Quantitative Evaluation of White & Black Box Interpretability Methods for Image Classification

Oscar Stanchi[1,3][0000−0003−0294−2053], Franco Ronchetti[1,2][0000−0003−3173−1327], Pedro Dal Bianco[1,4][0000−0001−7197−8602], Gastón Rios[1,4][0000−0003−0252−7036], Waldo Hasperué[1,2][0000−0002−9950−1563], Domènec Puig[5][0000−0002−0562−4205], Hatem Rashwan[5][0000−0001−5421−1637], and Facundo Quiroga[1,2][0000−0003−4495−4327]

[1] Instituto de Investigación en Informática LIDI - Universidad Nacional de La Plata, La Plata, Argentina

[2] Comisión de Investigaciones Científicas de la Pcia. de Bs. As. (CIC-PBA), La Plata, Argentina

[3] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), La Plata, Argentina

[4] Universidad Nacional de La Plata (UNLP), La Plata, Argentina

[5] Department of Computer Engineering and Mathematics - Universitat Rovira i Virgili, Tarragona, España

✉ ostanchi@lidi.info.unlp.edu.ar

Abstract. The field of interpretability in Deep Learning faces significant challenges due to the lack of standard metrics for systematically evaluating and comparing interpretability methods. The absence of quantifiable measures impedes practitioners ability to select the most suitable methods and models for their specific tasks. To address this issue, we propose the Pixel Erosion and Dilation Score, a novel metric designed to assess the robustness of model explanations. Our approach involves applying iterative erosion and dilation processes to heatmaps generated by various interpretability methods, thereby using them to hide and show the important regions of a image to the network, allowing for a coherent and interpretable evaluation of model decision-making processes. We conduct quantitative ablation tests using our metric on the ImageNet dataset with both VGG16 and ResNet18 models. The results reveal that our new measure provides a numerical and intuitive means for comparing interpretability methods and models, facilitating more informed decision-making for practitioners.

Keywords: Ablation · Black Box · Computer Vision · Deep Learning · Interpretability · Quantitative Measure · White Box.

1 Introduction

Deep Learning models are frequently perceived as black boxes due to their intricate structures and limited interpretability [6]. This challenge, though not new

[17], has recently garnered widespread attention from various sectors, including researchers and policymakers [18,13]. Nonetheless, the importance of making these models interpretable is increasingly recognized, with ongoing research dedicated to elucidating the reasoning behind their outputs [6,4]. Gaining this understanding is essential for build trust among users and stakeholders [11].

Feature attribution methods aim to assign an importance value to each feature based on its contribution to the prediction. These methods are arguably the most extensively studied and benchmarked within interpretable Deep Learning, as seen in [2,12]. Commonly, these methods generate importance maps, which serve as visualization tools to highlight critical regions within an input image that influence the model's decisions. These maps are typically presented as heatmaps [15]. Nevertheless, these maps do not clarify how the model uses this relevant information for its predictions [12]. Deep Learning models with similar accuracy can produce vastly different attribution maps [19,21].

However, interpretability is fundamentally a subjective issue [9,12]. Explanations are context-dependent and the perceived quality of an explanation is influenced by the backgrounds of both the provider and the receiver, as well as the type of information that interests the receiver [8]. Most interpretability research uses qualitative measures that complicate cross-study comparisons [16,12]. Also, a common assumption among authors [7,20] is that the explanations being evaluated are true, as discussing a false explanation would be meaningless from an epistemological standpoint.

Yet another issue in this field is that determining how different methods can be systematically compared remains uncertain without relying on qualitative evaluations or user-studies. Hence, the absence of standard metrics for evaluating interpretability quality poses challenges for practitioners selecting among various interpretability methods and Deep Learning models [13]. Evaluating, validating, comparing, and improving these aspects require quantifiable metrics [13,1,5]. Therefore, utilizing quantitative indicators can assist practitioners in choosing the most suitable method for their specific tasks [9]. These are preferable due to their numerical nature, as they provide an intuitive means of comparing different explanations through proxy metrics [3]. The absence of measures in this field will hinder efficient development in the field, ultimately impeding its rapid growth [9,13]. In Figure 1, we illustrate various interpretability methods that generate heatmaps for an image, highlighting the challenge of determining the most appropriate approach.

Explanations must be robust, sensitive to both model and data, and consistent. The robustness of attribution maps from different methods cannot be judged solely through qualitative means. Thus, careful evaluation by quantitative metrics of post-hoc explanations is necessary before they can be integrated into critical workflows [12]. Various metrics and evaluation strategies have been proposed to address the absence of ground truth for explanations, but it is impossible to devise quantitative metrics applicable to all interpretability methods [22].

Previous works have discussed current approaches to measuring and comparing the interpretability of different methods. Several authors have focused on specific

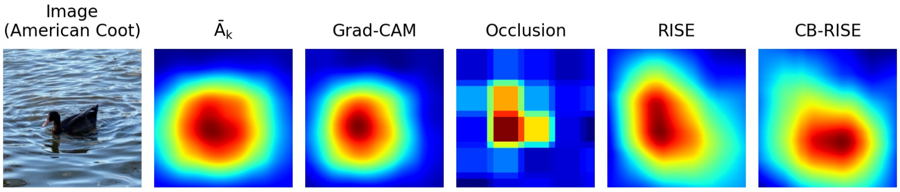


Fig. 1: Illustration of interpretability methods for an image classified as an American Coot using a ResNet18 architecture. Each column demonstrates a different interpretability method (except for the first, which shows the original image): $\bar{A}_k$ average activations from the last convolutional layer, GradCAM, Occlusion, RISE, and CB-RISE [15,12].

metrics for high-risk systems, such as clinical workflows, to address the actual limitations of this field [12].

In a more general context, there are studies like Petsiuk et al. [10], Wilson et al. [14], Schmidt et al. [13], and Nguyen et al. [9] that contribute to this discussion by providing quantitative assessments and methodologies for evaluating interpretability across various domains. Petsiuk et al. [10] proposed Causal Metrics for Explanations, specifically the *deletion* and *insertion* metrics. These metrics automatically evaluate explanations by measuring changes in the model's predicted probability when important pixels are removed or introduced. The deletion metric measures the decrease in probability as important pixels are removed, while the insertion metric measures the increase in probability as important pixels are added.

In this article, we propose the metric **Pixel Erosion and Dilation Score** (PE/DS), which evaluates the robustness of model explanations through iterative *erosion* and *dilation* processes applied to the heatmaps generated by interpretability methods. The core idea of our approach is to assess how the model's predictions change as we progressively remove or add relevant pixels identified by the heatmap.

The metric proposed in [10] aligns closely with the approach we propose, as they provide an objective evaluation free from human bias and more accurately reflect the model's decision-making process. Our method, however, uses morphological erosion and dilation, which offers a more structured approach compared to Petsiuk's previous method. In the prior approach, the heatmap is flattened into a one-dimensional array and sorted by pixel importance. Then, pixels are sequentially occluded based on their importance, starting with the most significant ones. By using morphological operations, our method maintains the spatial structure of the heatmap, leading to a more coherent and interpretable evaluation of the model's decision-making process.

This manuscript is organized as follows. Section 1 introduces the topic and reviews related work, highlighting the research motivation and objectives. Section 2 provides a comprehensive description of our proposed metric, including a detailed explanation of its operation. In Section 3, we present the experimental setup and the quantitative and aggregated results used to evaluate the interpretability of different architectures. The paper concludes in Section 4, where we summarize our findings and propose directions for future research.

2 Pixel Erosion and Dilation Score (PE/DS)

The **PE/DS** measure tests the robustness of model explanations by applying iterative morphological erosion and dilation to heatmaps, evaluating how model predictions change as relevant pixels are added or removed.

The **erosion** process begins by applying a threshold to the heatmap to create a binary mask, highlighting the most relevant pixels. This mask is then iteratively eroded, reducing the number of white pixels (relevant areas) until the specified fraction of remaining pixels is reached. At each iteration, we calculate the model's output by doing element-wise multiplication with the original image and the current eroded mask. This allows us to record the model's response to the eroded heatmap at each step, capturing the relationship between the number of relevant pixels and the model's output. The erosion process is illustrated in Figure 2.

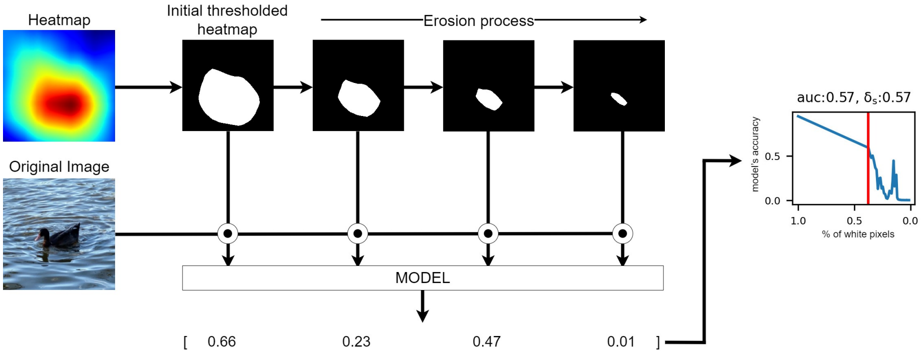


Fig. 2: Erosion process applied to the heatmap. The figure displays the initial heatmap, the resulting binary mask after the threshold is applied, and the effect of iterative erosion on the mask, along with the model's response as the number of relevant pixels decreases.

Similarly, the **dilation** process starts with the binary mask and iteratively adds pixels until the desired fraction of white pixels is achieved. The model's output is calculated at each step, providing a complementary perspective to the erosion process by observing how the inclusion of white pixels influences the model's predictions.

This new metric quantifies the changes in model output by plotting the model's prediction against the fraction of white pixels. The area under the curve

(AUC) is calculated as an aggregation metric, summarizing the overall change in model output across the erosion and dilation processes. The calculation of the AUC is relative to the curve, allowing comparison of which method causes a more rapid decline or increase in the curve. For the erosion process, a sharp decline in the probability curve, resulting in a lower AUC, indicates a more effective explanation, as it shows that removing relevant pixels significantly impacts the model's predictions. Conversely, a slower decrease in the AUC suggests a less impactful explanation. For the dilation process, the opposite is true. This comprehensive measure provides insight into how sensitive the model's predictions are to the inclusion or exclusion of key pixels, offering a robust evaluation of the explanation quality.

Therefore, the **PE/DS** metric not only ensures a more objective evaluation by removing human biases but also provides a detailed understanding of the impact of relevant pixels on the model's decision-making process.

Additionally, our approach is significantly more efficient; while the previous method [10] removes or adds pixels one by one, our morphological operations remove or add multiple pixels at each step. This results in fewer iterations and model inferences, enhancing computational efficiency. Although this may lead to a less smooth curve, the gain in efficiency is substantial, making the process faster and more scalable[6].

3 Experiments and Results

3.1 Experiments

For the quantitative evaluation using our automatic evaluation metric, we conducted an ablation test using 17 images from the ImageNet dataset. We tested two architectures: ResNet18, with 11,689,512 parameters, and VGG16, with 138,357,544 parameters. The images underwent preprocessing, including cropping to a size of 224 × 224, which is the standard input dimension for ImageNet. Additionally, we applied z-score normalization to the input images, using a mean of $[0.485, 0.456, 0.406]$ and a standard deviation of $[0.229, 0.224, 0.225]$. All experiments were conducted on an NVIDIA GTX 1060 GPU.

We evaluated the robustness of five interpretability methods: $\bar{A}_k$ average activations from the last convolutional layer, GradCAM, Occlusion, RISE, and CB-RISE. The following settings were used: for Grad-CAM, the last convolutional layer of both networks was utilized; for RISE, 2048 iterations (masks) were performed and we used a `initial_mask_size` of 4 × 4; and for CB-RISE, the parameters were `isotropic_sigma=10`, `patience=64`, `d_epsilon=3`, and `threshold=0.3`. The `threshold` parameter used for **PE/DS** was 0.5, and the algorithm stops either after a maximum of 100 iterations or when the desired amount of white pixels left (1%) is reached.

[6] The code for the **PE/DS** metric and the complete results from the testing can be found in the repository: https://github.com/indirivacua/peds-measure/tree/main

3.2 Quantitative Results

Figure 3 presents the results of the quantitative ablation test performed on 3 ImageNet images using both VGG16 and ResNet18 models. The top subfigures displays the erosion 3a and dilation 3b processes applied to heatmaps generated by VGG16, while the bottom subfigures shows the same processes (3c and 3d) applied to ResNet18. Each subfigure includes the input image and columns representing the five different interpretability methods selected. For each method, the corresponding **PE/DS** metric graph and AUC value are shown. The red line in each graph represents the percentage of white pixels (x-axis) and the model's accuracy (y-axis) when the thresholded heatmap is applied to the image during the **PE/DS** metric calculation. δ_s denotes the slope between the output with the original image and the output after applying the first threshold. This illustrates how the model's predictions change as key pixels are progressively removed or added.

Particularly for the results of RISE algorithm shown in Figure 3a:

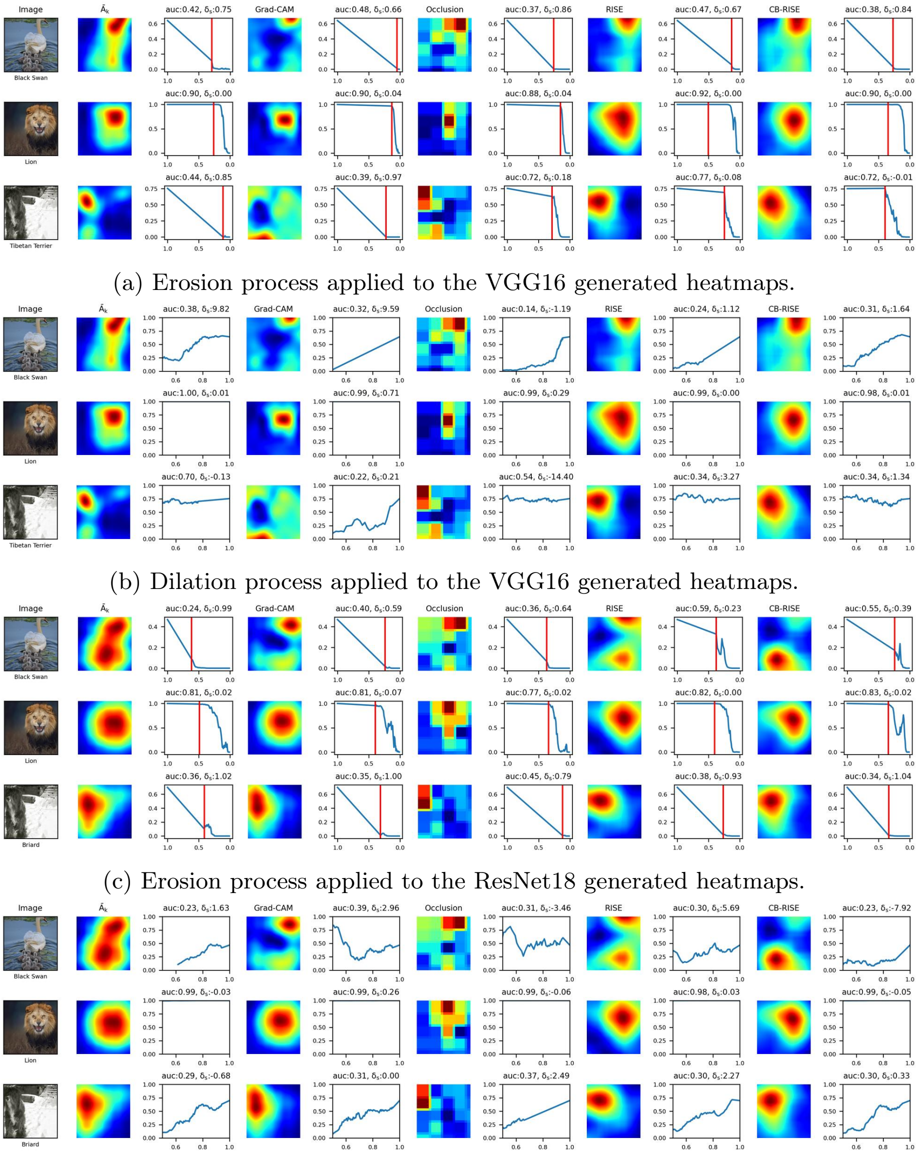
- The heatmap generated for the Tibetan Terrier is considered effective. Once the pixels are removed with the threshold of 0.5, the accuracy decreases very slightly at first and then begins to drop sharply after a few iterations. This indicates that the important features for classification are well-concentrated.
- The heatmap generated for the Lion could be improved. When pixels are removed starting from the 0.5 threshold, it takes a considerable number of iterations before there is a noticeable reduction in accuracy. A heatmap more concentrated on the face of the Lion would be more effective in this case.
- The heatmap generated for the Black Swam is deemed ineffective. Immediately after applying the 0.5 threshold and blending it with the image, the accuracy significantly drops. This suggests that the heatmap does not accurately capture the critical features needed for classification.

These observations are not exclusive to the RISE method and can be extrapolated to the other interpretability methods as well. However, the AUC does not always reflect these nuances accurately, highlighting a known limitation and area for improvement in this metric.

Therefore, if the curve generated by the **PE/DS** metric drops quickly for the erosion process, it indicates that the generated importance map is poorer, as it removed pixels that were essential for classification. This is a critical insight provided by the **PE/DS** metric in evaluating the quality of heatmaps. For the dilation process, it is evident that this operation introduces a certain level of "noise" to the image by revealing more pixels to the model as the heatmap dilates. This is reflected in the resulting graph, which often exhibits small jumps and lacks smoothness, indicating the perturbation caused by the additional information being presented to the model.

3.3 Aggregation Results

Table 1 presents the aggregation mean of quantitative results of our ablation tests using our **PE/DS** metric on both VGG16 and ResNet18 architectures, evaluated



(a) Erosion process applied to the VGG16 generated heatmaps.

(b) Dilation process applied to the VGG16 generated heatmaps.

(c) Erosion process applied to the ResNet18 generated heatmaps.

(d) Dilation process applied to the ResNet18 generated heatmaps.

Fig. 3: Quantitative ablation test results using VGG16 and ResNet18 on 3 ImageNet images. The top subfigures shows erosion (a) and dilation (b) processes on VGG16 heatmaps, while the bottom subfigures displays erosion (c) and dilation (d) processes on ResNet18 heatmaps. Each subfigure shows the input image, followed by columns representing interpretability methods: $\bar{A}_k$ average activations from the last convolutional layer, GradCAM, Occlusion, RISE, and CB-RISE. The **PE**/**DS** metric and AUC value for each method are displayed.

Table 1: Averaged AUC scores for different interpretability methods applied to VGG16 and ResNet18 models. The table compares the effectiveness of white-box and black-box methods, with the iterations required for each method included. Higher Erosion AUC and lower Dilation AUC values indicate better interpretability.

Model	**Method**	**Black-Box**	**Iterations**	**Erosion AUC ↑**	**Dilation AUC ↓**
VGG16	Activations	X	1	0.694335	0.673663
	Grad-CAM	X	2	0.664539	0.672096
	Occlusion	✓	26	0.696151	0.666154
	RISE	✓	4096	**0.753361**	**0.558094**
	CB-RISE	✓	1566.12 (avg)	0.743734	0.587257
ResNet18	Activations	X	1	0.632872	**0.455677**
	Grad-CAM	X	2	**0.664206**	0.485391
	Occlusion	✓	26	0.570779	0.578191
	RISE	✓	4096	0.646453	0.461413
	CB-RISE	✓	1615.06 (avg)	0.653912	0.466274

with the aforementioned interpretability methods. The table includes the Erosion AUC and Dilation AUC mean scores for the 17 images from ImageNet, which indicate the effectiveness of each method in capturing the critical features of the input images.

The results demonstrate that ResNet18 is generally less interpretable compared to VGG16, as evidenced by lower Erosion AUC. For VGG16, the black-box method RISE comes out on top, followed by CB-RISE and Occlusion, with white-box methods such as Activations and Grad-CAM performing the least effectively. This highlights that the choice of interpretability method can greatly depend on the specific model being used, with black-box methods proving to be more effective in certain architectures like VGG16. For ResNet18, white-box methods such as Activations and Grad-CAM tend to perform slightly better than black-box methods like Occlusion, RISE, and CB-RISE. However, the differences in performance are not substantial.

White-box methods often yield better interpretability results for ResNet18, but they come with certain disadvantages. These methods require access to the internal workings of the model, which can limit their applicability and flexibility. In contrast, black-box methods, despite showing slightly lower interpretability scores, offer a more general approach that does not rely on model-specific information, making them more versatile for different neural network architectures.

4 Conclusions and Future Work

4.1 Conclusions

The challenges associated with the absence of standard metrics for evaluating interpretability quality in Deep Learning models are significant. This paper addresses these challenges by introducing the **Pixel Erosion and Dilation Score** as a robust metric for evaluating model explanations. Through the application of iterative erosion and dilation processes on heatmaps generated by various interpretability methods, our **PE/DS** metric provides a quantifiable and intuitive means to assess the robustness of model predictions.

Our quantitative ablation tests demonstrate the efficacy of this approach. The results indicate that VGG16 generally exhibits higher interpretability compared to ResNet18, as evidenced by our **PE/DS** metric, that effectively highlights the quality of the heatmaps by indicating how well the important features are captured and concentrated.

The proposed **PE/DS** metric not only facilitates a more systematic comparison of interpretability methods but also aids practitioners in selecting the most suitable approach for their specific tasks. By offering a numerical and intuitive way to evaluate interpretability, our method contributes to the advancement and more efficient development of the field, ultimately fostering its rapid growth.

4.2 Future Work

An avenue for future research is enhance our **PE/DS** metric and its applicability. Firstly, integrating additional types of perturbations into the metric could offer more comprehensive evaluations. For instance, applying blurring techniques or altering color schemes could provide new insights into the robustness of model explanations. Another potential improvement is the automation of threshold selection for the **PE/DS** metric. Currently, a fixed threshold is used, but optimizing this threshold could lead to more accurate results. Additionally, it would be valuable to develop methods for comparing model explanations and to validate the proposed metric with human judgment [5]. Finally, we plan to systematically apply the **PE/DS** metric to a range of Deep Learning models to identify potential issues, such as overfitting.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)
2. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv preprint arXiv:1711.06104 (2017)
3. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)
4. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
5. Hakkoum, H., Abnane, I., Idri, A.: Interpretability in the medical field: A systematic mapping and review study. Applied Soft Computing **117**, 108391 (2022)
6. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
7. Lombrozo, T.: Explanatory preferences shape learning and inference. Trends in cognitive sciences **20**(10), 748–759 (2016)

8. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
9. Nguyen, A.p., Martínez, M.R.: On quantitative aspects of model interpretability. arXiv preprint arXiv:2007.07584 (2020)
10. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
11. Ribeiro, M.T., Singh, S., Guestrin, C.: " 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. pp. 1135–1144 (2016)
12. Salahuddin, Z., Woodruff, H.C., Chatterjee, A., Lambin, P.: Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in biology and medicine **140**, 105111 (2022)
13. Schmidt, P., Biessmann, F.: Quantifying interpretability and trust in machine learning systems. arXiv preprint arXiv:1901.08558 (2019)
14. Silva, W., Fernandes, K., Cardoso, M.J., Cardoso, J.S.: Towards complementary explanations using deep neural networks. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1. pp. 133–140. Springer (2018)
15. Stanchi, O., Ronchetti, F., Dal Bianco, P., Rios, G., Ponte Ahon, S., Hasperué, W., Quiroga, F.: Cb-rise: Improving the rise interpretability method through convergence detection and blurred perturbations. In: Conference on Cloud Computing, Big Data & Emerging Topics. Springer (2024)
16. Strumbelj, E., Kononenko, I.: An efficient explanation of individual classifications using game theory. The Journal of Machine Learning Research **11**, 1–18 (2010)
17. Thagard, P.: Philosophical and computational models of explanation. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition **64**(1), 87–104 (1991)
18. Vishwanath, T., Kaufmann, D.: Toward transparency: New approaches and their application to financial markets. The World Bank Research Observer **16**(1), 41–57 (2001)
19. Yilmaz, E.B., Mader, A.O., Fricke, T., Peña, J., Glüer, C.C., Meyer, C.: Assessing attribution maps for explaining cnn-based vertebral fracture classifiers. In: Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3. pp. 3–12. Springer (2020)
20. Ylikoski, P., Kuorikoski, J.: Dissecting explanatory power. Philosophical studies **148**, 201–219 (2010)
21. Young, K., Booth, G., Simpson, B., Dutton, R., Shrapnel, S.: Deep neural network or dermatologist? In: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9. pp. 48–55. Springer (2019)
22. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5), 593 (2021)