Discover Artificial Intelligence

Discover

Case Study

Improving deep learning performance by using Explainable Artificial Intelligence (XAI) approaches

Vitor Bento[1] · Manoela Kohler[1] · Pedro Diaz[1] · Leonardo Mendoza[2] · Marco Aurelio Pacheco[1]

Received: 5 April 2021 / Accepted: 3 August 2021
Published online: 06 October 2021
© The Author(s) 2021 OPEN

Abstract
In this work we propose a workflow to deal with overlaid images—images with superimposed text and company logos—, which is very common in underwater monitoring videos and surveillance camera footage. It is demonstrated that it is possible to use Explaining Artificial Intelligence to improve deep learning models performance for image classification tasks in general. A deep learning model trained to classify metal surface defect, which previously had a low performance, is then evaluated with Layer-wise relevance propagation—an Explaining Artificial Intelligence technique—to identify problems in a dataset that hinder the training of deep learning models in a wide range of applications. Thereafter, it is possible to remove this unwanted information from the dataset—using different approaches: from cutting part of the images to training a Generative Inpainting neural network model—and retrain the model with the new preprocessed images. This proposed methodology improved F1 score in 20% when compared to the original trained dataset, validating the proposed workflow.

Keywords Explain Artificial Intelligence · XAI · Layer-wise relevance propagation · Deep learning

1 Introduction

Deep learning classifiers [1, 2] are being widely used in a vast range of applications with different objectives in areas such as scientific studies, industry and entertainment [3–5] successfully. Despite the revolutionary character of this technology, there are still challenges that diminish its expansion or prevent the consolidation of deep learning in certain areas. Some of the main challenges to be overcome are the great complexity of models that require high computational cost [6] as well as the lack of transparency and explicability [7–9], which weaken the confidence and verifiability of decisions taken by a deep learning system.

The absence of explicability and transparency in certain areas is not invariably a problem since state-of-the-art models have an extremely high accuracy [10]. Furthermore, any errors, to a large extent, do not result in such relevant consequences, e.g., in applications such as facial recognition in photos taken by smart cameras [11]. However, in areas such as autonomous cars [12], financial transactions [13] and mainly medical applications [14], failures are unacceptable, considering that erroneous decisions can have disastrous con-sequences, such as the loss of human lives. Due to this fact, these application areas have extreme interest in explaining and interpreting each decision made by deep learning models.

✉ Vitor Bento, vitorbds@hotmail.com; Manoela Kohler, manoela@ele.puc-rio.br; Pedro Diaz, pedro.diaz@puc-rio.br; Leonardo Mendoza, leofome@hotmail.com; Marco Aurelio Pacheco, marco@ele.puc-rio.br | [1]Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil. [2]Department of Electrical Engineering State, University of Rio de Janeiro, Rio de Janeiro, Brazil.

Explaining Artificial Intelligence (XAI) [15] is the area of study that aims to explain, interpret, and visualize the decisions made by deep learning models. Many studies have been developed to understand how models make its decisions, so that in sensitive practical applications, the specialists have more confidence in the model's predictions. Medical research is an area that widely uses XAI techniques [16], aiming to understand how models learn to identify certain clinical problems and important features are taken into account to make each decision.

This study finds high evidence that XAI techniques can also be used to improve deep learning models performance. It is very common that datasets provided by companies and institutions consist of overlaid images—images with superimposed text and company logos—, as show in Fig. 1. It is observed, by XAI technique, namely layer-wise relevance propagation (LRP) [17], that this unwanted extra information can consistently reduce model's performance.

We reproduce—synthetically—the conditions of overlaid images in the dataset GC10-DET, a public dataset [18], adding random information and company logos superimposed to the original images. The processed synthetic dataset can be found at our repository presented in data availability section. Such dataset consists of ten classes of metal surface defects collected by an industry. A deep learning model was trained to classify these defects, obtaining a low accuracy which is considered our baseline. Then, the LRP technique was used to analyze the model's inferences. From acquired results, it can be observed that the model learned to solve the problem by identifying patterns in the text and logos superimposed on the image and not by the actual surface defect itself. Therefore, computer vision techniques were used to remove the superimposed text and logos from the images and the model was retrained, thereby, identifying the defect of interest. This new model achieved a F1 score 20% higher than the baseline.

The main contribution of this work is to show how XAI techniques can be used to improve performance of deep learning models. In addition, a problem of practical interest was solved using deep learning and XAI. Several works, such as [19, 20], also explore deep uncertainty learning [21] to improve deep models robustness and interpretability. The advantage of the proposed approach is to that it is straightforward and can be applied to any deep learning model.

This paper is organized as follows: Sect. 2 explains the theoretical concepts of the LRP technique and the computer vision techniques used. Section 3 presents the dataset and details of the experimental procedure. Section 4 evaluates the performance of the proposed workflow in a real case study. Finally, Sect. 5 summarizes the conclusions obtained in this work.

2 Background

LRP technique informs relevance of each pixel for the decision made by a deep neural network. Even though it is an oversimplified form of explanation compared to the human conception of explanation, this information is valuable to illustrate the behavior of deep learning models. This technique works by back-propagating the predicted output in the deep neural network using a set of rules.

Fig. 1 Example of images with overlay text. The red box highlights the region with overlay. **A** Surveillance camera. **B** Underwater monitoring

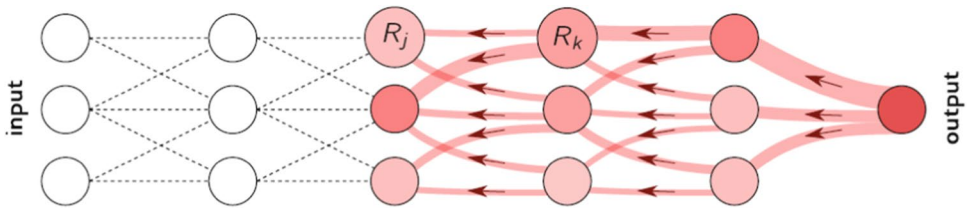

Fig. 2 Illustration of the LRP procedure for neuron k back-propagating the relevance score R_k, image taken from [9]

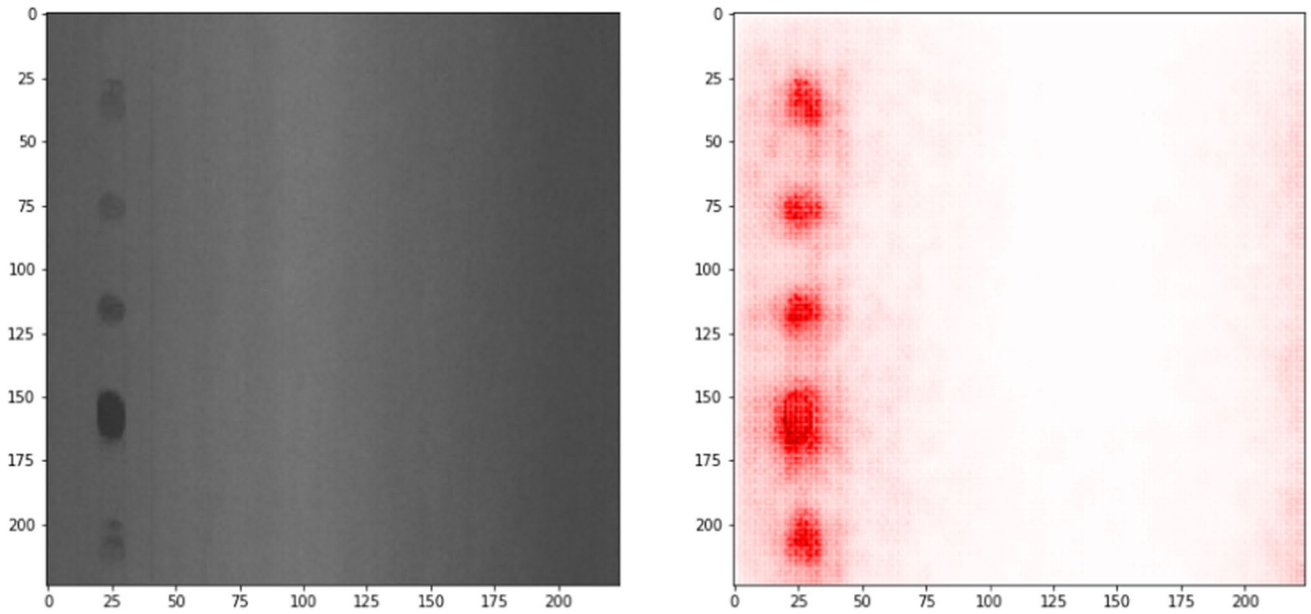

Fig. 3 LRP application example. Input image (left) with a metal surface defect, and the LRP output (right)

2.1 Layer-wise relevance propagation

This technique works with the conservation idea, resembling the Kirchoff's conservation laws in electric circuit theory [22]. Let *j* and *k* be the neurons at two consecutive layers of a deep neural network, where neurons in *k* are in a lower layer than neurons in *j* layer. Neurons in *j* will receive a relevance score from neurons in *k*. The relevance scores (*Rj*) at a given neuron in *j* is achieved by applying the following rule:

$$R_{j=} \sum_{1}^{k} \frac{z_{jk}}{\sum_{1}^{j} z_{jk}} R_k \tag{1}$$

where $z_{ik} = a_{jk} w_{jk}$, being a_{jk} the output values of the activation function of the neurons in *j*, and *w* the weight learned during training between the neuron *j* and *k*. The relevance in the input layer is between the neurons of the first layer and the pixels of the input image of the network, and the relevance in each pixel of the image is the final process. Figure 2 shows the processes of LRP back propagation.

Figure 3 illustrates an application example of the LRP technique in a deep learning model trained to detect metal surface defects. The left image is the input image, and the right image is the LRP output. Red pixels indicate high relevance in the process of image classification while the white region indicates low relevance. In such example, it is possible to observe that, based on its focus—the red pixels—, the network is in fact learning to identify the defect itself.

2.2 Computer vision techniques for inpainting

Four computer vision techniques to remove text and logos and withdraw the attention of the model from them were employed: (i) Gaussian Blur [23], (ii) image cropping, (iii) censor bars, and (iv) Generative Inpainting [24]. The first three

techniques were applied to the upper and lower regions of the images, details of their application are explained in experimental section.

For the fourth technique—Generative Inpainting—a generative model [25] was applied to the images. The generative model used was based in the Generative Adversarial Net-works (GAN) [26], such model estimates generative models via an adversarial process, in which it simultaneously trains two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than the generator G. GAN was introduced in [26] as a framework and the G and D are approximated by two neural networks.

These two networks compete in a min–max game, that under ideal conditions, converges and the generator learns the distribution of the given data. In another words, training a GAN is equivalent to minimizing Jensen–Shannon Divergence (JSD) [27]—other divergences could be possible too– between the generator and data distributions.

Conditional GANs (CGAN) [28], was adapted in this work for image inpainting [29], it was trained using the overlaid image as input to the generator network. The discriminator receives the preprocessed image as a real input image, i.e., computer vision techniques are applied to preprocess these images, and overlaid image as fake ones. As the discriminator associates real images with image without the superimposed text and logos, the generator is forced to learn to remove this information from the input image.

Formally, the generator is not simply maximizing the likelihood of single samples but minimizing the overall distance between real—image with no overlay—and the generated distribution—overlaid image—. Therefore, it is learning the real images' probability distribution, which is accomplished by minimizing the JSD under ideal conditions, as shown below:

$$min_G max_D V(D,G) = E_{x\sim p_{data}(x)}[logD(x|y)] - E_{z\sim p_z(z)}[\log(1 - D(G(z|y)))] \tag{2}$$

where G is the generator, D is the discriminator, x is a sample from a given dataset with probability density function p_{data}, y is the condition input of the model, z is the random noise from normal distribution p_z, and E is the expected value.

3 Experimental setup

We first train a deep learning model to classify 10 common metallic surface defects, the classes present on the dataset are: punching, welding line, crescent gap, water spot, oil spot, silk spot, inclusion, rolled pid, crease and waist folding. An example of each class present in the dataset is shown in Fig. 4.

3.1 Network architecture, optimizer and training details

MobileNet [30] with pre-trained weights from ImageNet [31] was empirically chosen for training, as it achieved better results with features extracted for this particular dataset. Adam optimizer with a momentum value equal to 0.9, initial learning rate equal to 0.001 and batch size of 16 was used. Thirty experiments were performed for each study case.

The dataset, containing 2306 samples, was split into train, validation and test sets with a distribution of 70%, 20% and 10%, respectively. Since this dataset is not balanced, F1 score, Precision and Recall metrics were used to evaluate our model's performance. The model was trained through 450 epochs with an early stopping patience of 60 epochs without improvement and a 0.001 tolerance over the F1 score in validation samples.

3.2 Computer vision techniques

LRP results show that text and logos superimposed to the images prevents models from learning relevant features. To solve this problem, computer vision techniques were used to eliminate such elements in the dataset. The techniques used were: (i) Gaussian Blur, (ii) Image Cropping, (iii) Censor Bars, and (iv) Generative Inpainting, as mentioned in the background chapter.

The first three techniques were applied to the upper and lower regions of the images, an area of 35×224 pixels at the top and at the bottom of the image, as show in Fig. 5. For the Gaussian Blur technique, the kernel size was of 17×17 pixels with an standard deviation of 20 in both directions, horizontal and vertical. In the Image Cropping technique, the same area was cropped and the resulting image was resized to its original size—224×224 pixels. The Censor Bars technique instead of cropping that area, it employs a black stripe to replace them and cover the text and logos. Finally, in the

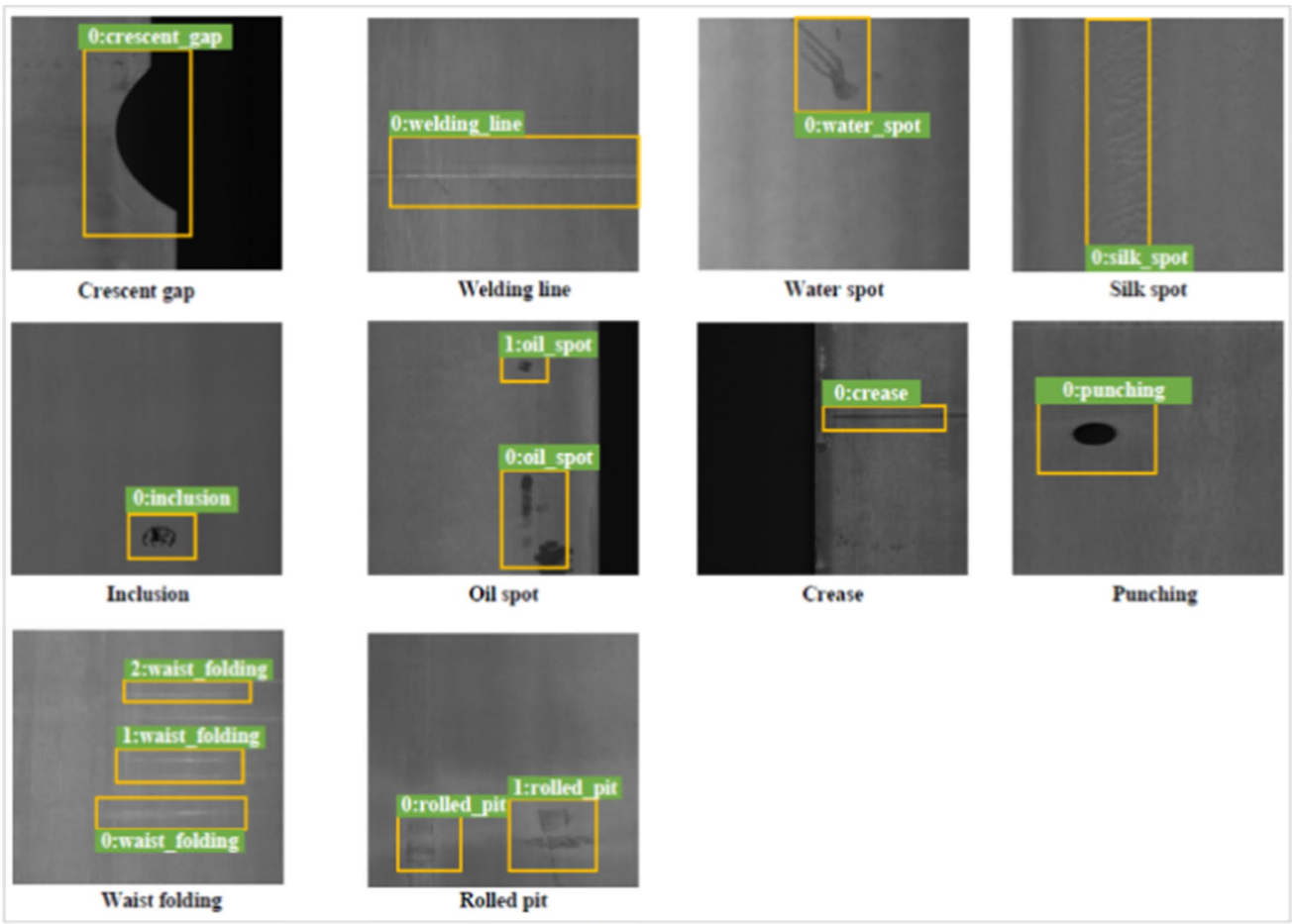

Fig. 4 Metallic Surface Defects. (Taken from original dataset, available at https://github.com/lvxiaoming2019/)

Generative Inpainting technique, the whole image was generated by the generative model, Fig. 5E shows an example of Generative Inpainting. The original image for all cases is illustrated in Fig. 5A. The complete dataset—with preprocessed images—and the original dataset are being made publicly available at our repository (data availability section).

4 Results

Table 1 shows the performance metrics for the first trained model—with the original dataset, with text and logos—, hereafter called Original Model.

As previously stated, to understand the low performance of the Original Model, a LRP technique was used. In Fig. 6, we show two examples belonging to the *Inclusion* class. It can be observed that the model focused on text and logos to make its decision. LRP results over all samples are publicly available at our repository (data availability section).

By prior knowledge, text and company logos do not have any useful information for class labelling. Furthermore, it is expected for the model to make its decision based on relevant features of the image and avoid these irrelevant patterns. In order to complete understand the learning model, we evaluate the *Original Model* (trained on images with text and logo) with the images preprocessed with the *Generative Inpainting* approach (tested on images without logos and text). In Fig. 7 we show the results of LRP over the same images shown in Fig. 6, but without text and logos. The result of the LRP technique shows that although the model used the *Inclusion* class features in this scenario, the model was also extremely noisy, giving high relevance to the borders of the image. In Table 2, performance metrics of the *Original Model* are presented: the poor performance shows that the model was focusing in texts and logos to make inferences. Thus, computer vision techniques, such as *blur*, *crop*, *censor bars* and *Generative Inpainting* are applied to prevent the model from using such information during the learning phase.

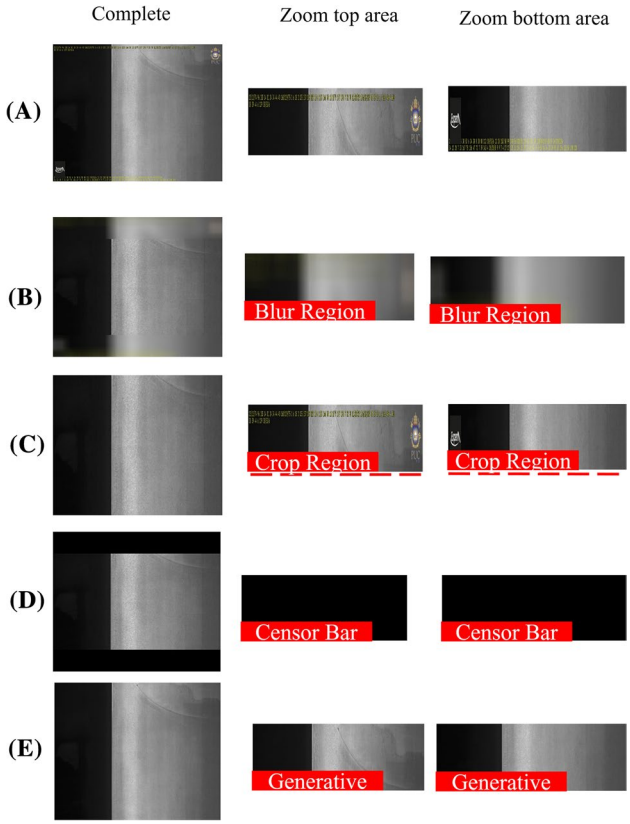

Fig. 5 **A** Original image with text and logos, **B** Gaussian Blur, **C** Image Cropping, **D** Censor Bars, b Generative Inpainting

Table 1 Test dataset performance metrics (Original model)

	F1-score	Recall	Precision
Original (test)	0.50 ± 0.02	0.51 ± 0.02	0.52 ± 0.02

4.1 Results after preprocessing images

Table 3 shows F1-score, Recall and Precision in the test samples for each approach compared to the *Original Model*. It is noteworthy that the *Crop*, *Blur* and the *Generative Inpainting* techniques increases the model's performance. This suggests that these techniques remove text and logos with no meaningful modification to the images.

It is clear, from the results presented, that when removing unnecessary information, such as text and logos, the model is able to significantly improve its performance. In addition, the model is also inferring the correct class based on relevant features from the images, as shown in Fig. 8. In Fig. 8, it is shown two examples of the *Oil spot* class, such defect is usually caused by contamination of mechanical lubricant, which will affect the appearance of the metal surface. Analyzing LRP output image, it is clear that it is exactly what the model is focusing on to make the correct prediction. So now, it is evident that the model is using the correct patterns in the image to make predictions, adhering to expectations that models use relevant features from images to label them. LRP results over all preprocessed samples can be accessed at our repository (data availability section).

Lastly, it is shown in Fig. 9 that during the training phase, all models have similar performance, indicating that a deep learning model can learn to solve the same problem based on different features from the samples; however, there is a possibility, as seen in the Fig. 6, that these learned features include noise data (from overlay) resulting in overfitting. This arise a poor performance in real applications as shown previously.

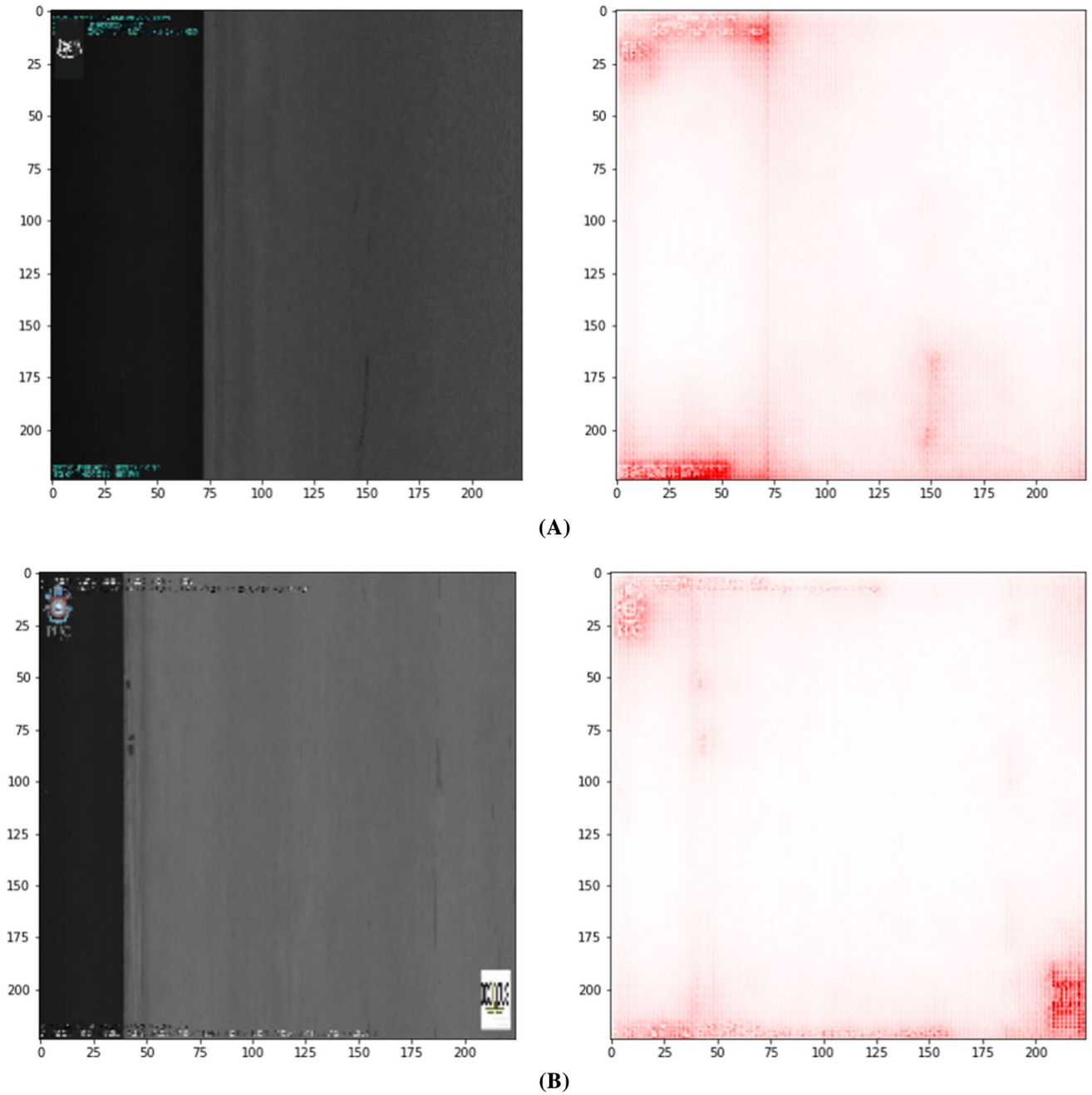

Fig. 6 Results of LRP over two images belonging to the *Inclusion* class (using the *Original Model*)

5 Conclusions

In this work we show that it is possible to use XAI techniques not only to understand the model's behavior, but also to improve its performance.

We observed that, by using XAI, the trained model was using information from the images' superimposed text and logos to infer data classes. By prior knowledge, text and company logos do not have any useful information for class labelling, furthermore, it is expected for the model to make its decision based on relevant features of the image. Thus, computer vision techniques, such as blur, crop, censor bars and generative inpainting are applied in order to prevent the model from using such information during the learning phase, obtaining the best results with

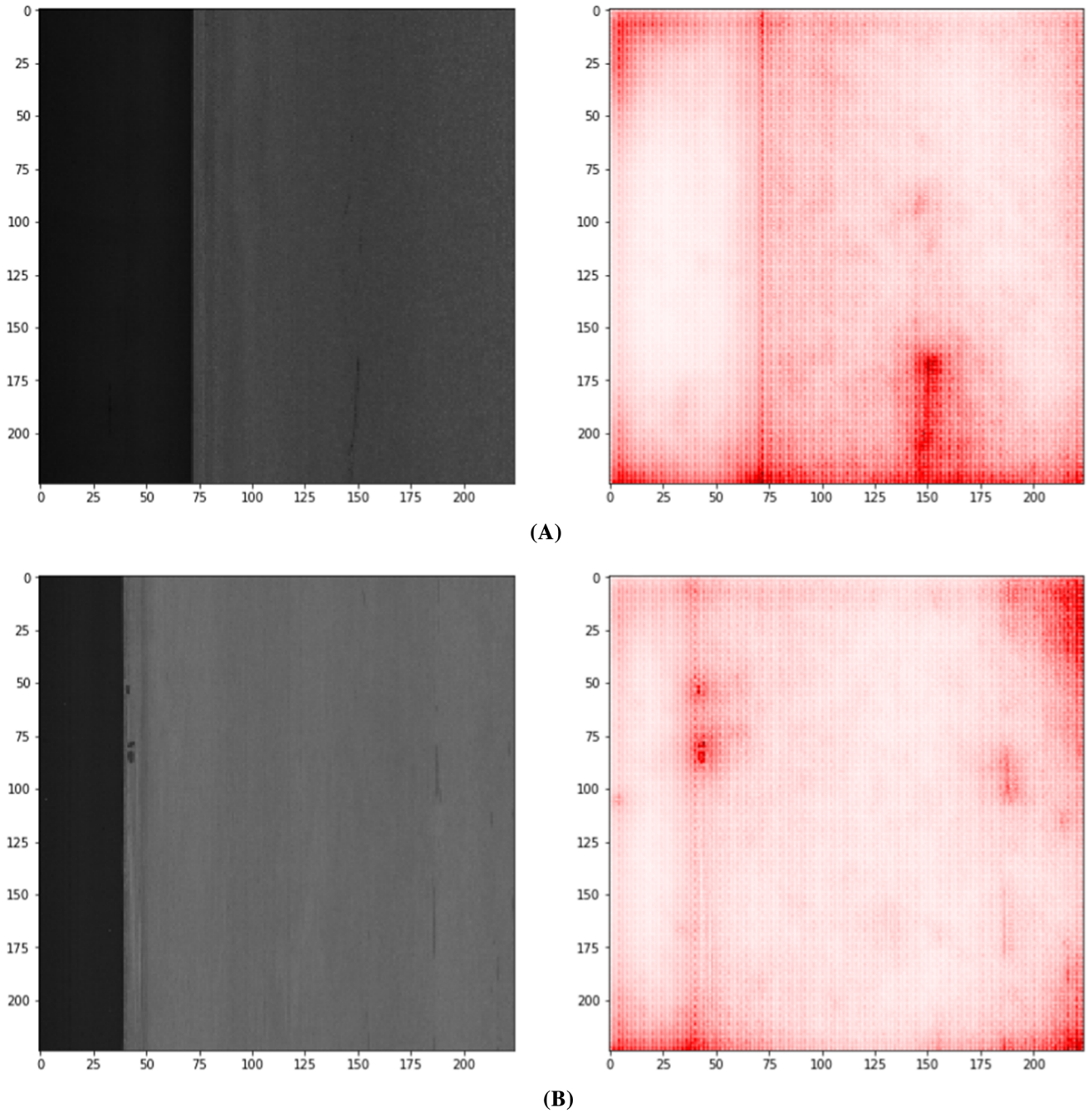

Fig. 7 Results of LRP over two images belonging to the *Inclusion* class (using the *Original Model* over the Generative samples)

Table 2 Test dataset performance metrics (*Original Model* with images preprocessed with the Generative approach)

	F1-score	Recall	Precision
Original model over Generative samples	0.33 ± 0.02	0.31 ± 0.02	0.64 ± 0.02

Table 3 Teste dataset performance metrics: F1-Score, Recall and Precision (results in bold highlight superior results when compared to the original approach)

	F1-score	Recall	Precision
Original	0.50 ± 0.02	0.51 ± 0.02	0.52 ± 0.02
Blur	**0.54 ± 0.03**	**0.52 ± 0.03**	**0.61 ± 0.05**
Black Stripe	0.47 ± 0.01	0.47 ± 0.01	0.48 ± 0.01
Crop	**0.51 ± 0.04**	0.50 ± 0.04	**0.56 ± 0.03**
Generative	**0.65 ± 0.02**	**0.64 ± 0.02**	**0.69 ± 0.03**

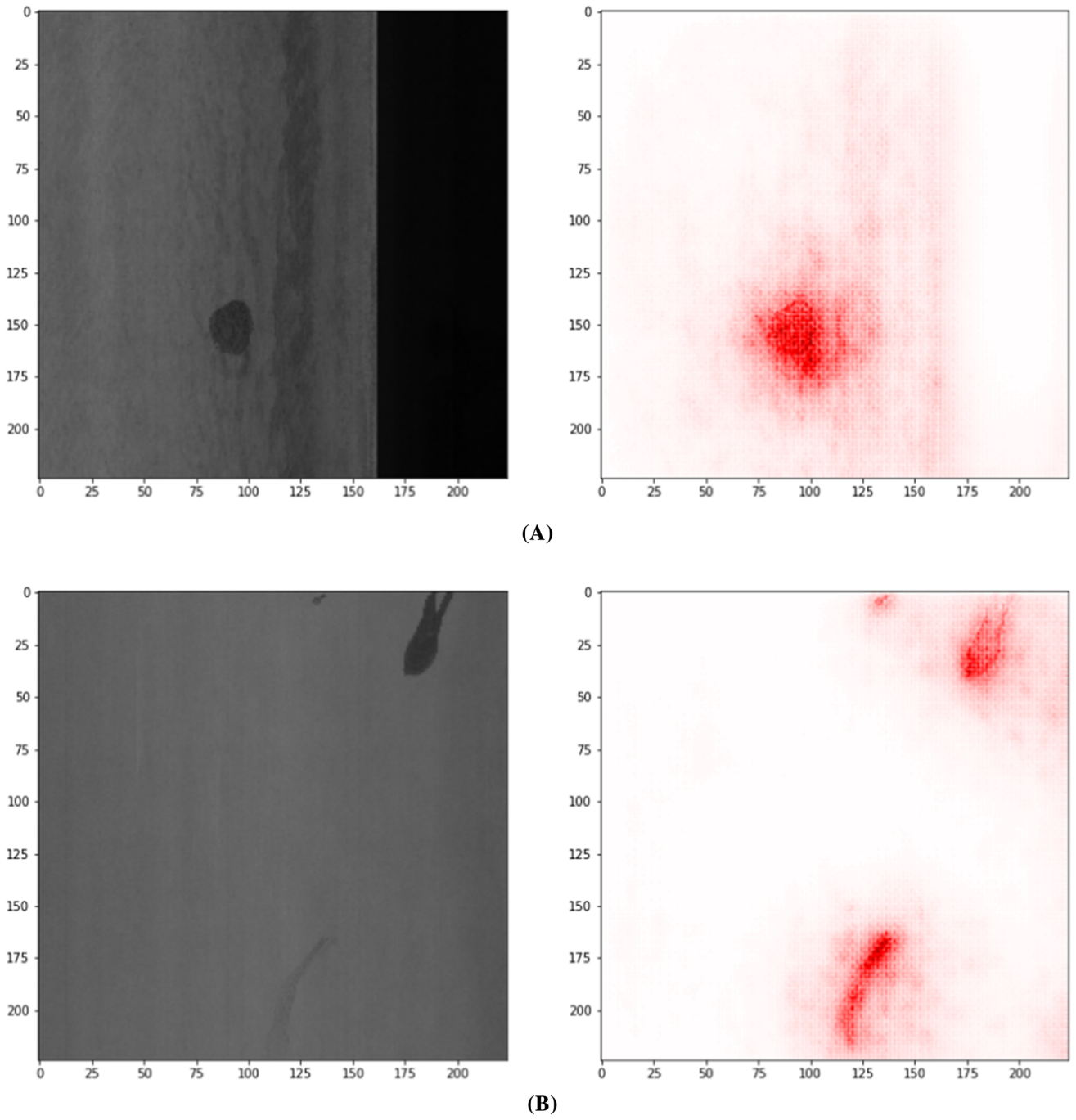

Fig. 8 Two examples of LRP results using images belonging to the *Oil Spot* class (preprocessed with the Generative Model)

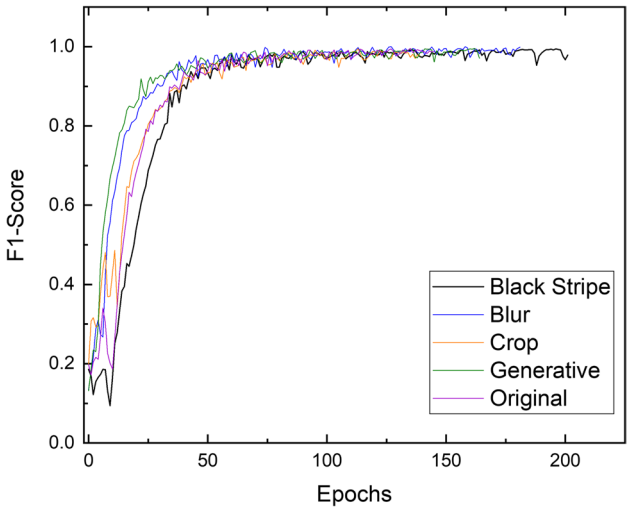

Fig. 9 Evolution of F1-socre during the training of all evaluated approaches

generative inpainting. Retrained models achieved better results than the original one, improving F1 score by 20% for the best preprocessing technique.

Acknowledgements The authors would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) for their financial support.

Authors' contributions Conceptualization VB and MK; Formal analysis: VB, MK and PD; Methodology MK, LM and PD; Software VB and PD; Supervision MK, LM and MAP; Writing VB, MK, PD, LM, MAP. All authors read and approved the final manuscript.

Funding Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio).

Data availability The datasets for the presented study case are available in the following repository: https://drive.google.com/drive/folders/1jnxBKwM7GKDrVObC0KC5NCXe1QjH69OK.

Code availability All source code can be found at https://github.com/ICA-PUC/Improving-Deep-Learning-Performance-By-Using-xAI.

Declarations

Competing interests There is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016. https://doi.org/10.1109/CVPR.2016.90.
2. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: 3rd international conference on learning representations ICLR 2015—conference tracking proceedings. 2015. p. 1–14. https://arxiv.org/abs/1409.1556.
3. Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. J Comput Chem. 2017;38:1291–307. https://doi.org/10.1002/jcc.24764.
4. Luckow A, Cook M, Ashcraft N, et al. Deep learning in the automotive industry: Applications and tools. In: Proceedings 2016 IEEE international conference big data, Big Data; 2016. p. 3759–3768. https://doi.org/10.1109/BigData.2016.7841045
5. Jin Y, Zhang J, Li M, et al (2017) Towards the automatic anime characters creation with generative adversarial networks. arXiv 92:1–16

6. Lindholm E, Nickolls J, Oberman S, Montrym J. Nvidia Tesla: a unified graphics and computing architecture to enable flexible, programmable graphics and high-performance computing. IEEE Micro. 2008;28:39–55.
7. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv: 1702.08608 [Preprint]. 2017. Available from: https://arxiv.org/abs/1702.08608
8. Holzinger A, Langs G, Denk H, et al. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9:1–13. https://doi.org/10.1002/widm.1312.
9. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv: 1708.08296 [Preprint]. 2017. Available from: https://arxiv.org/abs/1708.08296
10. Gonzalez TF. ImageNet classification with deep convolutional neural networks. Handb Approx Algorithms Metaheurifile. 2007. https://doi.org/10.1201/9781420010749.
11. Balaban S, Labs L. Deep learning and face recognition: the state of the art. arXiv; 2019. p. 1–9. https://doi.org/10.1117/12.2181526
12. Tian Y, Pei K, Jana S, Ray B. DeepTest: automated testing of deep-neural-network-driven autonomous cars. In: Proceedings of the 40th international conference on software engineering. 2018. p. 303–314. https://doi.org/10.1145/3180155.3180220.
13. Dornadula VN, Geetha S. Credit card fraud detection using machine learning algorithms. Procedia Comput Sci. 2019;165:631–41. https://doi.org/10.1016/j.procs.2020.01.057.
14. Wang D, Khosla A, Gargeya R, et al. Deep learning for identifying metastatic breast cancer. arXiv: 1606.05718 [Preprint]. 2016. Available from: https://arxiv.org/abs/1606.05718
15. Lecture Notes in Artificial Intelligence. Explainable AI: interpreting, explaining and visualizing deep learning. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, editors. Springer Nature. https://doi.org/10.1007/978-3-030-28954-6.
16. Lee S, Lim S, Lee T, et al. Cancer subtype classification and modeling by pathway attention and propagation. Bioinformatics. 2020;36:3818–24. https://doi.org/10.1093/bioinformatics/btaa203.
17. Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE. 2015;10:1–46. https://doi.org/10.1371/journal.pone.0130140.
18. Lv X, Duan F, Jiang JJ, et al. Deep metallic surface defect detection: The new benchmark and detection network. Sensors (Switzerland). 2020. https://doi.org/10.3390/s20061562.
19. Kraus F, Dietmayer K. Uncertainty estimation in one-stage object detection. 2019 IEEE Intell Transp Syst Conf ITSC. 2019. https://doi.org/10.1109/ITSC.2019.8917494.
20. Yu T, Li D, Yang Y, et al (2019) Robust person re-identification by modelling feature uncertainty. In: Proceedings of the IEEE international conference on computer vision 2019-October. p. 552–61. https://doi.org/10.1109/ICCV.2019.00064.
21. Chang J, Lan Z, Cheng C, Wei Y. Data uncertainty learning in face recognition. In: Proceedings of the IEEE computer society conference on computer vision pattern recognition; 2020. p. 5709–18. https://doi.org/10.1109/CVPR42600.2020.00575.
22. Smale S. On the mathematical foundations of electrical circuit theory. In: On the mathematical foundations of electrical circuit theory, vol 2. The Collected Papers of Stephen Smale; 2000. p. 951–68.
23. Gedraite ES, Hadad M. Investigation on the effect of a Gaussian Blur in image filtering and segmentation. Proceedings of the Elmar—international symposium electron Mar; 2011. p. 393–6.
24. Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention. In: Proceedings of the IEEE computer society conference computer vision Pattern Recognition; 2018. p. 5505–14. https://doi.org/10.1109/CVPR.2018.00577.
25. Salakhutdinov R. Learning deep generative models. Annu Rev Stat Its Appl. 2015;2:361–85. https://doi.org/10.1146/annurev-statistics-010814-020120.
26. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Commun ACM. 2020;63:139–44. https://doi.org/10.1145/3422622.
27. Fuglede B, Topsoe F. Jensen-Shannon divergence and Hilbert space embedding BT—IEEE international symposium on information theory; 2004. p. 31. https://doi.org/10.1109/ISIT.2004.1365067.
28. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv: 1411.1784 [Preprint]. 2014. Available from: https://arxiv.org/abs/1411.1784
29. Takahashi T. Image inpainting. Kyokai Joho Imeji Zasshi/Journal Inst Image Inf Telev Eng. 2017;71:503–4. https://doi.org/10.3169/itej.71.503.
30. Howard AG, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861 [Preprint]. 2017. Available from: https://arxiv.org/abs/1704.04861
31. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115:211–52. https://doi.org/10.1007/s11263-015-0816-y.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.