Smart Agricultural Technology 12 (2025) 101286

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Enhancing explainability in pacu fish image segmentation using saliency maps and combined explainable AI methods

Juliana da C. Feitosa [a,*], Fabrício M. Batista [a], Juliana C.F. Catharino [a], Milena V. Freitas [b], Diogo T. Hashimoto [b], João Paulo Papa [a], José Remo F. Brega [a]

[a] Department of Computing, São Paulo State University (UNESP), Bauru, SP, Brazil
[b] Aquaculture Center of Unesp, São Paulo State University (UNESP), Jaboticabal, SP, Brazil

ARTICLE INFO

Dataset link: https://github.com/jufeitosa10/XAI_Segmentacao_Pacu.git

Keywords:
Aquaculture
Artificial intelligence
EXplainable artificial intelligence
Evaluation
Pixels perturbation

ABSTRACT

Advances in Artificial Intelligence (AI) have sparked concerns regarding the transparency of model outputs, necessitating the development of eXplainable Artificial Intelligence (XAI) techniques. This study presents an improved evaluation of XAI methods applied to Pacu fish image segmentation. We compare and evaluate four XAI methods - Grad-CAM, Saliency Map, CNN Filters, and Layer Grad-CAM - using pixel perturbation techniques applied in 100 fish images. Our experiments reveal that through the images generated by the pixel perturbation techniques (118 to the white noise, 144 to the dark noise, and 123 to the random noise), the Saliency Map achieves the best results, highlighting the most relevant regions for AI model predictions. Furthermore, by combining Saliency Maps with other XAI methods, we demonstrate substantial improvements in explainability and segmentation accuracy. In this case, the largest number of images generated in the second experiment was 108, by combining the between Saliency Map and Grad-CAM to the white noise pixel perturbation. Finally, this work not only advances the state-of-the-art in XAI for image segmentation but also underscores the importance of combining XAI techniques to achieve superior explanatory power in areas such as Aquaculture.

1. Introduction

The advancement of research focused on eXplainable Artificial Intelligence (XAI) [13,55], as well as the emergence of new interpretable solutions [20], has contributed to the increase in evaluation methods that allow us to understand the numerous existing techniques, in order to avoid meaningless solutions and results [9]. Despite this advancement, explainability methods still cannot provide guarantees that an individual decision is correct, since the evaluation of explainable results is mostly non-existent. Another problem is the lack of guarantee that the explanations resulting from XAI techniques match the reality presented by the prediction process, as well as the difference in the behavior of these methods when subjected to global and local problems. Moreover, explanations that are understandable to human beings are currently sought, which can be safely used in decision-making, especially in health-related contexts [16].

The visualizations produced by XAI methods can often help data scientists discover problems in data sets or AI models [27]. However, such visualizations can lead to overconfidence by the user, along with misuse of these methods. In image segmentation, where the aim is to highlight one or more regions of interest, the need for a good explanation is even greater, since the aim is to understand which pixels are most relevant to the decision-making of the model [33]. [51] presents an example of this need in Aquaculture, where disease or anomaly detection can be accomplished by applying XAI methods that provide information about critical areas in images, influencing AI models' decisions. Therefore, combining methods such as Grad-CAM and Confusion Matrix increase model interpretability, ensuring acceptable qualitative and quantitative results.

The main problem investigated here concerns the lack of guarantee regarding the explanations during the prediction of the AI model applied to image segmentation. One strategy involves understanding explainable techniques, analyzing the influence of external agents on their results, and improving explanations based on the analysis performed. This paper focuses on the analysis of the explainability of XAI methods applied to Pacu fish image segmentation. The explainable methods were subjected to adverse situations, such as pixel disturbance, to find the best method based on its prediction. Experiments were implemented in an attempt to improve the results of the XAI methods based on the

* Corresponding author.
E-mail address: juliana.feitosa@unesp.br (J. da C. Feitosa).

https://doi.org/10.1016/j.atech.2025.101286
Received 28 February 2025; Received in revised form 3 August 2025; Accepted 4 August 2025

Available online 8 August 2025

best explainable technique found. The main contributions of this study include:

- A systematic literature review on the evaluation and implementation of enhancements to XAI methods;
- The implementation of explainable methods for segmenting images of Pacu fish;
- The development of an evaluation technique using pixel disturbance for XAI methods;
- Identifying the optimal XAI method for the proposed scenario; and
- Refining existing XAI methods based on the best method identified.

This paper is organized as follows: Section 2 presents the studies resulting from the systematic review, and the Section 3 presents the fundamental concepts used. Section 4 presents the methodology, and Section 5 discusses the experiments. The section 6 presents the results and a critical analysis of the experiments. Finally, Section 7 states conclusions, limitations and future work.

2. Systematic literature review

The findings from the review process are organized based on the researched questions. The analysis focused on the number of questions addressed in the examined documents, ending up in 31 documents from 2019 to 2023, providing answers to the research inquiries. The questions aim to understand current evaluation techniques for explainable methods, such as the use of pixel perturbation to analyze the explainability of these methods, further to existing ways to improve the explanations generated. The findings have been extracted and are outlined below.

2.1. Q1: Can we assess whether XAI methods explain what happened in the AI model's prediction?

Much attention has been given to enhancing confidence in AI models. The importance of aligning explanations from eXplainable AI methods with human-understandable explanations highlights the effort to boost trust in such systems [7]. Therefore, there has been a notable increase in the quest for techniques to evaluate explainability, particularly in contexts necessitating completely reliable explanations such as the analysis of medical images [43,25,42]. Across the materials scrutinized in this assessment, the majority proposed methodologies for appraising the explainability of XAI approaches. Consequently, the focus on generating high-quality explanations has broadened to encompass diverse domains, incorporating interdisciplinary evaluation approaches [35] involving disciplines such as Psychology, Philosophy, and Social Sciences [54].

The assessment of explainability relies on various dimensions that define its quality, including fidelity, comprehensibility, robustness, and complexity [53]. Alongside computational measures, there are human-centric metrics designed to gauge explanation quality by considering laypeople's trust in the topic. These metrics fall into two categories: subjective metrics gauging human trust in explanations, and objective metrics assessing human behavioral state and task performance [22].

Different criteria and perspectives have led to the presentation of varied methods and metrics for assessing explainability. Categorization was performed based on the evaluation aspect, measures used, and construction method, offering an overview of evaluative solutions for explainable AI. The evaluation aspect can be divided into mental models, effectiveness and satisfaction, trust and dependence, human-AI performance, and functional evaluation. Currently, AI regulatory bodies universally acknowledge the necessity of rigorously evaluating automated explanations. Therefore, the findings in this study underscore the significance of evaluating XAI methods in this research.

Currently, all regulatory bodies focused on AI agree on the need to carefully evaluate the quality of automated explanations [2] [32]. Thus, the results presented for this issue support the importance of evaluating the XAI methods applied to this present research.

2.2. Q2: Are XAI methods influenced by the disturbance of pixels in the model's input images?

Disturbances allow us to examine the relationship between the input and output of a model, making it possible to observe which part of the input a model attaches greater importance to [23]. The relationship between explainable methods and the different perturbation techniques can be described in several ways. The first is through the classification of XAI methods, which can be divided into two types: gradient and perturbation-based. Thereby, methods such as Occlusion have as their main characteristic the perturbation of the input to define whether the AI model identifies, during the inference process, the most relevant points [40]. As a result, most of the studies analyzed in this review consider pixel perturbation to be the basis of an explainable method.

An additional option has been identified in studies utilizing input perturbation to assess the accuracy of AI models [26] and the interpretability of explainable techniques [34]. In both scenarios, demonstrating an impact from the modified inputs validates the tool's effectiveness. For explainable methods, if this impact is not evident, it suggests that the explanations produced do not align with the model's actual inference process [19].

The use of the perturbation techniques extends to scenarios other than classification [31] or image segmentation [17]. There is research in which this technique is considered for input to time series models, and is even used as a way of evaluating the explainability of XAI methods applied to this type of model [44,52,1]. It can thus be seen that input perturbation is a commonly used technique and helps to detect flaws in explainability, or even in the AI model itself.

For image-based models, perturbation techniques enable the identification of various noise types in different image categories [46]. In practical contexts like medical imaging, diverse noise sources can arise during image acquisition, leading to challenges for interpretable methods in pinpointing the most relevant regions [36]. Consequently, certain scenarios may not be adequately addressed in model training, potentially resulting in erroneous predictions.

Disturbance techniques can help detect whether the prediction process will be carried out correctly in the face of these changes. Furthermore, when explainability methods are applied, the result should be in line with the disturbance of the most relevant pixels [36]. In other words, when these are modified, explainability should be influenced. Therefore, according to the answers found to this research question, it can be said that the explainability method with good performance should indeed be influenced by the disturbance made to the AI model input, including images and visual explanations.

2.3. Q3: How can XAI methods be improved to provide explanations that are consistent with the prediction of the AI model?

The last research question refers to the improvements in XAI methods, the aim of which is to improve the quality of explainability. This question helps to distinguish the difference between the methodology applied in this research and what has been found in the literature. Moreover, the search for more reliable explanations has motivated researchers in various fields (e.g., Medicine) to look for techniques that improve the explanations provided by XAI methods [47]. Therefore, although the issue is considered pertinent in the current scenario, only nine of the 31 articles analyzed presented suggestions for improvement.

Of the solutions presented, the most common is related to considerable modifications to existing XAI methods [28,58]. Some of the research found shows modifications more than one method [19], such as the development of an end-to-end explainability approach for image subtitling architectures based on Bottom-Up (BU) visual resources [11].

The LIME explainable method appears in several research projects related to improving explainability. The EvEx method, for instance, has been developed as a new explainable AI that uses explanations provided by LIME and which has been combined with a multi-objective genetic algorithm [50]. In another example found, the authors propose an automatic Ensemble-based Genetic Algorithm Explainer (EGAE), whose function is to improve the explainability of LIME [38]. Furthermore, the CAM method is one of the methods that undergoes the most changes. For example, in one of the studies analyzed, the authors present another variation called SA-CAM, which uses the self-attention mechanism in its methodology [30].

Another proposed improvement is the integration of XAI techniques to enhance explainability. For example, combining Grad-CAM, Integrated Gradient, Gradient Input, and Occlusion methods for generating multiple visual maps has been suggested [6]. These visualizations were evaluated for pest detection in rural areas, with the overlap measurement between image pairs conducted using AND, OR, and XOR operators. The findings underscore the importance of enhancing explainable methods.

The results presented reinforce the ideology of seeking to improve explainable methods. Furthermore, one can observe that according to the research analyzed during the review, most of the suggested improvements are based on specific modifications to existing methods. However, only one document suggests combining methods to improve explainability. Finally, it can be stated that the methodology used in this paper is consistent with the ideology proposed for explainable methods, and the proposed issue was clarified by RSL.

2.4. Discussion

During the review process, only the study by [19] addressed all three research questions. Although similar, the responses analyzed led to a distinct proposal, thus not diminishing the significance of the current study. Most of the documents presented answers analogous or similar to what was proposed in this paper for questions 1 and 2. Therefore, one can observe that evaluating XAI methods using pixel perturbation is a correct approach supported by the literature.

Unlike other works, a noticeable disparity lies in the third research question focusing on enhancing explainable methods. The array of approaches for explainability enhancements in the nine documents varies, each emphasizing efficacy in explanations. However, this study aimed to enhance XAI methods by selecting the optimal approach for a given scenario and then integrating it with other methods in use. The study most akin to our proposition was conducted by [6], who explored combinations of local and global XAI methods to enhance. Despite the authors' evaluation of method explainability quality, they did not combine the best method with others to achieve superior explanatory outcomes. Essentially, their proposed combination involved methods selected for the study without regard to their comparative efficacy. The subsequent sections explored into whether our study's methodology goes beyond enhancing methods deemed inferior, aspiring instead to elevate the best XAI method to generate superior results.

3. Theoretical background

3.1. XAI methods

To improve interpretability and reliability, the use of XAI techniques has been considered in several areas of study, e.g., medicine, where the need to understand the model's decisions is essential for professionals in the field. For neuroradiologists [37], for example, this transparency is invaluable because it allows for greater confidence in the AI model's outputs, allowing for the correct diagnosis of brain tumors through magnetic resonance imaging. In another example, researchers use XAI to identify content that may have adverse effects on video game players [60]. Due to this increased interest in explainability, ways to classify these methods have emerged in order to aid in their implementation. Therefore, an XAI model can be classified according to the explanations provided, which can be spoken or visualized [55]. However, in a more complete way, DARPA's XAI program presents four modes of explanation that facilitate the user's decision [20]:

- **Analytical statements:** are stated in natural language to describe the elements and context that support a decision;
- **Visualizations:** directly highlight the parts of the data that support a decision and allow users to form their own perceptual understanding;
- **Cases:** use of specific examples or stories that help in decision-making; and
- **Rejections of alternative choices:** arguing against certain biased responses on the basis of analysis, cases, and data.

In this context, one can observe the conversion of analytical statements into visualizations, as well as the use of this area as a tool to generate explanations to users [10]. Concerning this work, four explainable methods were applied, whose results are based on visualization techniques: Grad-CAM [45], Layer Grad-CAM [5], Saliency Map [48], and CNN Filters [12].

In an attempt to better understand CNNs (Convolutional Neural Network), several methods have appeared in the literature for visualizing this internal representation, e.g., Class Activation Maps (CAM) [61], the aim of which is to use Global Average Pooling (GAP) in the CNN. CAM combines the activation maps A of the convolution layer, which contains K filters, in addition to the weights $w_{k,c}$ a fully connected layer (FC), whose values *(k,c)* represent the specific weighted connection between the convolution layer and the FC used to create the relevance map score [40], as follows:

$$map_c = \sum_k^K w_{k,c} A_k. \tag{1}$$

Grad-CAM is a variation of the CAM method using the gradients of the network's output regarding the last convolutional layer of the CNN to obtain the class activation map. This allows Grad-CAM to be implemented in more types of CNNs compared to the CAM method, requiring only that the final activation function used for network prediction be a differentiable function, such as the softmax function. For each feature map in the final convolution layer, a gradient of the score y_c (*logit*) of the class concerning every node in A_k is calculated to obtain an importance score $\alpha_{k,c}$ for the feature map A_k. Grad-CAM then linearly combines these importance scores and forwards them through the ReLU function to obtain a relevance map, as follows:

$$map_c = ReLU(\sum_k^K \alpha_{k,c} A_k). \tag{2}$$

The main difference between CAM and Grad-CAM lies in the way the weights for the feature maps are generated. In CAM, the heat maps are generated by calculating the weighted sum of the activations of the last convolutional layer from the weights of the FC layer. In Grad-CAM, the gradients of any layer are used to generate these weights [34].

When the Grad-CAM method is applied to a particular layer, it is called Layer Grad-CAM [45]. Therefore, according to [5], the output gradients are calculated concern to the chosen layer. The resulting gradients are averaged for each output channel and then multiplied by the layer activations. The results are finally summed across all channels. Concerning the weights w referring to features in the class, the GAP is performed on the A feature maps:

$$S^c = \sum_k w_k^c \frac{1}{Z} \sum_i \sum_j A_{i,j}^k. \tag{3}$$

Saliency or Heat Maps are a visualization method commonly used to explain the prediction process of a network. A transparent colored

heat map is often superimposed on the original input image to generate the visualization. Furthermore, these maps identify characteristics of the input that are more relevant (salient). In other words, it highlights regions that most stimulate the network to influence the output of the model [34].

According to [48], class saliency is extracted given an image I_0 (with m rows and n columns) and a class c. The Saliency Map of the class $M \in R^{mxn}$ is calculated from the derivative w by backpropagation. After that, the Saliency Map is obtained by reorganizing the elements of the w vector. If the image is in grayscale, the number of elements in w is equal to the number of pixels in I_0, allowing the map to be calculated as in Equation (4), whose value $h(i, j)$ equals the index of the element of w corresponding to the pixel in the image in the i-th row, and j-th column:

$$M_{i,j} = |w_{h(i,j)}|. \tag{4}$$

Concerning a colored image, it is assumed, for example, that the color channel c of pixel (i, j) of image I corresponds to the element of w with index $h(i, j, c)$. To derive a unique class saliency value for each pixel (i, j), it is necessary to reach the maximum magnitude of w in all color channels, as follows:

$$M_{i,j} = max_c |w_{h(i,j,c)}|. \tag{5}$$

A simple and common qualitative way of comparing features extracted by the first layer of a CNN is to look at the filters learned by the model [12]. In other words, it is necessary to analyze the linear weights in the weight matrix represented in the input space. This approach becomes even more important when the inputs are images that can be viewed. These filters can take the form of trace detectors when trained on digit data, or edge detectors when trained on fragments of natural images.

According to [21], the convolutional layers of a CNN consist of filters that are applied to the input of the layer and produce maps that represent the presence of specific local characteristics of the input, which have been learned by the network. As the filters are linear functions of the input, the weights w_f of the filter and an additive bias b_f determine the contribution of the input x to the activation function σ, whose output is represented as follows:

$$output_f = \sigma(w_f x + b_f). \tag{6}$$

Based on the assumption that a unit can be characterized by the filters of the previous layer to which it is most strongly connected, by obtaining a weighted linear combination of the filters of the previous layer, one can observe that a network with expansion restrictions in the activations, trained on natural images, will tend to learn the so-called corner detectors in the second layer [12].

CNN Filters is a simple and efficient technique; however, one shall note that there is a link between the gradient updates to maximize the activation of a unit and finding the linear combination of weights as described by [29]. Moreover, visualization can vary by layer, e.g., the units in the top layer representing more complex features, and corresponding to combinations of features from the lower layers [12]. Multifaceted patterns can be recognized by the filters in the intermediate layers. Finally, each filter in a layer, except the first, defines a non-convex function with multiple local maxima in the image space, demonstrating that visualization is a useful tool for interpreting the black-box of CNNs [56].

3.2. Aquaculture image segmentation

The measurement of animal characteristics is commonly carried out using digital image analysis, the aim of which is to help improving fish quality and its measurements from manually obtained images. This process can be time-consuming and exhausting for the animal, especially when implemented routinely [4]. Based on this scenario, AI models are being used to make this industry more intelligent [57], enabling numerous applications such as species classification and animal recognition [59]. Furthermore, advances in Aquaculture production have contributed significantly to the search for new technologies, the aim of which is to help produce fish in a sustainable and effective way [15].

According to [14], the search for fast and non-invasive alternatives for measuring fish characteristics has also been crucial for the implementation of AI models. The authors even used automated image segmentation to extract biometric measurements, predict body weight, carcass weight and fillet yield of Nile tilapia fish. Concerning the Pacu fish species, [15] presented a Computer Vision System (CVS) that contains a combination of DL strategies for segmenting animal images. A Region-based CNN (R-CNN) Mask was used to select the regions (body, head, fins, and pelvis) and also to extract the features from the original image.

The fish's characteristics can be measured routinely, without causing major damage to the animal [15]. The results obtained show that creating CVS from AI techniques is efficient for estimating morphometric measurements of Pacu, mainly because of the resilience shown when faced with different lighting conditions and image background [15]. However, according to [14], the AI models used can be considered complex, even though they are state-of-the-art.

In this context, it is possible to observe that AI concepts are increasingly present in society, including biological areas, such as Aquaculture. However, XAI systems are still new to humans, although they have proven to be necessary since the search for explanations in AI systems has ceased to be an additional feature and has become essential. Technological advances have brought benefits to users, but they have also raised concerns regarding the decisions that a machine can make. The large amount of data involved in these decisions raises concerns regarding the security of users and their respective data. Therefore, XAI systems emerged to address these concerns regarding the challenges presented.

4. Methodology

As aforementioned, the results of XAI methods were analyzed in the face of pixel disturbance applied to the input images, to find the best explainable method and combine it with other methods to improve them. A number of steps were taken to achieve this objective: implementation of an AI model for segmenting images of Pacu fish; implementation of four XAI methods compatible with the applied model; application of three types of pixel perturbation to the input images; search for the best XAI method based on an analysis of the results obtained; implementing a combination of explainable methods to improve results; and analysis of the results obtained.

4.1. AI model

The segmentation model was created from a CNN, more specifically the Mask R-CNN, using a variant of ResNet with 18 layers as a feature extractor. The aim of the model is to segment regions of interest in the fish such as head dimensions, body dimensions, and fin area for phenotyping purposes and subsequent genetic selection. For each fish image, the model generates as output the classes (head, fins, body, and background), the masks of each region of the fish, a bounding box and the confidence score in the prediction of the class. This model was determined on the basis of experiments carried out previously, which served to determine the best approach for this work.

In total, there were 100 images of Pacu fish provided by the Laboratory of Genetics in Aquaculture and Conservation (LaGeAC) at Unesp Jabotical. All images were taken in portrait mode and the same environment so as not to alter the model's results. Although the images had different dimensions (images were not taken from the same device), it did not hinder the performance of the methodology, since each method followed the dimensions of the image being analyzed at the time. For the training stage, each image was segmented manually using a spe-

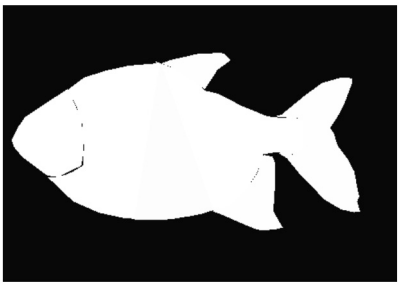


Fig. 1. Example of original mask generated from manual segmentation using the Labelbox tool (https://labelbox.com/).

cific segmentation tool called Labelbox.[1] The model was then trained with the aim of highlighting the parts of the fish and classifying them correctly according to the segmentation performed. The dataset used consists of approximately 1,802 Pacu and 365 Tambaqui animals. It was randomly divided, stratified to preserve the proportions of the different species, into three subsets: training (80%), validation (20%), and 5% of the training set was used to tune hyperparameters during experimentation, avoiding the direct use of the validation set and preventing possible overfitting [3].

For this work, the AI model was subjected to the inference process to make predictions based on what was learned during the training stage. The aim was therefore to segment the images to delimit the fish from the background area, rather than segmenting the animal into parts, as was done previously. A black and white mask was generated based on the manual segmentation, resulting from combining the segmentations of the fish's regions of interest. As shown in Fig. 1, the white areas represent the Pacu area, while the black areas represent the background and other objects in the image.

Each dataset image comes with its own mask, making it possible to compare them with the masks generated by the model during the experiments. In this context, the images were subjected to pixel perturbation techniques and the masks (original and generated during inference) were compared to analyze the difference between them. One can thus observe the influence of pixel perturbations on the results of explainable methods. These analyses are discussed further in this paper.

4.2. XAI methods

The XAI methods were applied to explain the most important regions for the Mask R-CNN prediction model. These methods were chosen based on their compatibility with the AI model used in this work. Moreover, most of the explainable methods provide results through the classification process. However, for the experiments carried out, the aim was to obtain results from the segmentation process, the objective of which was to separate the fish from the rest of the image. Therefore, among the methods previously studied, four explainable visualization methods were implemented (Fig. 2): Grad-CAM, Saliency Map, CNN Filters, and Layer Grad-CAM. They all highlight the regions of the image that are most relevant for the model to identify the regions of the fish. Furthermore, it is possible to check whether the result presented by the CNN matches the important regions highlighted by the explainable methods.

The Grad-CAM is based on gradients and uses colors to highlight the most relevant regions of the image during the segmentation process. The regions presented in warm colors (e.g., red) are the most important for the model during the inference process. This relevance decreases as the colors get colder. The blue regions are the least relevant when it comes to separating the fish from the other parts of the image (Fig. 2a). The second XAI method used was the Saliency Map, the aim of which is to highlight relevant regions based on the brightness of each pixel. For the experiments in this work, the output image has black and red pixels. As a result, the black regions represent the image's background, while the red regions represent the Pacu areas, with the brightest pixels being the most important for the inference process (Fig. 2b).

The CNN Filters method reuses the filters applied by CNN to explain the model's own inference process, making it also an XAI method. The number of images generated depends on the number of convolution layers used by the network, i.e., 17 in this work. Each image shows the most relevant regions for the corresponding layer, so the results are different from each other, allowing for different explanations. For this work, only the result of the first CNN layer was used to explain the model, since it would be impractical to use all outputs (Fig. 2c).

The last method applied was Layer Grad-CAM, which is based on the gradient calculation used in the Grad-CAM method, implemented in the last convolution layer of the CNN. For this work, the output of the method is an image in green and red according to the classes considered for segmentation during the model's inference process. According to Fig. 2d, the color red was used to represent the background of the image, while the color green was used to describe the head, body, and fins of the fish, thus highlighting the animal from the rest of the image. One can observe in green which pixels are considered relevant to determine the regions of the Pacu. It is important to highlight that the output image generated by the method suffers a decrease in size since the output of the last layer of the CNN is used as input for the explainable method, justifying the loss of image quality.

It is worth noting that explanations can be visualized using groups of pixels or pixel-by-pixel of the input image. The Grad-CAM and Layer Grad-CAM methods use groups of pixels to explain the inference process, while the Saliency Map and CNN Filters methods use each point in the image to generate the explanation. As a result, the use of these methods proved to be important due to their variability and distinctiveness. In other words, despite the similarity between them in terms of the type of visualization, and despite the fact that there are only four of them, the XAI methods implemented present totally different explainable results, enabling a greater scope in the experiments carried out.

4.3. Pixel perturbation

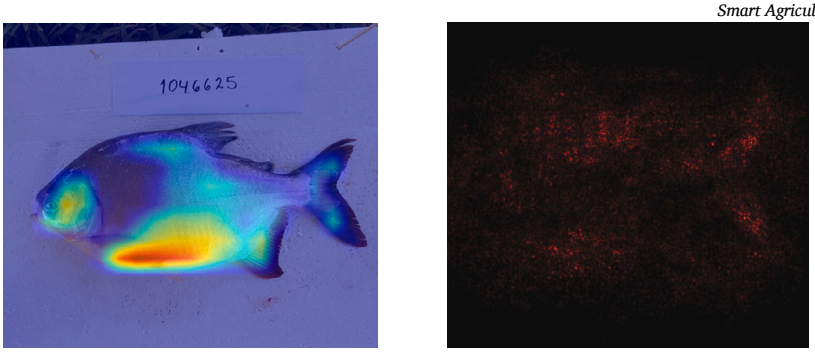
The performance of AI models can be analyzed based on changes to the input images. This is even the premise used in explainable methods based on pixel perturbation, the aim of which is to explain which regions or points of the image are considered relevant to the model. For this work, perturbation techniques were implemented to analyze the influence of these changes on the performance of AI models, and how this affects the explainability.

According to [18], the construction of perturbed images can be performed by changing the pixel to a specific color. Moreover, one can observe the use of pixel perturbation methods to identify the part of the image that has the greatest responsibility in relation to the decision of the classification process, or according to more relavants pixels for a XAI method specific [41].

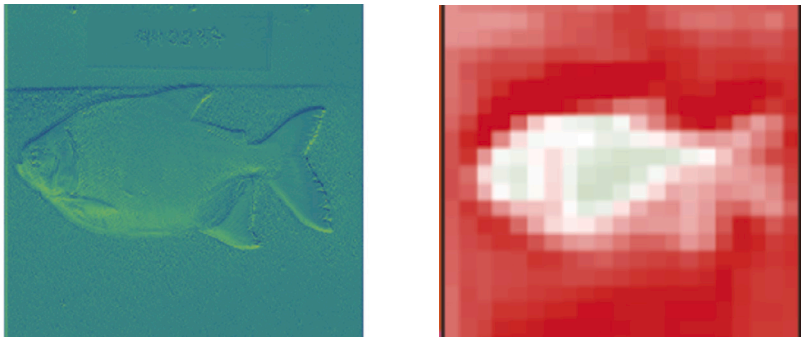
The changes caused by pixel disturbance can occur in different ways, resulting in different techniques. Three disturbance models were considered for this work: white, dark, and random noise. These techniques were implemented using the pixel regions highlighted by the explainable methods. In other words, considering a single input image, for each XAI method implemented, three different pixel perturbation techniques were used, generating three different output images.

Concerning the white noise technique, the changes are characterized by pixels randomly painted in dark and white. Depending on the explainable method in which this disturbance was implemented, white noise can appear in denser regions, or only in random points of the image. This also occurs with the other perturbation techniques mentioned. Fig. 3 illustrates the same image subjected to white noise according to the different XAI methods. Despite being the same type of disturbance, it can be seen that in the Grad-CAM (Fig. 3a) and Layer Grad-CAM (Fig. 3c)

[1] Labelbox: https://labelbox.com/.

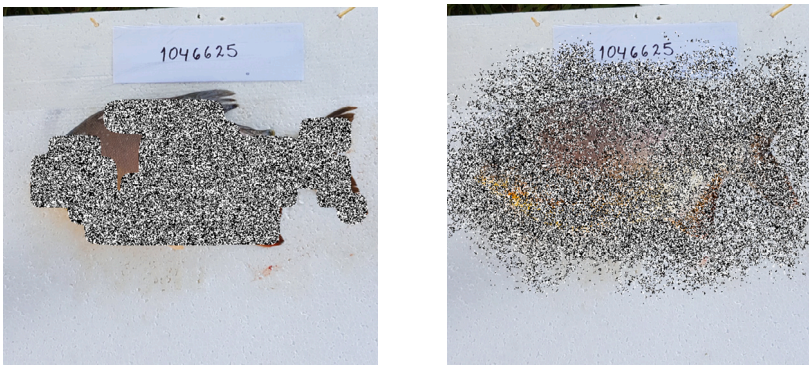


(a) Grad-CAM (b) Saliency Map



(c) CNN Filters (d) Layer Grad-CAM

Fig. 2. Example of XAI methods results.



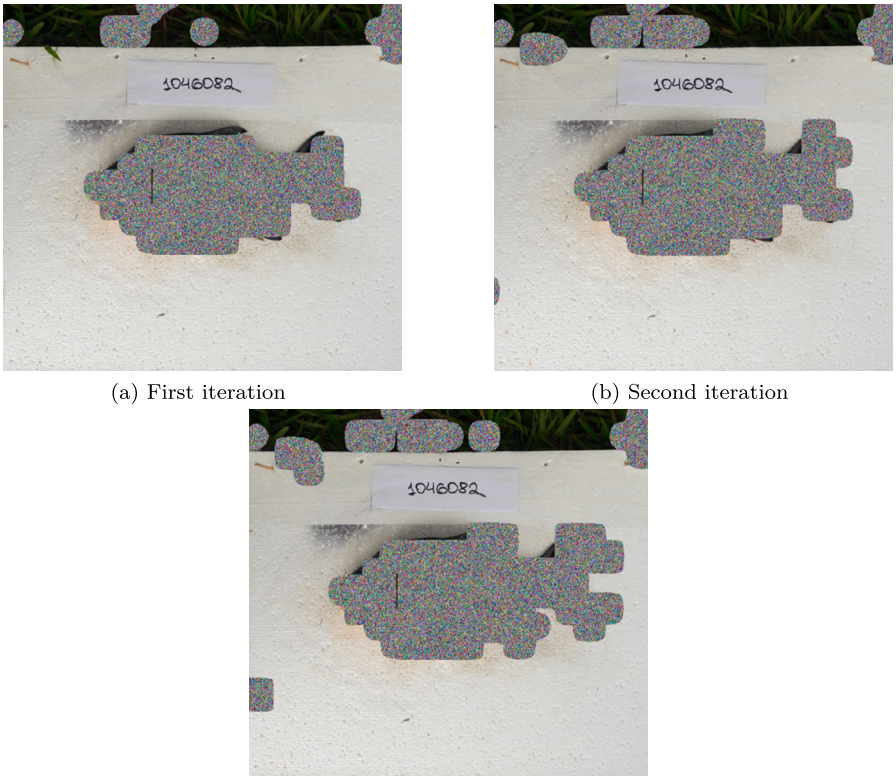
(a) Grad-CAM (b) Saliency Map

(c) Layer Grad-CAM (d) CNN Filters

Fig. 3. Example of white noise pixel disturbance for the four XAI methods.

methods, the changes occur inside the object. Saliency Map (Fig. 3b) and CNN Filters (Fig. 3d) placed points inside and outside the fish's body. Furthermore, the pixel perturbation was only applied to the regions or pixels that the explainable methods characterized as relevant to the AI model.

The second perturbation technique is responsible for changing the color of the pixels of the input image to dark. Based on each of the explainable methods, the dark is used both in regions and at specific points in the image, according on their relevance to the AI model's inference process. The last perturbation technique applied alters the color



(a) First iteration (b) Second iteration

(c) Third iteration

Fig. 4. Example of iterations for the Grad-CAM method on the influence of pixel disturbance randomly colored.

of pixels randomly, with RGB colors between 0 and 255, making them colorful. As such, the most relevant regions or points for the inference process are colored according to the XAI method applied. As a result, unlike previous techniques, even for explainable methods with denser visualization, whose relevance is given by regions, it is still possible to observe pixel-by-pixel explanations due to the wide variety of colors.

When we talk about a good explainable method, this complicates the inference process, since the relevant regions are "covered" and may not be identified by the AI model. When the model indicates a segmentation error during inference, it means that the XAI method used is sufficiently good, because there is no longer any region of the fish that can be segmented. If the model continues with its iterations, it is correct to say that the regions that have been disturbed are not important for the model, indicating that the XAI method used is not satisfactory, as it has highlighted regions that are not really relevant.

4.4. The best XAI method

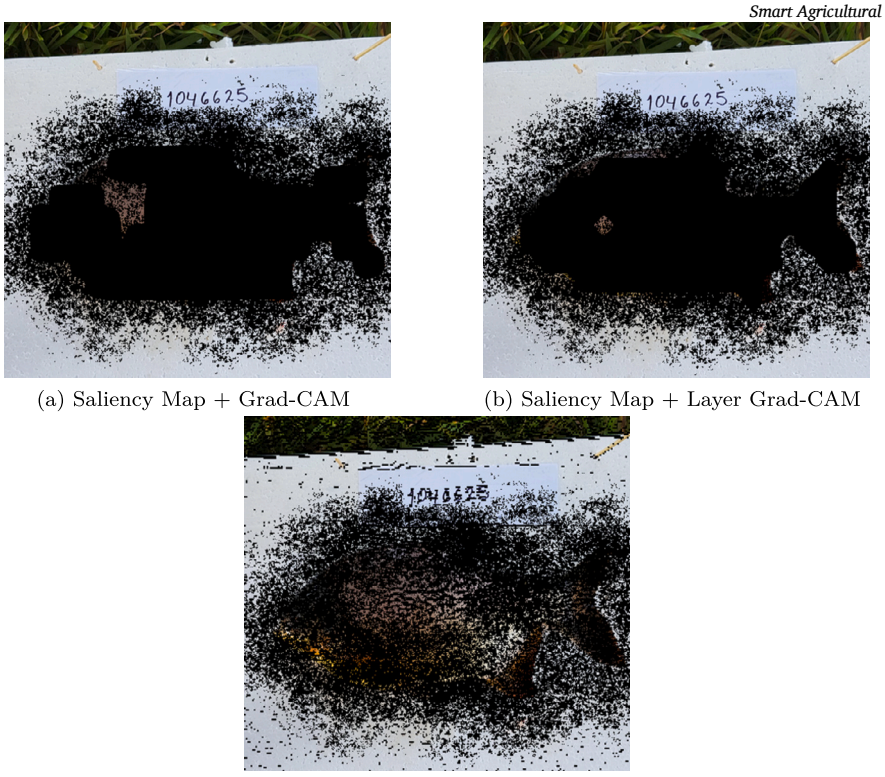
The search for the best method is based on the explanations generated concerning the inference process of the AI model. Even before understanding how the analysis of the four methods applied to this work was carried out, it is necessary to understand the step-by-step process that each image underwent. The following steps were carried out for each XAI method: (1) input image submitted to the AI model; (2) the model inference process was submitted to the XAI method; (3) most relevant pixels are disturbed according to an explainable method, generating an output image; (4) output image is used as input image; and (5) repeating the steps until an error occurs in the inference process.

The execution of steps 1 to 4 characterizes what we call an "iteration". The number of iterations varies depending on the combination of the explainable method and the type of pixel disturbance implemented. With each iteration performed, the more disturbed the pixels of greatest relevance to the model are, which are determined according to the XAI method (Fig. 4). Fig. 4a shows the first iteration of the Grad-CAM method on the influence of the disturbance of randomly colored, and Fig. 4b shows more randomly colored regions, compared to the previous image, representing the second iteration. Fig. 4c, which represents the third iteration, has the most disturbed regions, including areas that are not part of the fish's body, head or fins.

When we talk about a good explainable method, this complicates the inference process, since the relevant regions are "covered" and may not be identified by the AI model. When the model indicates a segmentation error message during inference, it means that the XAI method used is good, because there is no longer any region of the fish that can be segmented. If the model continues with its iterations, it is safe to say that the regions that have been disturbed are not important for the model, indicating that the XAI method used is not suitable, as it has highlighted regions that are not really relevant. The influence of the pixel perturbation applied to the model also determines the performance of the XAI methods. In other words, for each method, it is necessary to analyze how the AI model behaves in the face of each perturbation technique. The first experiment was based on finding the best method, whose perturbation techniques were tested by combining them.

4.5. Combination of XAI methods

Once the best explainable method had been chosen, it was suggested that this method be combined with the other techniques. The aim of this process was to identify whether or not combining the best with the worst could improve the results. In case of improvement, this answers the question "How can XAI methods be improved to present explanations that are consistent with the prediction of the AI model?". To analyze whether the combination was efficient, it was necessary to carry out the same steps presented in the search for the best method. However, the difference lies in the application of the explainable method: instead of



(a) Saliency Map + Grad-CAM (b) Saliency Map + Layer Grad-CAM

(c) Saliency Map + CNN Filters

Fig. 5. Example of an image of a Pacu fish subjected to the combination of the Saliency Map with the other XAI methods, on the effect of the pixels disturbance on dark coloring.

using just one method, the best method was combined with each of the other three remaining ones.

The results were analyzed in a similar way to the search for the best XAI method. In other words, pixel perturbation techniques were implemented to check their influence on the inference process and the explainability in the combination. Therefore, based on the number of iterations obtained for each image, it was possible to observe whether or not there was an improvement in explainability. Moreover, the combination was based on the masks generated by each method during the inference process. As a result, the model's output image contained the most important pixels and regions highlighted according to the perturbation technique. Fig. 5 illustrates an example of a fish image that has been subjected to the combination of the Saliency Map with Grad-CAM (Fig. 5a), Layer Grad-CAM (Fig. 5b) and CNN Filters (Fig. 5c), based on the implementation of the dark coloring perturbation technique.

The process of combining explainable methods one can perform using the logical operators OR, AND, and XOR. Therefore, the AND operator corresponds to the intersection between the images resulting from the methods, while the OR operator corresponds to their union. The XOR operator corresponds to the exclusive disjunction between the images of the combined methods [6].

For this work, the logical operator chosen was OR because of its ability to join the important areas between the images, filling the undefined space with zeros. The union allows us to explore the complementarity of XAI methods, jointly utilizing the information that characterizes each of them. In other words, the location of objects in an image, which is not addressed by the Grad-CAM method, for example, can be satisfied by the Saliency Map method [6].

4.6. Analysis of results

The last stage of the methodology applied aimed at analyzing the results obtained from the experiments. Analysis of the combination of methods was also carried out in an analogous way, making it possible to understand their behavior based on this adjustment. The process of disturbing the input images took place according to the scores obtained for each pixel in the explainability methods. Each method considers a different scale to highlight the most relevant regions, for which the pixels with the values shown in Table 1 were considered. Visual inspection of some images in the dataset was used to identify borderline values for each method. For other XAI methods, a new score analysis should be considered according to the relevant regions of the images. The objective was to establish values that delimited the area belonging to the fish, ignoring pixels from other regions. The broader the threshold established, the more pixels will be considered, which may result in loss of accuracy in the XAI methods.

Table 1
Values of the most important pixels for each XAI method.

XAI Method	Values
Grad-CAM	Greater than 0.01
Saliency Map	Greater than de 0.045
Layer Grad-CAM	Greater than -0.5
CNN Filters	Less than 0

The Sorensen Dice coefficient (SD) is a statistical method that aims to analyze the similarity between two samples [49,8]. The original formula used for this method is presented in Equation (7), the result of which is presented in the range [0, 1]. Therefore, X and Y represent the number of elements in the two samples to be compared, e.g., images. For this work, the similarity between the original fish segmentation mask and the mask generated by the AI model using the XAI method and the pixel perturbation technique applied is compared. This calculation was carried out at each iteration, for each of the explainable methods, on the influence of each perturbation technique. Furthermore, the intersection

between the samples indicates which pixels are shared by the two. The closer the result of the equation is to 1, the more similar the compared masks are.

$$\frac{2|X \cap Y|}{|X| + |Y|}. \tag{7}$$

The Intersection Over Union (IoU), also known as the Jaccard index [24], is a metric used in DL algorithms to check the similarity between two images. In this case, the similarity between the set of pixels in the original mask and the set of pixels in the mask generated by the model was analyzed. This metric is calculated by dividing the overlap between the sets in relation to their union, as presented in Equation (8). The sets are represented by A and B, and the result obtained by the equation is in the range [0, 1]:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{8}$$

Based on these values, it was possible to identify the best explainable method and also the best combination applied. Moreover, one can observe which pixel perturbation technique was most influential concerning the AI model's inference process and the explainability of the XAI method. Finally, it was also possible to observe that in the same method, the perturbation techniques can influence in different ways.

5. Experiments

This section describes the experiments and results obtained based on the methodology presented above. Two experiments were conducted: the first related to the search for the best explainable method and the second related to the improvement of XAI methods by combining the best with others. Both were implemented using PyTorch [39] and designed in Google Collaboratory, using a Python 3 runtime environment with a CPU. Furthermore, 12 GB of RAM and 100 GB of disk space were provided for each of the experiments. These experiments are described in detail and discussed in the following sections.

5.1. Experiment I

The first experiment (Fig. 6) used a model that had already been trained to segment Pacu fishes. The model was created from an R-CNN Mask using a variant of ResNet-18 as a feature extractor. Furthermore, the XAI methods implemented (Grad-CAM, Saliency Map, CNN Filters and Layer Grad-CAM) were considered during the prediction process to analyze which pixels were most important. Finally, the disturbance techniques (white, dark and random noise) served as a method of evaluating the explainable methods.

The process of disturbing the input images took place according to the scores obtained for each pixel in the explainability methods. Each method considers a different scale to highlight the most relevant regions. Visual inspection of some images in the dataset was used to identify borderline values for each method. To determine which was the best explainable method, it was necessary to compare the original mask, generated during manual segmentation, with the mask generated during model prediction after successive pixel perturbations. The results obtained were tabulated according to the perturbation technique and the explainability method used.

All steps were carried out in a maximum of five iterations per image. In some cases, the model itself interrupted the process even before reaching the maximum. In these situations, the model was unable to make the segmentation after the pixels had been disturbed, thus interrupting the process. For each iteration, the image used as input was the image generated as output by the previous iteration, except for the first iteration where the input was the original fish image. Furthermore, with each pass, the most relevant pixels were increasingly disturbed to make the process even more difficult and, consequently, impair the model's ability to predict. The results obtained from these experiments were computed and are presented below.

Table 2
Values obtained for the IoU and SD indices concerning the XAI methods.

XAI Method	Perturbation Technique	IoU	SD
Grad-CAM	Random	0.2821	0.2102
	Dark	0.1654	0.1503
	White noise	0.2864	0.2426
Saliency Map	Random	0.2000	0.1763
	Dark	0.1985	0.1705
	White noise	0.2069	0.1816
Layer Grad-CAM	Random	0.3688	0.2288
	Dark	0.2421	0.2040
	White noise	0.3435	0.2468
CNN Filters	Random	0.0544	0.0671
	Dark	0.0566	0.0721
	White noise	0.0549	0.0685

5.2. Experiment II

The second experiment was carried out based on the results obtained the previous experiment (Fig. 7). The aim was to improve the XAI methods by combining them with the best method previously determined in the first experiment. Based on the three pixel perturbation techniques, the best explainable method was the Saliency Map, which showed to be influenced by the modifications made to the Pacu fish images. In other words, the method showed the truly relevant regions for the prediction process. The methods were combined according to the methodology presented, i.e., with the logical OR operation. This operation joins the images resulting from the XAI methods, whose pixels are highlighted according to their relevance to each methods, resulting in the union of the regions more relevant (Table 1).

The AI model and the dataset used were the same as in the first experiment, the aim of which was to segment the input images based on the body, head, and fins of the Pacu fish. Differently from the previous experiment, the iterations are determined by combining the explainable methods with the Saliency Map, rather than simply implementing each of them. As a result, the three pixel perturbation methods were applied and the resulting mask was used for comparison, just as before.

A maximum of five iterations per image was also considered for this experiment to assess whether the combination of the Saliency Map and the other methods was influenced by the disturbance of pixels classified as significant for the model. Moreover, the same criteria were used to analyze the results. Finally, from the results it was possible to observe the behavior of the methods when combined, even comparing them with the results obtained previously, in which the XAI methods were applied separately.

6. Discussion

From the first experiment, it was possible to determine the influence of the perturbation techniques on the results presented by the XAI methods, in order to identify the best method among the four presented. Table 2 shows the values obtained for the indices IoU and SD. These results show that the masks generated for each XAI method highlight pixels or regions that go beyond the fish's head, body, and fins. The CNN Filters method, for example, obtained the lowest values for both indices, which indicates that the regions highlighted in white, for the most part, are not part of the Pacu fish. However, for the Layer Grad-CAM method, it can be seen that the values are the highest, indicating that most of the areas highlighted in white belong to the fish. Remember that the mask generated during the prediction process ignores characteristics of the methods, such as pixel color or brightness. These characteristics are extremely important for identifying the most relevant regions of the image and therefore prevent the methods from being correctly evaluated using the IoU and SD indices alone.

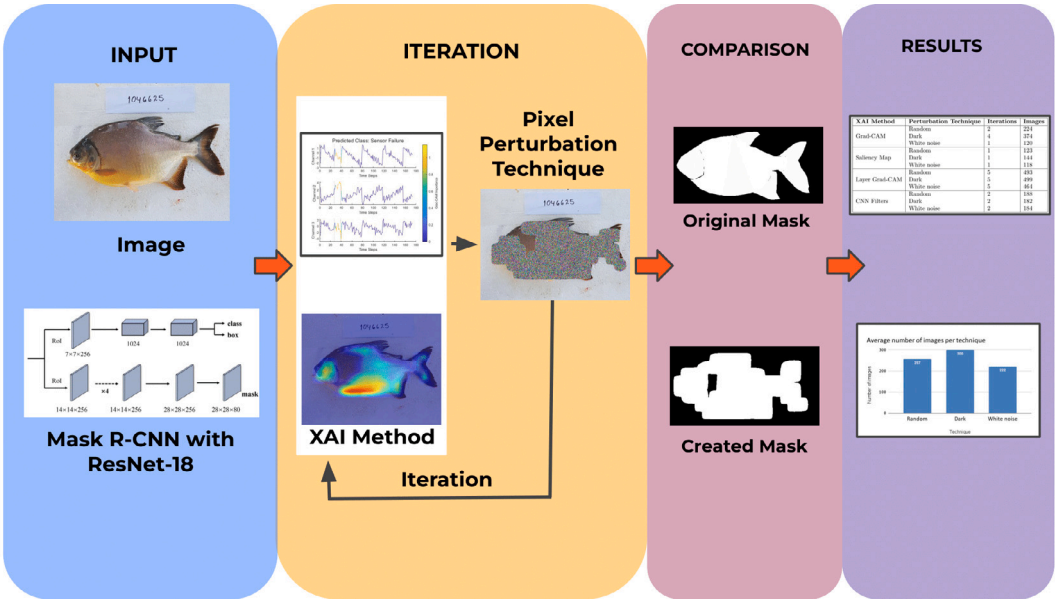


Fig. 6. Diagram of the methodology used for the first experiment.

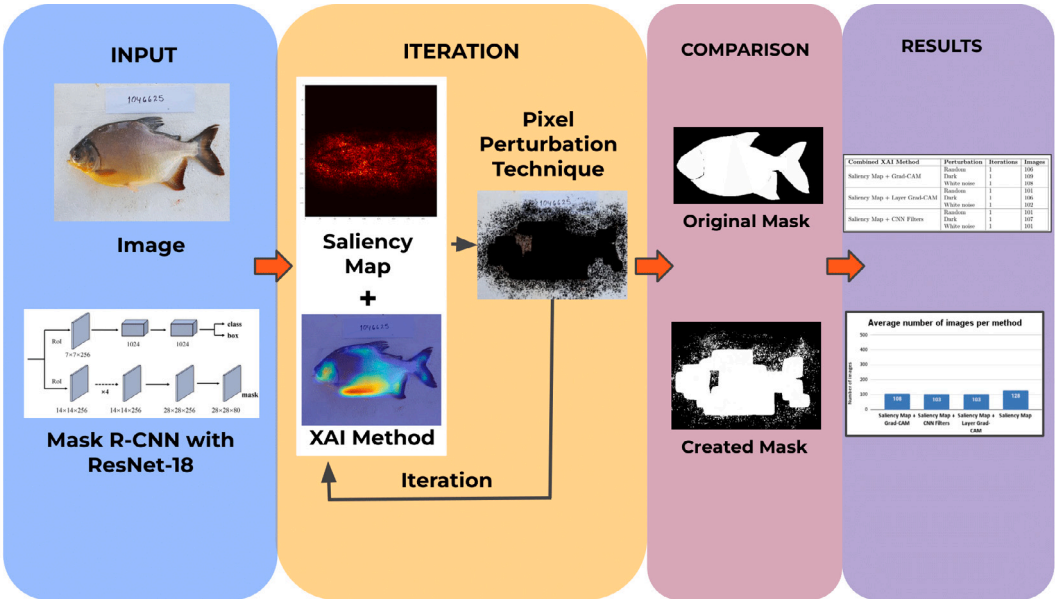


Fig. 7. Diagram of the methodology used for the second experiment.

To determine the results, in addition to these values, the average number of iterations and, consequently, the average number of images generated for each of the methods were observed. These values were used to determine whether or not the method was influenced by the pixel disturbance techniques. Table 3 shows the average number of iterations and the number of images obtained for each XAI method and perturbation techniques. The latter refers to the number of images generated in all iterations for a single image. For example, when the maximum of five iterations for a single input image is reached, five output images are generated. Therefore, based on the 100 images in the dataset used, implementing an XAI method, together with a given pixel perturbation technique, can generate a maximum of 500 output images. It is also worth noting that both values help determine the performance of an explainable method in the face of changes to the pixels in the input image. The methods that achieved the lowest number of iterations and, consequently, lowest number of images, were the ones that showed the best results. It can be seen that not all methods behaved in the same way, and the methods that showed the best results did not necessarily achieve the same result given the number of iterations and images generated.

Given the results presented, the dark color technique had the least effect on the method, as the explanations presented did not highlight the pixels of real importance to the model. The white noise technique had the greatest influence on the Grad-CAM method, resulting in a smaller number of images, with an average of just one iteration per input image. With regard to the methods that reached the maximum number of iterations (five), it can be said that the model continued to be able to perform segmentation even after the most relevant pixels for the current method were disturbed. In other words, the disturbance did not affect the model and the prediction process continued normally. This means that the explainable method erroneously characterized pixels or regions of pixels as important. Therefore, by continuing the prediction even in the face of the disturbance, it means that the change in pixels did not significantly

Table 3
Average number of iterations and number of images per XAI method concerning the pixel perturbation techniques implemented.

XAI Method	Perturbation Technique	Iterations	Images
Grad-CAM	Random	2	224
	Dark	4	374
	White noise	1	120
Saliency Map	Random	1	123
	Dark	1	144
	White noise	1	118
Layer Grad-CAM	Random	5	493
	Dark	5	499
	White noise	5	464
CNN Filters	Random	2	188
	Dark	2	182
	White noise	2	184

affect the AI model. For example, the Layer Grad-CAM method had the highest average number of iterations, making it the worst of the four methods tested. These results can be seen in the Table 3, which shows the average number of iterations for each type of disturbance, together with the number of resulting images, the maximum possible number of which would be 500 (maximum of five iterations for each of the 100 input images).

The CNN Filters method obtained satisfactory results with all pixel perturbation techniques, compared to the Layer Grad-CAM and Grad-CAM method. The average number of iterations remained stable in relation to the changes made to the input images, while the number of images was two per iteration. Despite this, these results show that the model still performed one more iteration, indicating that the pixels highlighted by the method were not all relevant to the prediction process. Therefore, it can be seen that only after the second iteration did the method manage to highlight all regions that were truly significant for the model.

With regard to the methods that interrupted the prediction process even before reaching the maximum of five iterations, this means that the joint implementation of the XAI method with the pixel perturbation technique affected the model and therefore prevented the prediction process altogether. Furthermore, it can be said that the explainable method highlighted the truly most relevant pixels for the AI model. The Saliency Map achieved this result as shown in Table 3, as it obtained the lowest average number of iterations while showing less sensitivity to different types of pixel disturbance, making it the best method of the four. This was also reflected in the number of images obtained.

Of the pixel perturbation techniques implemented, it can be seen that three of the four XAI methods were more influenced by the dark coloring technique than the other white and random coloring techniques (Fig. 8). It is possible that this phenomenon is related to the use of padding, which is generally implemented with zero-value pixels, thus making the model more resilient to disturbances of this nature. For these methods, the number of iterations and images generated was greater for this perturbation technique than for the others. However, the CNN Filters method was the only one to show less sensitivity to this disturbance technique, making it the method with the least variation among the techniques, when considering the number of resulting images.

For the second experiment, the combined evaluation of the explainable methods was performed similarly to experiment I. For the IoU and SD indices, the variations were more subtle, as combining the Saliency Map methods with the other methods highlighted more pixels and revealed more differences between the original mask and the one generated by the model. It is worth noting that not all regions considered relevant by the explainable methods belong to the fish regions. The three combinations obtained similar results for the IoU index. However, the

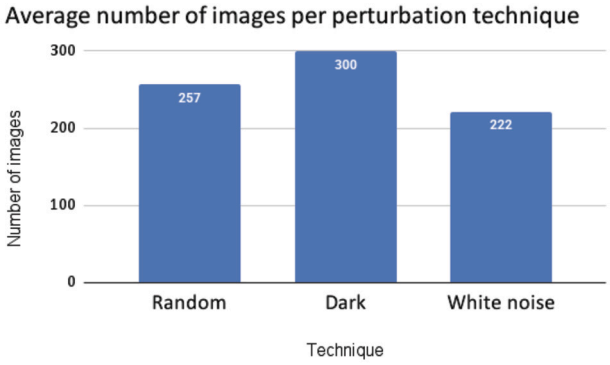


Fig. 8. Bar graph of the average number of images per pixel perturbation technique for the first experiment.

Table 4
Values obtained for the IoU and SD indices concerning the combined XAI methods.

Combined XAI method	Perturbation Technique	IoU	SD
Saliency Map + Grad-CAM	Random	0.1125	0.3072
	Dark	0.1143	0.3048
	White noise	0.1104	0.3015
Saliency Map + Layer Grad-CAM	Random	0.0969	0.3035
	Dark	0.0970	0.2972
	White noise	0.0969	0.3021
Saliency Map + CNN Filters	Random	0.0931	0.1651
	Dark	0.0902	0.1619
	White noise	0.0930	0.1654

Table 5
Average number of iterations and number of images per combined XAI method concerning the pixel perturbation techniques implemented.

Combined XAI Method	Perturbation	Iterations	Images
Saliency Map + Grad-CAM	Random	1	106
	Dark	1	109
	White noise	1	108
Saliency Map + Layer Grad-CAM	Random	1	101
	Dark	1	106
	White noise	1	102
Saliency Map + CNN Filters	Random	1	101
	Dark	1	107
	White noise	1	101

average value achieved for the index was lower for the combination of the Saliency Map and CNN Filters compared to the others.

For the SD index, the combination of the Saliency Map and CNN Filters showed the lowest similarities between the masks. The combination of the Saliency Map and Grad-CAM showed less variation compared to the previous experiment, indicating an improvement in the results. Table 4 shows the average indices obtained for each combination of XAI techniques and pixel perturbation approaches. However, as in the previous experiment, the values obtained for the average number of iterations and the number of images were taken into account.

Table 5 shows these values for each pixel perturbation technique applied to each XAI method in combination with the Saliency Map. These values demonstrate the improvement of the methods compared to the previous experiment, in which it was possible to observe a significant difference between the methods. However, for this experiment, it can be seen that the logical OR operation stabilized the values (amount of iteraction and images).

Table 5 demonstrates the improvement of the Grad-CAM method when combined with the Saliency Map. It can be seen that unlike the pre-

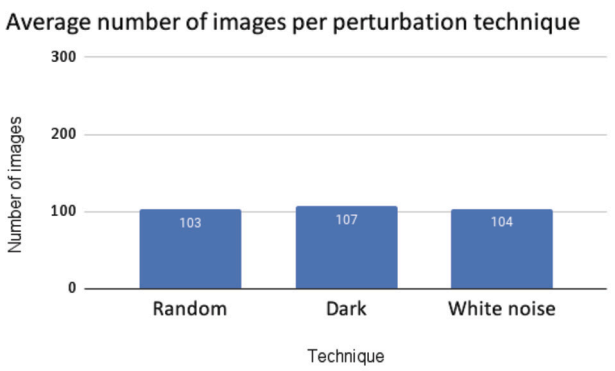


Fig. 9. Bar graph of the average number of images per pixel perturbation technique for the second experiment.

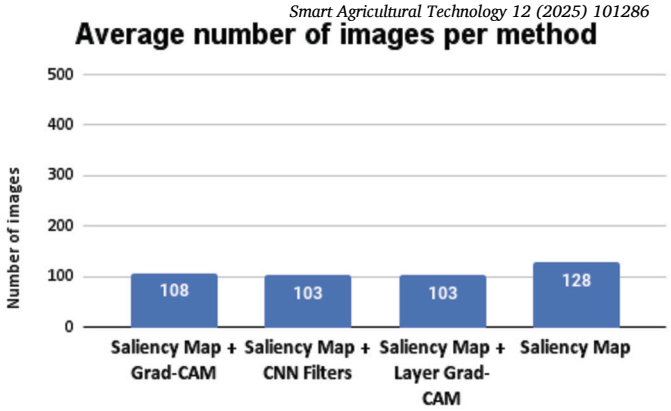


Fig. 10. Bar chart of the average number of images by combination of methods.

vious experiment, Grad-CAM remained stable in the face of the types of perturbations applied. In addition, the black coloring technique, which previously produced results close to 400 images, was considerably reduced to just 109 images. Even with the white noise technique, which had produced a satisfactory result of 120 images, the method managed to produce even better results, obtaining a total of 108 images. In view of this, for all perturbation techniques, the average number of iterations was one, differently what was presented previously. This explains the drop in the number of images generated.

The Layer Grad-CAM, previously considered the worst explainable method for the context of segmenting Pacu fish images, was the one that obtained the most significant change in its results. Therefore, in both perturbation techniques, the method had presented values above 400 for the number of resulting images, while for experiment II, the method behaved much better and with results way below; the lowest value being 101 images for the random noise technique. It's worth noting that the minimum quantity possible in this model is 100 images (with one resulting image for each of the 100 fish input images). For all types of disturbance, the method performed, on average, only one iteration (Table 5).

Following the same performance obtained with the other methods, the results of the combination with the Saliency Map for the CNN Filters method were also satisfactory. One can observe that the values were better than those achieved in the first experiment. While the lowest number of images obtained by the method was 182 for the black color disturbance in experiment I, the highest number obtained in experiment II was 107 for the same type of disturbance. This shows that even the worst value obtained exceeds the best result achieved previously.

Despite the good results, the dark color perturbation technique remained the worst (Fig. 9). On the other hand, compared to the first experiment, this technique achieved a significant improvement. While the average number of images for the first experiment was 300 concerning the dark noise, the average obtained in the second experiment was 107. Furthermore, a significant difference was previously observed between this technique and the others, whereas for the present experiment, this difference was subtle (maximum difference of four images from the average of each technique).

In all cases, one can confirm that the combination served as a determining factor for the improvement of the three other explainable methods. Moreover, it was observed that even the worst results of this experiment were superior to the results obtained by the Salience Map method itself, considered the best. Fig. 10 shows a graph of the average number of images obtained by the combinations, compared to the average obtained by the Saliency Map method alone. It can be seen that the value of 128 images, considered a good result in the first experiment, was exceeded by the other three combinations. These results show that the objectives of this work were achieved, since the Saliency Map method, which was evaluated as the best, was decisive for the improvement of the other XAI methods implemented in this experiment.

7. Conclusions

XAI has been considered as a possible solution in the search for transparency in black-box models. However, just as it is necessary to verify the origin of AI model predictions, it is also necessary to analyze the explanations generated by explainable methods. The need therefore arose to evaluate these methods to check the quality of the explanations in relation to the results presented during the prediction process. This makes it possible to determine the best methodology to consider in a specific scenario, such as the health sector, where medical images need to be explained accurately.

One commonly used technique for evaluation is pixel perturbation, especially for models using image inputs. This involves altering key pixels in the image to confirm if the prediction remains accurate and if the explanations are influenced by these changes. Successful results indicate that the explainable method highlights the crucial regions considered by the model during predictions. Contrary results, suggest that the XAI method identifies different pixels than those deemed significant by the AI models.

The main limitations encountered during the development of this work are:

- Lack of standardization of output images: each explainable model generated images in different dimensions, making comparison between models difficult. Therefore, it was necessary to standardize the images, which contributed to a loss of quality in the generated images, especially in the Layer Grad-CAM method;
- Different images from the CNN Filters method: the method generates an output image for each neural network filter. Therefore, these images differ from each other, both in dimension and in the methodology used to highlight the image pixels;
- Lack of compatible explainable methods: difficulty in finding XAI methods that were compatible with the image segmentation model used, limiting the experiments to the four explainable methods mentioned; and
- Dataset standardization: images of fish of the same species, in the same environment, and the same position. Based on this dataset, results may not be compatible with the variability existing in aquaculture, hindering the accuracy of methods about the different contexts presented by the area.

For the present work, two experiments were conducted to analyze the segmentation of Pacu fish images, whose Grad-CAM, Saliency Map, CNN Filters and Layer Grad-CAM techniques were evaluated. In the first experiment, the results indicated that Saliency Map had the best performance, revealing the most relevant pixels according to Mask R-CNN, even in disturbances such as black color, randomness and white noise. These results addressed two of the three research questions presented in this study.

The black color perturbation produced more iterations and, consequently, more images in three out of the four XAI methods discussed. Thus, it is inferred that this method has the least impact on the segmentation model among those tested. Analysis of the average number of iterations and images generated suggests that the Saliency Map method was the least affected by different perturbation techniques, while CNN Filters showed the least sensitivity to perturbation types, exhibiting minimal variation in image averages. In contrast, Grad-CAM was the most responsive to perturbation techniques among the four.

The exploration of XAI methods unveiled a plethora of explainable methodologies across various domains. This diversity spurred the quest for techniques enhancing XAI methods to elevate the quality of explanations. Consequently, the second experiment in this study aimed to enhance explanations by combining the best XAI method identified in the first experiment with other implemented methods through the intersection of resulting masks.

From the results of the second experiment, it was observed that the methodology used not only improved Grad-CAM, CNN Filters, and Layer Grad-CAM methods but also enhanced the Saliency Map method itself, surpassing the values of the best method. This result addresses the final research question and supports the aims initially proposed. Thereby, it is concluded that enhancing the explanation quality of an explainable method applied to an image segmentation model is feasible. Furthermore, the combination of methods can potentially diminish the explainability of the superior method when paired with a lower-quality one. However, in this experiment, Layer Grad-CAM considered the weakest method in the first experiment, did not compromise the effectiveness of the best method. Moreover, the Saliency Map also exhibited a notable improvement in results, indicating the successful synergy of methods.

In comparison with other studies found in the literature, this work presents superior results regarding the evaluation of explainable methods. The success of the combination of techniques in the face of the presented pixel perturbations was also proven. However, it is still necessary to compare the results achieved through other existing image segmentation models, as well as to evaluate other XAI methods from different evaluation techniques. Finally, the scientific contribution of this research is notable, which aims to question the explainability of existing XAI methods.

Given the widespread popularity of intelligent models, the results show how important it is to work with reliable XAI methods to better understand the steps involved in AI experiments. For the Aquaculture sector, this prevents erroneous and misleading results related to aquatic organisms and their characteristics. In addition, it prevents waste and provides greater confidence to professionals in the field who see AI as an opportunity to improve and optimize farming activities.

With these results, the explainability of XAI methods can be investigated in other contexts, such as image classification. Therefore, for future research, one recommends replicating these experiments with various AI models (e.g., U-Net, DeepLabV3+, or Swin-UNet), with different dataset partitioning strategy (e.g., K-Fold Cross Validation), and explainability methods. To further validate the quality of explanations, one expects to use alternative mechanisms, such as automated thresholding methods (e.g., Otsu's algorithm or percentile-based algorithm) and statistical methods (e.g., p-values or confidence intervals), instead of visual inspection to determine relevant regions. Also one expects to explore different perturbation techniques, such as alternative distortions (e.g., Gaussian blur, sliding-window occlusion, or adversarial patches), and evaluation methodologies, such as objective metrics (e.g., deletion AUC, insertion AUC, or relevance mass accuracy).

Testing logical combinations (e.g., AND or XOR) between the Saliency Map and other methods may yield new insights into improvements in explainability. Furthermore, applying the image segmentation model to identify fish body parts, such as the body, head, and fins, represents a potential future experiment. Applying assessment indicators and quantitative analysis, and, finally, extending these research hypotheses to other contexts, such as classification and diverse datasets (with images in different environments and dimensions), can broaden the findings of this study, including in other branches of Aquaculture. Finally, one expects to consult with Aquaculture experts to assess the feasibility of the study and its contributions to the field.

CRediT authorship contribution statement

Juliana da C. Feitosa: Writing – original draft, Software, Resources, Methodology, Investigation, Conceptualization. **Fabrício M. Batista:** Software, Investigation. **Juliana C.F. Catharino:** Writing – review & editing. **Milena V. Freitas:** Data curation. **Diogo T. Hashimoto:** Data curation. **João Paulo Papa:** Writing – review & editing, Supervision, Methodology. **José Remo F. Brega:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition.

Ethics approval

All animal handling procedures were performed in accordance with the guidelines established by the Ethics Committee on Animal Use of the Faculty of Agrarian and Veterinary Sciences of UNESP (CEUA, 19005/2017).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Juliana da Costa Feitosa reports financial support was provided by Coordination for the Improvement of Higher Education Personnel (CAPES), Process No. 88887.512829/2020-00. This project was partially supported by Huawei do Brasil Telecomunicações Ltda (Fundunesp Process 3123/2020). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was financed with support from the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES). This project was partially supported by Huawei do Brasil Telecomunicações Ltda.

Data availability

Supplementary data to this article can be found online at https://github.com/jufeitosa10/XAI_Segmentacao_Pacu.git

References

[1] A. Abanda, U. Mori, J.A. Lozano, Ad-hoc explanation for time series classification, Knowl.-Based Syst. 252 (2022) 109366, https://doi.org/10.1016/j.knosys.2022.109366.

[2] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (xai): what we know and what is left to attain trustworthy artificial intelligence, Inf. Fusion 99 (2023) 101805.

[3] F.M. Batista, J.R.F. Brega, Image segmentation applied to multi-species phenotyping in fish farming, in: Computational Science and Its Applications – ICCSA 2024, Springer Nature Switzerland, Cham, 2024, pp. 96–111.

[4] A.J.d.S. Cardoso, C.A.L.d. Oliveira, E.C. Campos, R.P. Ribeiro, G.J.d.F. Assis, F.F.e. Silva, Estimation of genetic parameters for body areas in nile tilapia measured by digital image analysis, J. Animal Breed. Genet. 138 (2021) 731–738, https://doi.org/10.1111/jbg.12551.

[5] S. Chatterjee, A. Das, C. Mandal, B. Mukhopadhyay, M. Vipinraj, A. Shukla, R. Nagaraja Rao, C. Sarasaen, O. Speck, A. Nürnberger, Torchesegeta: framework for interpretability and explainability of image-based deep learning models, Appl. Sci. 12 (2022) 1834, https://doi.org/10.3390/app12041834.

[6] S. Coulibaly, B. Kamsu-Foguem, D. Kamissoko, D. Traore, Explainable deep convolutional neural networks for insect pest recognition, J. Clean. Prod. 371 (2022) 133638, https://doi.org/10.1016/j.jclepro.2022.133638.

[7] N. Díaz-Rodríguez, A. Lamas, J. Sanchez, G. Franchi, I. Donadello, S. Tabik, D. Filliat, P. Cruz, R. Montes, F. Herrera, Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: the monumai cultural heritage use case, Inf. Fusion 79 (2022) 58–83, https://doi.org/10.1016/j.inffus.2021.09.022.

[8] L.R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (1945) 297–302.

[9] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, preprint, arXiv:1702.08608, https://doi.org/10.48550/arXiv.1702.08608, 2017.

[10] D.M. Eler, M.P. Batista, R. Garcia, D. Pereira, W. Marcilio, Visual approach to support analysis of optimum-path forest classifier, in: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), IEEE Computer Society, Los Alamitos, CA, USA, 2019, pp. 777–782, https://doi.ieeecomputersociety.org/10.1109/BRACIS.2019.00139.

[11] S. Elguendouze, A. Hafiane, M.C. de Souto, A. Halftermeyer, Explainability in image captioning based on the latent space, Neurocomputing 546 (2023) 126319, https://doi.org/10.1016/j.neucom.2023.126319.

[12] D. Erhan, Y. Bengio, A. Courville, P. Vincent, Visualizing Higher-Layer Features of a Deep Network, University of Montreal 1341, 2009, p. 1.

[13] J.M. Fellous, G. Sapiro, A. Rossi, H. Mayberg, M. Ferrante, Explainable artificial intelligence for neuroscience: behavioral neurostimulation, Front. Neurosci. 13 (2019), https://doi.org/10.3389/fnins.2019.01346. cited by 0.

[14] A.F. Fernandes, E.M. Turra, E.R. de Alvarenga, T.L. Passafaro, F.B. Lopes, G.F. Alves, V. Singh, G.J. Rosa, Deep learning image segmentation for extraction of fish body measurements and prediction of body weight and carcass traits in nile tilapia, Comput. Electron. Agric. 170 (2020) 105274, https://doi.org/10.1016/j.compag.2020.105274.

[15] M.V. Freitas, C.G. Lemos, R.B. Ariede, J.F. Agudelo, R.R. Neto, C.H. Borges, V.A. Mastrochirico-Filho, F. Porto-Foresti, R.L. Iope, F.M. Batista, et al., High-throughput phenotyping by deep learning to include body shape in the breeding program of pacu (piaractus mesopotamicus), Aquaculture 562 (2023) 738847, https://doi.org/10.1016/j.aquaculture.2022.738847.

[16] M. Ghassemi, L. Oakden-Rayner, A.L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, Lancet Digit. Health 3 (2021) e745–e750, https://doi.org/10.1016/S2589-7500(21)00208-9.

[17] R. Gipiškis, D. Chiaro, M. Preziosi, E. Prezioso, F. Piccialli, The impact of adversarial attacks on interpretable semantic segmentation in cyber-physical systems, IEEE Syst. J. (2023), https://doi.org/10.1109/JSYST.2023.3281079.

[18] O. Gorokhovatskyi, O. Peredrii, Multiclass image classification explanation with the complement perturbation images, in: Data Stream Mining & Processing: Third International Conference, DSMP 2020, Lviv, Ukraine, August 21–25, 2020, Proceedings 3, Springer, 2020, pp. 275–287.

[19] N. Gumpfer, J. Prim, T. Keller, B. Seeger, M. Guckert, J. Hannig, Signed explanations: unveiling relevant features by reducing bias, Inf. Fusion (2023) 101883, https://doi.org/10.1016/j.inffus.2023.101883.

[20] D. Gunning, D. Aha, Darpa's explainable artificial intelligence program, AI Mag. 40 (2019) 44–58, https://doi.org/10.1609/aimag.v40i2.2850, cited by 6.

[21] J. Hochuli, A. Helbling, T. Skaist, M. Ragoza, D.R. Koes, Visualizing convolutional neural network protein-ligand scoring, J. Mol. Graph. Model. 84 (2018) 96–108, https://doi.org/10.1016/j.jmgm.2018.06.005.

[22] R. Ibrahim, M.O. Shafiq, Explainable convolutional neural networks: a taxonomy, review, and future directions, ACM Comput. Surv. 55 (2023) 1–37, https://doi.org/10.1145/3563691.

[23] M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: a survey, Pattern Recognit. Lett. 150 (2021) 228–234, https://doi.org/10.1016/j.patrec.2021.06.030.

[24] P. Jaccard, The distribution of the flora in the alpine zone. 1, New Phytol. 11 (1912) 37–50.

[25] W. Jin, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable ai in medical image analysis, Med. Image Anal. 84 (2023) 102684, https://doi.org/10.1016/j.media.2022.102684.

[26] M.A. Kadir, A. Mohamed Selim, M. Barz, D. Sonntag, A user interface for explaining machine learning model explanations, in: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, Association for Computing Machinery, New York, NY, USA, 2023, pp. 59–63.

[27] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, J. Wortman Vaughan, Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–14.

[28] J.P. Kucklick, O. Müller, Tackling the accuracy-interpretability trade-off: interpretable deep learning models for satellite image-based real estate appraisal, ACM Trans. Manag. Inf. Syst. 14 (2023) 1–24, https://doi.org/10.1145/3567430.

[29] H. Lee, R. Grosse, R. Ranganath, A.Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: Proceedings of the 26th Annual International Conference on Machine Learning, Association for Computing Machinery, New York, NY, USA, 2009, pp. 609–616.

[30] Y. Liang, M. Li, C. Jiang, Generating self-attention activation maps for visual interpretations of convolutional neural networks, Neurocomputing 490 (2022) 206–216, https://doi.org/10.1016/j.neucom.2021.11.084.

[31] Y.S. Lin, W.C. Lee, Z.B. Celik, What do you see? Evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1027–1035.

[32] P. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, T. Villmann, The coming of age of interpretable and explainable machine learning models, Neurocomputing 535 (2023) 25–39.

[33] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2021) 3523–3542, https://doi.org/10.1109/TPAMI.2021.3059968.

[34] E. Mohamed, K. Sirlantzis, G. Howells, A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation, Displays 73 (2022) 102239, https://doi.org/10.1016/j.displa.2022.102239.

[35] 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, https://doi.org/10.1145/3387166.

[36] S.M. Muddamsetty, M.N. Jahromi, A.E. Ciontos, L.M. Fenoy, T.B. Moeslund, Visual explanation of black-box model: similarity difference and uniqueness (sidu) method, Pattern Recognit. 127 (2022) 108604, https://doi.org/10.1016/j.patcog.2022.108604.

[37] H. Mzoughi, I. Njeh, M. BenSlima, N. Farhat, C. Mhiri, Vision transformers (vit) and deep convolutional neural network (d-cnn)-based models for mri brain primary tumors images multi-classification supported by explainable artificial intelligence (xai), Vis. Comput. (2024) 1–20, https://doi.org/10.1007/s00371-024-03524-x.

[38] H. Nematzadeh, J. García-Nieto, I. Navas-Delgado, J.F. Aldana-Montes, Ensemble-based genetic algorithm explainer with automized image segmentation: a case study on melanoma detection dataset, Comput. Biol. Med. 155 (2023) 106613, https://doi.org/10.1016/j.compbiomed.2023.106613.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: an imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2019, https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[40] G. Ras, N. Xie, M. Van Gerven, D. Doran, Explainable deep learning: a field guide for the uninitiated, J. Artif. Intell. Res. 73 (2022) 329–396, https://doi.org/10.1613/jair.1.13200.

[41] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[42] Z. Salahuddin, H.C. Woodruff, A. Chatterjee, P. Lambin, Transparency of deep neural networks for medical image analysis: a review of interpretability methods, Comput. Biol. Med. 140 (2022) 105111, https://doi.org/10.1016/j.compbiomed.2021.105111.

[43] H. Saleem, A.R. Shahid, B. Raza, Visual interpretability in 3d brain tumor segmentation network, Comput. Biol. Med. 133 (2021) 104410, https://doi.org/10.1016/j.compbiomed.2021.104410.

[44] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D.A. Keim, Towards a rigorous evaluation of xai methods on time series, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 4197–4201.

[45] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2020) 336–359, https://doi.org/10.1007/s11263-019-01228-7.

[46] R. Shi, T. Li, Y. Yamaguchi, Understanding contributing neurons via attribution visualization, Neurocomputing 126492doi (2023), https://doi.org/10.1016/j.neucom.2023.126492.

[47] S. Shojaei, M.S. Abadeh, Z. Momeni, An evolutionary explainable deep learning approach for Alzheimer's mri classification, Expert Syst. Appl. 220 (2023) 119709, https://doi.org/10.1016/j.eswa.2023.119709.

[48] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps, preprint, arXiv:1312.6034, https://doi.org/10.48550/arXiv.1312.6034, 2013.

[49] T. Sorensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, Biol. Skr. 5 (1948) 1–34.

[50] I.P. de Sousa, M.M. Vellasco, E.C. da Silva, Evolved explainable classifications for lymph node metastases, Neural Netw. 148 (2022) 1–12, https://doi.org/10.1016/j.neunet.2021.12.014.

[51] K. Sriputhorn, R. Pitakaso, S. Matitopanum, P. Luesak, S. Khonjun, R. Kraiklang, C. Chueadee, S. Gonwirat, Revolutionizing climbing perch disease management: ai-driven solutions for sustainable aquaculture, Smart Agric. Technol. 10 (2025) 100746, https://www.sciencedirect.com/science/article/pii/S2772375524003502, https://doi.org/10.1016/j.atech.2024.100746.

[52] M. Veerappa, M. Anneken, N. Burkart, M.F. Huber, Validation of xai explanations for multivariate time series classification in the maritime domain, J. Comput. Sci. 58 (2022) 101539, https://doi.org/10.1016/j.jocs.2021.101539.

[53] C.P. Vieira, L.A. Digiampietri, Machine learning post-hoc interpretability: a systematic mapping study, in: Proceedings of the XVIII Brazilian Symposium on Information Systems, Association for Computing Machinery, New York, NY, USA, 2022.

[54] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Inf. Fusion 76 (2021) 89–106, https://doi.org/10.1016/j.inffus.2021.05.009.

[55] R.O. Weber, A.J. Johs, J. Li, K. Huang, Investigating textual case-based XAI, in: M.T. Cox, P. Funk, S. Begum (Eds.), Case-Based Reasoning Research and Development, vol. 11156, Springer International Publishing, 2018, pp. 431–447, http://link.springer.com/10.1007/978-3-030-01081-2_29.

[56] G. Xie, K. Yang, J. Lai, Filter-in-filter: low cost cnn improvement by sub-filter parameter sharing, Pattern Recognit. 91 (2019) 391–403, https://doi.org/10.1016/j.patcog.2019.01.044.

[57] X. Yang, S. Zhang, J. Liu, Q. Gao, S. Dong, C. Zhou, Deep learning for smart fish farming: applications, opportunities and challenges, Rev. Aquacult. 13 (2021) 66–90, https://doi.org/10.1111/raq.12464.

[58] M.R. Zafar, N. Khan, Deterministic local interpretable model-agnostic explanations for stable explainability, Mach. Learn. Knowl. Extr. 3 (2021) 525–541, https://doi.org/10.3390/make3030027.

[59] S. Zhao, S. Zhang, J. Liu, H. Wang, J. Zhu, D. Li, R. Zhao, Application of machine learning in intelligent fish aquaculture: a review, Aquaculture 540 (2021) 736724, https://doi.org/10.1016/j.aquaculture.2021.736724.

[60] F. Zhipeng, H. Gani, Interpretable models for the potentially harmful content in video games based on game rating predictions, Appl. Artif. Intell. 36 (2022) 2008148, https://doi.org/10.1080/08839514.2021.2008148.

[61] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.