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI, 2021, pp. 11639–11647.
- [40] J.H. Sejr, P. Schneider-Kamp, N. Ayoub, Surrogate object detection explainer (SODEx) with YOLOv4 and LIME, *Mach. Learn. Knowl. Extr.* 3 (2021) 662–671.
- [41] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE Conference on Computer Vision*, IEEE, 2017, pp. 618–626.
- [42] L. Shen, H. Tao, Y. Ni, Y. Wang, V. Stojanovic, Improved YOLOv3 model with feature map cropping for multi-scale road object detection, *Meas. Sci. Technol.* 34 (2023) 045406.
- [43] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint, arXiv:1409.1556, 2014.
- [44] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, SmoothGrad: removing noise by adding noise, arXiv preprint, arXiv:1706.03825, 2017.
- [45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2015, pp. 1–9.
- [46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2016, pp. 2818–2826.
- [47] J. Terven, D. Cordova-Esparza, A comprehensive review of YOLO: from YOLOv1 to YOLOv8 and beyond, arXiv preprint, arXiv:2304.00501, 2023.
- [48] M. Tulio Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, arXiv:1606.05386, 2016.
- [49] K. Uehara, M. Murakawa, H. Nosato, H. Sakanashi, Prototype-based interpretation of pathological image analysis by convolutional neural networks, in: *Asian Conference on Pattern Recognition*, Springer, 2019, pp. 640–652.
- [50] A. Vedaldi, S. Soatto, Quick shift and kernel methods for mode seeking, in: *European Conference on Computer Vision*, Springer, 2008, pp. 705–718.
- [51] C. Wang, A. Bochkovskiy, H. Liao, YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, arXiv preprint, arXiv:2207.02696, 2022.
- [52] Y. Yang, V. Tresp, M. Wunderle, P.A. Fasching, Explaining therapy predictions with layer-wise relevance propagation in neural networks, in: *IEEE International Conference on Healthcare Informatics*, IEEE, 2018, pp. 152–162.
- [53] Y. Yao, Information granulation and rough set approximation, *Int. J. Intell. Syst.* 16 (2001) 87–104.