Influence of Pixel Perturbation on eXplainable Artificial Intelligence Methods

Juliana da Costa Feitosa[a], Mateus Roder[b], João Paulo Papa[c] and José Remo Ferreira Brega[d]
Departmant of Computing, School of Science, Sao Paulo State University (UNESP), Brazil

Keywords: eXplainable Artificial Intelligence, Pixel Pertubation, Artificial Intelligence.

Abstract: The current scenario around Artificial Intelligence (AI) has demanded more and more transparent explanations about the existing models. The use of eXplicable Artificial Intelligence (XAI) has been considered as a solution in the search for explainability. As such, XAI methods can be used to verify the influence of adverse scenarios, such as pixel disturbance on AI models for segmentation. This paper presents the experiments performed with fish images of the Pacu species to determine the influence of pixel perturbation through the following explainable methods: Grad-CAM, Saliency Map, Layer Grad-CAM and CNN Filters. The perturbed pixels were considered the most important for the model during the segmentation process of the input image regions. From the existing pixel perturbation techniques, the images were subjected to three main techniques: white noise, color black noise and random noise. From the results obtained, it was observed that the Grad-CAM method had different behaviors for each perturbation technique tested, while the CNN Filters method showed more stability in the variation of the image averaging. The Saliency Map was the least sensitive to the three types of perturbation, as it required fewer iterations. Furthermore, of the perturbation techniques tested, Black noise showed the least ability to impact segmentation. Thus, it is concluded that the perturbation methods influence the outcome of the explainable models tested and interfere with these models in different ways. It is suggested that the experiments presented here be replicated on other AI models, on other explainability methods, and with other existing perturbation techniques to gather more evidence about this influence and from that, quantify which combination of XAI method and pixel perturbation is best for a given problem.

1 INTRODUCTION

Artificial Intelligence (AI) is commonly used today to describe the newest experiences of interaction between computer systems and their users (Kaufman, 2019). Through this interaction that AI systems and concepts can be found in countless areas of knowledge, such as law, medicine, engineering and mathematics (Russell and Norvig, 2004).

The aim of AI is to make machines partially simulate the workings of the human mind (Kistan et al., 2018). Therefore, there is still no AI system that completely simulates our brains and solves every type of problem solved by a human being. However, there is still no knowledge of all the problems that are capable of being solved by AI systems or of their total capacity (Teixeira, 2019).

From this context, Deep Learning (DL) was created with the aim of configuring the parameters of the input data so that the machine learns on its own, through pattern recognition, in several layers of artificial neurons (Goodfellow et al., 2016). In this way, DL is currently used for image recognition, speech, object detection and content description (Deng and Yu, 2014).

According to (Fellous et al., 2019), there are two possible classifications for existing ML models. The first is black-box models, whose decisions made by the machine are difficult for a human being to explain (e.g. DL). Black boxes are considered to be more complex and perform better. In contrast, there are white-box models whose decisions can be explained and are therefore more transparent (e.g. decision trees) (Camacho et al., 2018).

The advance of AI has also led to concerns about the transparency of decisions. As an example of this, it has recently been possible to observe the emergence of the legal requirement prescribed by art. 22 of

[a] https://orcid.org/0009-0005-6935-1022
[b] https://orcid.org/0000-0002-3112-5290
[c] https://orcid.org/0000-0002-6494-7514
[d] https://orcid.org/0000-0002-2275-4722

Feitosa, J., Roder, M., Papa, J. and Brega, J.
Influence of Pixel Perturbation on eXplainable Artificial Intelligence Methods.
DOI: 10.5220/0012424800003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 624-631
ISBN: 978-989-758-679-8; ISSN: 2184-4321

the General Data Protection Regulation (GDPR) described by the European Union jurisdiction (Wolf and Ringland, 2020) (Arnout et al., 2019), which ensures the transparency of an AI system's decisions. It was also possible to observe the emergence of the General Personal Data Protection Act (LGPD), which similarly demonstrates concern about advances in AI and data protection at the national level (Pinheiro, 2020). As a result, the search for secure and transparent alternatives has become a priority for researchers in the field.

Based on this context, eXplainable Artificial Intelligence (XAI) emerged, defined as a set of techniques that combine AI methods, ML and DL, with effective transparent approaches to generate explainable outputs (Fellous et al., 2019). In other words, the term XAI refers to techniques that make AI models understandable to humans (Wolf and Ringland, 2020). Despite gaining due attention recently, according to (Xu et al., 2019), concepts related to XAI date back 40 years, where rules were used to explain the functioning of expert systems. However, in 2017, the Defense Advanced Research Projects Agency (DARPA) of the United States of America (USA) created a program aimed at XAI, whose goal was to create AI systems capable of explaining their logic to humans, in order to characterize their strengths and weaknesses, and transmit future behavioral information (Gunning and Aha, 2019).

According to (Wolf and Ringland, 2020), explanations can be classified as global, whose purpose is to describe the representations of the model used, and local, whose purpose is to explain the input data. Moreover, explainability is based on the human being's need to understand the system (Wolf, 2019). Therefore, based on the user, who may or may not be an expert, the explanations provided by AI can be spoken or created to be visualized, as required (Weber et al., 2018). According to DARPA's definitions, explanations can be classified into four modes: analytical statements, visualizations, cases and rejections of alternative choices (Gunning, 2017). In both cases, explainability is achieved based on the prediction process of the AI model analyzed.

The classification of XAI methods can also be defined according to the methodology used to generate the explanations. According to (Ivanovs et al., 2021a), pixel perturbation techniques, e.g., make it possible to analyze the model's input in relation to its output. With this in mind, there are XAI methods based on pixel perturbation, such as LIME and Occlusion, the aim of which is to better understand the functioning of AI models (Ivanovs et al., 2021a). Therefore, the input image is repeatedly modified through blurring or random colors in specific regions of the image (Hendrycks and Dietterich, 2019). Furthermore, the results obtained are compared with the results of the original (undisturbed) input image. Therefore, the image region is considered significant if its removal results in a noticeable change in the result (Gupta et al., 2023).

Much has been said about the search for confidence in AI models. However, the need to align the explanations of XAI methods with the explanations of human beings demonstrates the concern to also increase confidence in explainable AI methods (Díaz-Rodríguez et al., 2022). Therefore, although XAI is presented as one of the solutions to the lack of transparency in AI models, it is also necessary to challenge the explanations generated by these methods, since these explanations can produce different results when subjected to the global and local processes of the models, for example (Ghassemi et al., 2021). According to (Doshi-Velez and Kim, 2017) we must be careful with interpretable methods, avoiding vague statements and considering factors relevant to the tasks performed and the method used. According to (Ghassemi et al., 2021), despite being attractive due to their explainability, XAI methods can have their explanations hindered by the presence of unrecognized confounding factors. It is also necessary to check that the results obtained by these methods are not altered when the AI model is subjected to external factors, such as changes in the input image.

Given this scenario, it is worth stating the importance of using explainable methods that provide reliable explanations to the user, based on concepts and studies related to the transparency of AI models. As well as the importance of examining the relationship between the input and output of a model based on pixel perturbation, identifying which part of the input the model attributes greater relevance to (Ivanovs et al., 2021a). This work aims to evaluate XAI methods by analyzing the influence of pixel perturbation. The experiments were applied to a model for segmenting regions of interest in Pacu fish. To train the model, a dataset containing the segmentation of 2000 different animals was required. MaskRCNN (He et al., 2017) from a feature extractor trained on ImageNet (Deng et al., 2009) based on the 18-layer variant of ResNet (He et al., 2015), (He et al., 2016) was the architecture applied to the original problem.

For the experiments, 100 fish images were used for three different disturbance techniques. In addition, four different explainable methods were used: Grad-CAM (Selvaraju et al., 2017), Saliency Map (Simonyan et al., 2013), Layer Grad-CAM (whose gradients resulting from Grad-CAM are calculated in re-

lation to the final convolution layer) (Chatterjee et al., 2022) and CNN Filters (Erhan et al., 2009). The results show which explainable model was the least sensitive to pixel perturbation techniques. In addition, it is possible to see which of the types of perturbation had the greatest influence on the model's predictive capacity, and whether these techniques affect each of the explainable methods differently.

2 RELATED WORKS

Disturbances make it possible to examine the relationship between the input and output of a model, allowing us to see which part of the input a model attributes greater importance to (Ivanovs et al., 2021b). Thereby, pixel perturbation can be used to assess both the accuracy of AI models (Kadir et al., 2023) and the explainability of explainable methods (Mohamed et al., 2022). In both cases, when there is influence from the changes made to the input, the good performance of the tool is proven. In the case of explainable methods, when this influence is not proven, it can be said that the explanations generated do not match the reality of the model's inference process. For example, to evaluate an explanation in terms of describing the model's behavior, there is a method that replaces pixels or regions of pixels based on the MoRF process, which is done in descending order based on their average relevance (Gumpfer et al., 2023).

The use of the perturbation technique extends to scenarios other than classification (Lin et al., 2021) or image segmentation (Gipiškis et al., 2023). In the works carried out by (Schlegel et al., 2019), (Veerappa et al., 2022) and (Abanda et al., 2022), for example, the technique is considered for input to time series models, and is even used as a way of assessing the explainability of XAI methods applied to this type of model. Therefore, it can 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 models that use images as input, the perturbation technique makes it possible to check for different types of noise in different types of images, such as (Shi et al., 2023). In real scenarios, such as medical images, these noises can be generated in different ways during image capture.

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 presented should be in line with the disturbance of the most relevant pixels, as can be seen in the figure shown. When these are modified, explainability should be influenced. Therefore, it can be said that an explainable method with good performance should be influenced by the disturbance made to the AI model input, even for images and visual explanations.

3 METHODOLOGY

The experiments were performed based on the steps shown in Figure 1. The codes developed and used in the process were created with the Pytorch library in Python, and executed in Google Collaboratory.

The segmentation model used was created from a Convolutional Neural Network (CNN), more specifically the MaskRCNN using as feature extractor a variant of ResNet with 18 layers. The goal of the model is to segment regions of interest in the fish such as head and body dimensions, and fin area for the purpose of phenotyping and subsequent genetic selection. For each fish image used as input, the model generates as output the classes, the masks of each fish region, a bounding box, and the confidence score in the class prediction.

In total, there were 100 fish images of the Pacu species provided by Laboratory of Genetics in Aquaculture and Conservation (LaGeAC) at Unesp Jaboticabal. Each image was manually segmented using a specific segmentation tool called LabelBox and then submitted to the AI model for inference. From this step, a black and white mask was created whose white areas represent the fish.

XAI methods were applied to explain the most important regions for the MaskRCNN prediction model. Therefore, four methods were applied: Grad-CAM, Saliency Map, CNN Filters and Layer Grad-CAM. Both highlight the image regions of greatest relevance for the model to identify the fish regions. Thus, it is possible to verify whether the result presented by CNN matches the important regions highlighted by the explainable methods.

The efficiency of the prediction process was tested from the implementation of pixel perturbation techniques, whose objective was to perturb the regions highlighted as important for the model from the XAI methods. As such, the perturbation was applied from the scores obtained for each pixel in the explainability methods. Each method considers a different scale to highlight the most relevant regions, as presented in Table 1. To identify borderline values for each of the methods, visual inspection was used on some of the dataset images.

From the determined values, three different pixel perturbation techniques were implemented. These

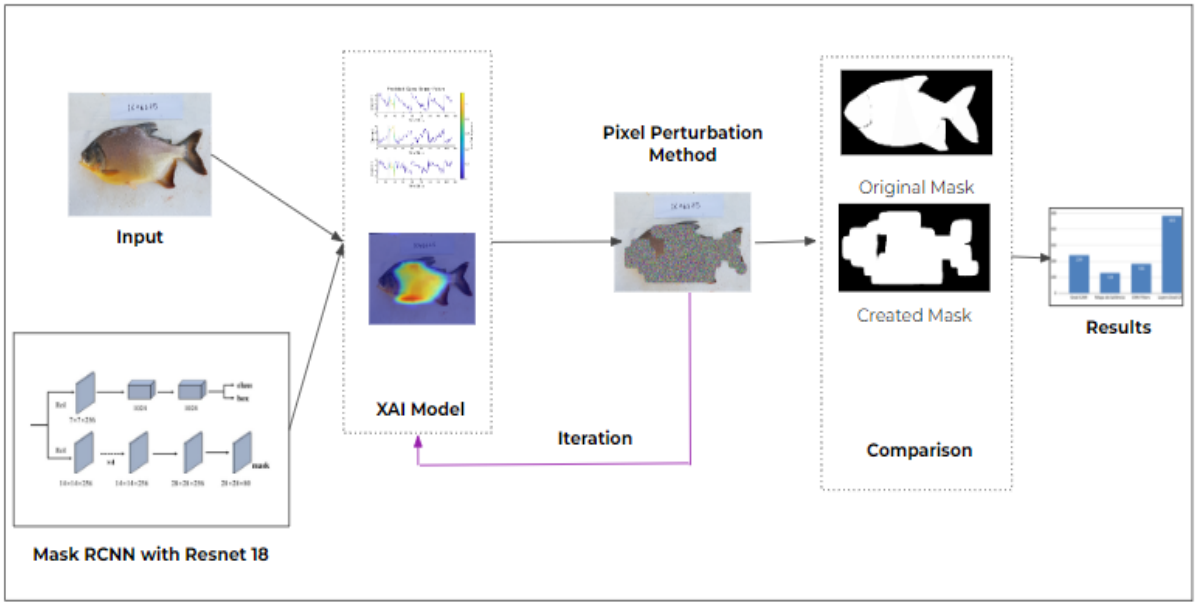

Figure 1: Diagram of the methodology used during the experiments.

Table 1: Table of most important pixel values for each XAI method.

XAI Methods	Values
Grad-CAM	Greater than 0,01
Saliency Map	Greater than 0,045
Layer Grad-CAM	Greater than -0,5
CNN Filters	Less than 0

Figure 2: Pixel perturbation techniques applied on the same fish by the Grad-CAM method.

are: white noise, black color noise and random noise. Each was applied to the input image from the most relevant regions identified by the previously presented XAI methods. Figure 2 illustrates the different perturbations applied to an image from regions extracted using the Grad-CAM method.

Next, it was necessary to perform a comparison of the original mask, generated during manual segmentation, with the mask generated during the model prediction after successive perturbations in the pixels. The latter was also created in black and white, whose white regions are equivalent to the important regions highlighted in the previous step, as illustrated in Figure 3.

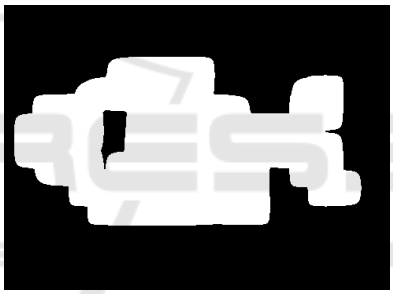

Figure 3: Mask generated during prediction from the Grad-CAM method.

The comparison was performed based on two indices: Intersection under Union (IoU) and Sorensen Dice (SD). Both aim to verify how similar the images to be compared are. In this case, how similar the original mask is to the mask generated during prediction after pixel perturbation. The results obtained were tabulated according to the perturbation technique and explainability method used.

All steps were performed in a maximum of five iterations per image. In some cases, the model itself stopped the process even before reaching the maximum. In these situations, the segmentation model was not able to perform the prediction after the pixel perturbation, thus interrupting the process.

For each iteration, the image used as input was the image generated as output by the previous iteration, except in the first iteration whose input was the original image of the fish. In this way, with each pass, the most relevant pixels were increasingly perturbed

with the intention of making the prediction even more difficult and consequently impairing the model's predictive ability. In Figure 4, the red arrows indicate the regions that were disturbed in the second iteration, increasing the disturbed region compared to the first iteration.

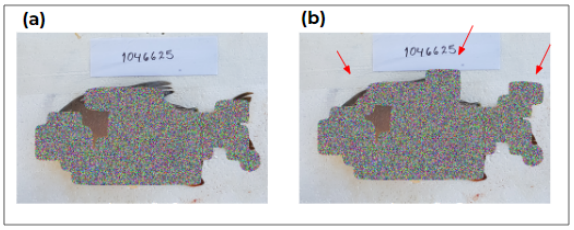

Figure 4: (a) is the image of the fish in the first iteration of the Grad-CAM method with random pixel disturbance, and (b) is the image of the fish in the second iteration of the Grad-CAM method with random pixel pertubation.

4 RESULTS

From the presented experiments, it was possible to determine the influence of the perturbation techniques on the XAI methods. Therefore, by observing the IoU and SD indices, it was possible to verify the different behavior of the methods. The indices vary approximately from 0 to 0.6, as presented 5. For the Grad-CAM method, in both indexes, the variation was considerable in relation to the other methods. It is worth mentioning the behavior of the Layer Grad-CAM method that presented a discrete variation in the IoU index and for the SD index it presented an opposite behavior, being the method with the greatest variation for this index. These variations show that the masks (original and prediction) suffered significant changes with the disturbance of pixels, making them increasingly different. This process happened in a single iteration or along the five iterations depending on the combination of the perturbation method with the XAI technique used.

To determine the results, besides the IoU and SD indexes, the number of iterations and number of images generated for each method were observed. It should be noted that the number of iterations determined the number of images generated throughout the experiments. Thus, for each input image, a maximum of five output images were generated, depending on the explainable method and perturbation technique applied.

Regarding the methods that interrupted the prediction process even before reaching the maximum of five iterations, it means that the combination of the XAI method with the pixel perturbation technique affected the model, and therefore impeded the prediction process. Therefore, the Saliency Map obtained the lowest average number of iterations while demonstrating a lower sensitivity to different types of pixel disturbance, making it the best method among the four.

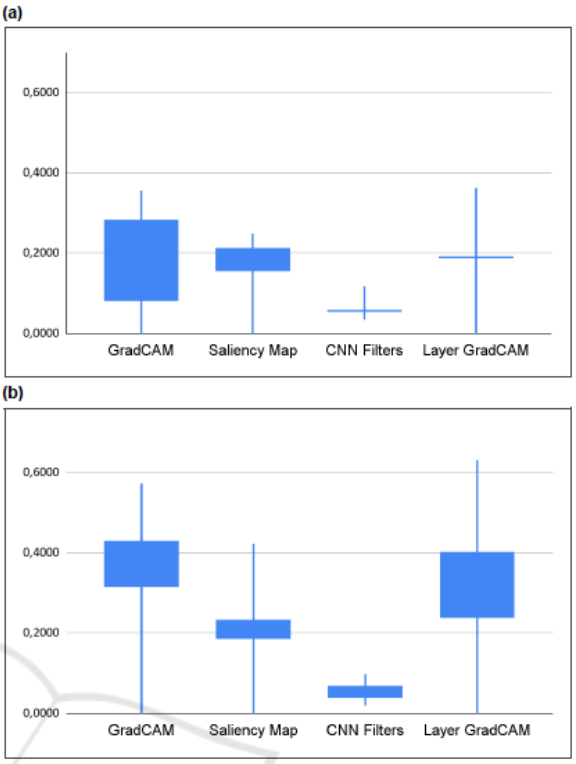

Figure 5: Boxplots of (a) IoU and (b) SD indices from XAI methods.

For the methods that reached the maximum number of iterations, this means that the model continued to be able to perform the segmentation even after the most relevant pixels were disturbed according to the method. That is, the perturbation did not affect the model to the extent that it continued to perform the prediction process. For example, the Layer Grad-CAM method that presented the maximum average number of iterations, becoming the worst among the four methods tested. These results can be observed through the graph in Figure 6, which shows the average number of iterations for each XAI method.

When analyzing each perturbation technique individually, it was possible to observe that each of them influences the same XAI method in different ways. As presented in Table 2, the Grad-CAM method, for example, obtained a significant variation in the number of images from one perturbation technique to another. It is also possible to observe this same variation in the average amount of iterations, as shown in Table 3.

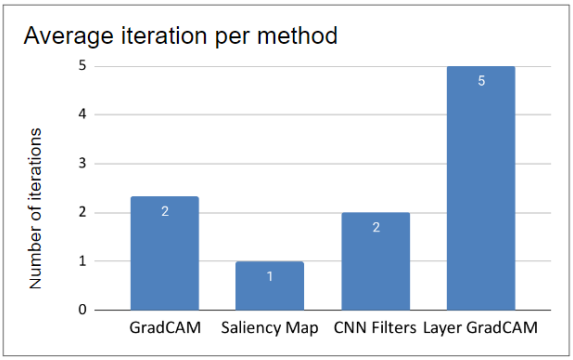

Figure 6: Bar graph of the average number of iterations per XAI method.

The other methods (Saliency Map, CNN Filters, and Grad-CAM), for example, did not vary much between perturbation techniques. Consequently, it is possible to understand that the methods are influenced in different ways according to the pixel perturbation technique applied.

Table 2: Average amount of images generated in each XAI method according to the pixel disturbance technique applied.

XAI Methods	Pertubation	Images
Grad-CAM	Random	224
	Black	374
	White noise	120
Saliency Map	Random	123
	Black	144
	White noise	118
Layer Grad-CAM	Random	493
	Black	499
	White noise	464
CNN Filters	Random	188
	Black	182
	White noise	184

From the implemented pixel perturbation methods, it is possible to observe that three of the four XAI methods suffered more influence from the black coloration technique, if compared to the other white noise and random techniques, as illustrated in Figure 7. It is possible that this phenomenon is related to the use of padding, which is usually implemented with pixels of value 0, thus making the model more resilient to perturbations of this nature. Therefore, for these methods, the number of iterations and images generated was higher for this perturbation technique than for the others. However, the CNN Layers method

Table 3: Average number of iterations by each XAI method according to the applied pixel perturbation technique.

XAI Methods	Pertubation	Iterations
Grad-CAM	Random	2
	Black	4
	White noise	1
Saliency Map	Random	1
	Black	1
	White noise	1
Layer Grad-CAM	Random	5
	Black	5
	White noise	5
CNN Filters	Random	2
	Black	2
	White noise	2

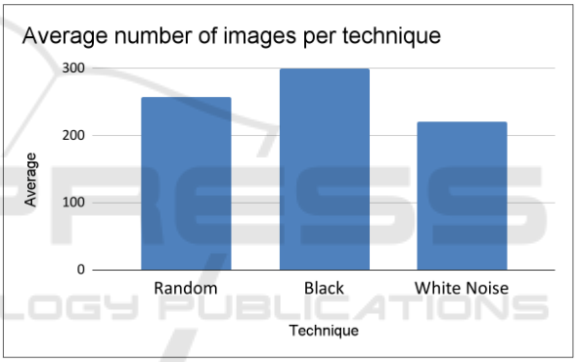

Figure 7: Bar graph of the average number of images per pixel disturbance technique.

was the only one to show less sensitivity to this perturbation technique, becoming the method with the least variation among the techniques, when considering the average number of images.

5 CONCLUSIONS

XAI methods are seen as solutions to the lack of transparency in AI models. The objective of these methods is to highlight which are the most relevant regions of an image during model prediction. Based on this and the experiments carried out in this work, the influence on the MaskRCNN prediction from pixel perturbation techniques and on the outputs obtained by explainability methods applied to the segmentation of images of fish of the Pacu species was verified. Thereby, through the results presented, it is possible to identify that the perturbation methods influence the results

of the tested explainability models, and also that the perturbation techniques interfere with the XAI methods in different ways, such as the Grad-CAM method, which had different behaviors. for each technique tested.

The black color perturbation technique generated more iterations, and consequently more images, in three of the four XAI methods presented. Therefore, it can be concluded that this technique has the least capacity to generate impact on the segmentation model among the techniques tested in this work. Still according to the average number of iterations and the average number of images generated, it is possible to conclude that the Saliency Map method was the least sensitive to the different perturbation methods and, therefore, the best among the XAI methods tested on the problem in question. The CNN Filters method was the least sensitive to the types of disturbance, presenting less variation in the average number of images, while the Grad-CAM method was the most sensitive among the four.

For future work, it is suggested that the experiments presented here be replicated in other AI models and other explainability methods, as well as in other scenarios beyond image segmentation. Furthermore, it is interesting to test other existing perturbation techniques and their combinations with explainability methods to identify their influence on the predictive capacity of the models. Finally, more evidence about this influence can be gathered and from this, it can be quantified which combination of the XAI method and pixel perturbation is best for a given problem.

REFERENCES

Abanda, A., Mori, U., and Lozano, J. (2022). Ad-hoc explanation for time series classification. *Knowledge-Based Systems*, 252:109366.

Arnout, H., El-Assady, M., Oelke, D., and Keim, D. A. (2019). Towards a rigorous evaluation of xai methods on time series. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4197–4201.

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell*, 173(7):1581–1592.

Chatterjee, S., Das, A., Mandal, C., Mukhopadhyay, B., Vipinraj, M., Shukla, A., Nagaraja Rao, R., Sarasaen, C., Speck, O., and Nürnberger, A. (2022). Torchesegeta: Framework for interpretability and explainability of image-based deep learning models. *Applied Sciences*, 12(4):1834.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Deng, L. and Yu, D. (2014). Deep learning: methods and applications. *Foundations and trends in signal processing*, 7(3–4):197–387.

Díaz-Rodríguez, N., Lamas, A., Sanchez, J., Franchi, G., Donadello, I., Tabik, S., Filliat, D., Cruz, P., Montes, R., and Herrera, F. (2022). Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. *Information Fusion*, 79:58–83.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1.

Fellous, J.-M., Sapiro, G., Rossi, A., Mayberg, H., and Ferrante, M. (2019). Explainable artificial intelligence for neuroscience: Behavioral neurostimulation. *Frontiers in Neuroscience*, 13. cited By 0.

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750.

Gipiškis, R., Chiaro, D., Preziosi, M., Prezioso, E., and Piccialli, F. (2023). The impact of adversarial attacks on interpretable semantic segmentation in cyber-physical systems. *IEEE Systems Journal*.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Gumpfer, N., Prim, J., Keller, T., Seeger, B., Guckert, M., and Hannig, J. (2023). Signed explanations: Unveiling relevant features by reducing bias. *Information Fusion*, page 101883.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2:2.

Gunning, D. and Aha, D. (2019). Darpa's explainable artificial intelligence program. *AI Magazine*, 40(2):44–58. cited By 6.

Gupta, L. K., Koundal, D., and Mongia, S. (2023). Explainable methods for image-based deep learning: a review. *Archives of Computational Methods in Engineering*, 30(4):2651–2666.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. (2017). Mask R-CNN. *CoRR*, abs/1703.06870.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. *CoRR*, abs/1603.05027.

Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Ivanovs, M., Kadikis, R., and Ozols, K. (2021a). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234.

Ivanovs, M., Kadikis, R., and Ozols, K. (2021b). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234.

Kadir, M. A., Mohamed Selim, A., Barz, M., and Sonntag, D. (2023). A user interface for explaining machine learning model explanations. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 59–63.

Kaufman, D. (2019). *A inteligência artificial irá suplantar a inteligência humana?* ESTAÇÃO DAS LETRAS E CORES EDI.

Kistan, T., Gardi, A., and Sabatini, R. (2018). Machine learning and cognitive ergonomics in air traffic management: Recent developments and considerations for certification. *Aerospace*, 5(4). cited By 3.

Lin, Y.-S., Lee, W.-C., and Celik, Z. B. (2021). 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*, pages 1027–1035.

Mohamed, E., Sirlantzis, K., and Howells, G. (2022). A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. *Displays*, 73:102239.

Pinheiro, P. P. (2020). *Proteção de dados pessoais: Comentários à lei n. 13.709/2018-lgpd*. Saraiva Educação SA.

Russell, S. J. and Norvig, P. (2004). *Inteligência artificial*. Elsevier.

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., and Keim, D. A. (2019). Towards a rigorous evaluation of xai methods on time series. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4197–4201. IEEE.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Shi, R., Li, T., and Yamaguchi, Y. (2023). Understanding contributing neurons via attribution visualization. *Neurocomputing*, page 126492.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Teixeira, J. (2019). *O que é inteligência artificial*. E-Galáxia.

Veerappa, M., Anneken, M., Burkart, N., and Huber, M. F. (2022). Validation of xai explanations for multivariate time series classification in the maritime domain. *Journal of Computational Science*, 58:101539.

Weber, R. O., Johs, A. J., Li, J., and Huang, K. (2018). Investigating textual case-based XAI. In Cox, M. T., Funk, P., and Begum, S., editors, *Case-Based Reasoning Research and Development*, volume 11156, pages 431–447. Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Wolf, C. T. (2019). Explainability scenarios: Towards scenario-based xai design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 252–257, New York, NY, USA. Association for Computing Machinery.

Wolf, C. T. and Ringland, K. E. (2020). Designing accessible, explainable ai (xai) experiences. *SIGACCESS Access. Comput.*, (125).

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In Tang, J., Kan, M.-Y., Zhao, D., Li, S., and Zan, H., editors, *Natural Language Processing and Chinese Computing*, volume 11839, pages 563–574. Springer International Publishing. Series Title: Lecture Notes in Computer Science.