Survey of Perturbation Approaches for Explainable ML in the Context of Flood Detection from SAR Images

Anastasia Schlegel and Ronny Hänsch
Microwaves and Radar Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany

Abstract

Machine learning and especially deep convolutional networks (ConvNets) are increasingly being used for various image analysis tasks in Earth observation. Despite their strong performance, ConvNets are considered black boxes lacking explainability of their predictions. Methods under the umbrella term “explainable machine learning” or more “explainable AI” (XAI) aim to provide human-interpretable reasoning for why a model made a particular prediction. Amongst them, perturbation techniques explore changes in the prediction when the input is locally distorted. We investigate the influence of different parameter choices on the quality of explanations in the context of flood detection using SAR images.

1 Introduction

In recent years, the use of machine learning (ML) methods has grown considerably in the remote sensing (RS) and Earth observation (EO) community. Especially deep convolutional networks (ConvNets) are the established go-to models for numerous image analysis tasks [1]. Being trained with regard to high accuracy, deep ConvNets typically are composed of millions of interdependent parameters. Their sheer number and extreme connectivity yield highly complex functional relationships [2]. Although the overall model structure, its individual components, as well as the general learning algorithm are commonly known, the created model complexity makes it impossible to even grasp what the model has actually learned, whether it works as intended, or if its prediction is sensible. This lack of interpretability and explainability gives deep ConvNets the reputation of being black box models [3].

In broader terms, the demand for explainability in ML can be linked to trust, accountability, robustness, etc. of the model and its predictions. As opposed to creating inherently interpretable models, under the umbrella term of "explainable AI" (XAI) methods have been developed that aim to open black box models and understand their decisions [4]. XAI techniques that provide explanations to already trained models in a post-hoc manner can broadly be categorized into gradient-based [5, 6, 7, 8] and perturbation-based [9, 10, 11, 12] approaches. Both provide local explanations, i.e. explanations concerning the prediction of single instances. While the former approaches leverage gradients computed at individual instances to explain predictions [13], the latter ones monitor changes in the prediction associated with locally distorted input with respect to the original input [14].

Compared to gradient-based methods, perturbation-based techniques do not require an understanding of the models' inner workings, i.e. they are model-agnostic and can be applied to any black box ML model. Moreover, perturbation-based approaches have the advantage of being particularly simple in terms of their implementation as well as in the interpretation of their explanations. With perturbation-based approaches an attribution, i.e. the relevance, of an input part or feature is computed by occluding it and investigating the changes this causes in the prediction with respect to the original input [14]. Thus, perturbation-based approaches only require forward passes to obtain predictions for different inputs and provide relevance scores that are directly linked to the models' output variables. However, not only does this come with the cost of computational load as the number of features to test increases, but more importantly, the result is strongly influenced by the methods' parameters, such as the size and type of the occluding patches.

This paper presents a survey of perturbation-based XAI methods used to enhance understanding of ML model predictions in the context of flood detection from Synthetic Aperture Radar (SAR) images. Flood events cause more damage than other natural hazards [15]. Model predictions can assist in mitigating the flood risk, monitoring the flood progress, optimizing emergency vehicle routing, etc.. For such critical applications, explainability becomes of upmost importance ensuring accurate and reliable predictions [1]. Particularly, local and global explanation methods have been used to analyze results for surface water detection [16] and SHAP [11] to interpret CNN-generated flood susceptibility maps [17].

In our survey we focus on the analysis of the occlusion-based methods. For that purpose, we employ a rather standard segmentation network, the U-Net [18], which we train and evaluate on the Sen1Floods11 [15] dataset. We extend the application of perturbation-based approaches from image classification, whereby an image is assigned a single label, to semantic segmentation, and thus to a pixel-wise prediction. In our analysis we primarily investigate the influence the two parameters, window size and occlusion value, have on the explanation.

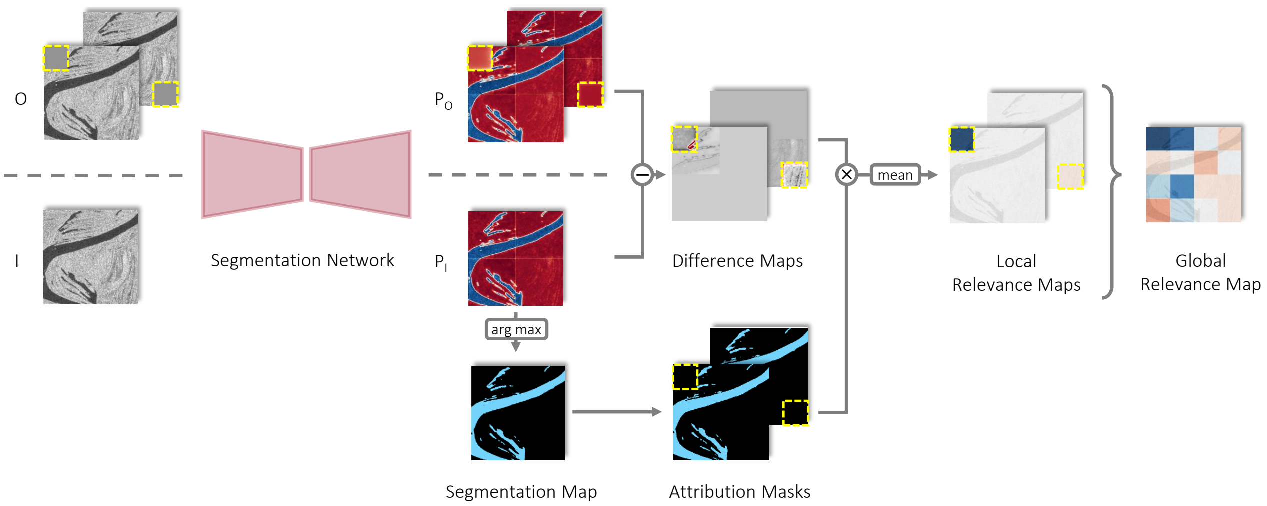

Figure 1 Schematic overview of the occlusion method to estimate the relative importance of single image patches for the prediction of water in SAR images. Within the probability maps P_I and P_O estimated from the segmentation network, blue areas represent high probability values for the presence of water, while low probability values are shown in red. In the computed difference maps, red is associated with a prediction drop, black with an improved prediction, and gray with no change in prediction. The individual attribution masks exclude water areas inside the occluded patch. Within the local and global relevance maps blue is attributed to positive and red to negative relevance scores.

2 Occlusion-based Approaches for Relevance Estimation

Occlusion [9] is a perturbation-based explanation method that monitors the change of a networks' output when instead of the original image an image with a locally distorted region is introduced. A probability decrease thereby indicates the relative importance of the occluded patch for the prediction. An increase in probability on the other hand is attributed to a disturbing influence of the occluded area on the prediction. This allows to investigate whether the prediction is based on the object itself or the surrounding context.

Figure 1 exemplary outlines how the relevance of individual image parts for the prediction of flooded areas in SAR images is determined using the occlusion method. First, given input images I and O are forwarded through the segmentation network. While I denotes the original image, O represents a sample from the set of occluded images created from I. By shifting the occlusion window over the original image, parts of the signal are blocked out causing local distortions. For both images I and O the segmentation network outputs the probability of each pixel belonging to the water class, here summarized in probability maps P_I and P_O, respectively. For each I and O pair, the relevance of the occluded image part R_O is then determined based on the difference between P_I and P_O

$$D_O = P_I - P_O. \tag{1}$$

To define the area to which the relevance of the occluded image parts is attributed, binary attribution masks are created from the segmentation output of the original image I. Averaging the difference values D_O over the water areas in the respective attribution mask yields a relevance score R_O of the locally distorted image part, which is visualized as a local relevance map. To avoid the bias on the relevance coming from large difference values in the occluded image parts, these regions are ignored in the attribution map. The global relevance map combines the individual relevance scores computed for each I and O image pair. Thereby blue image patches are associated with positive relevance scores and are considered important for the prediction of the water areas. Red image patches on the other hand represent negative relevance scores indicating a disturbing influence.

3 Experiments

3.1 Setup

Analyzing the occlusion method on a semantic segmentation task, we employ a standard U-Net which we train and evaluate on the Sen1Floods11 [15] dataset. The dataset contains 446 hand-labeled Sentinel-1 images in dual polarization (VV and VH) covering flood events across 11 countries. Since random cropping is used for data augmentation in the training stage, during inference the 512×512 images are split into four patches of size 256×256, which are later stitched back together. For the analysis of the occlusion method, this means that the influence of the distorted image part is limited to the prediction of water areas within the respective 256×256 patch and has no influence on the prediction in the remaining patches that belong to the original 512×512 image.

In the analysis, we consider two occlusion parameters, namely the size of the occluding patch as well as the occlusion type, which both have an influence on the resulting relevance maps and thus on the explanations the method provides.

	VV	VH
$\bar{x}$	0.837	1.013
$\bar{x}_{water}$	-1.639	-0.815
$\bar{x}_{non-water}$	1.439	1.457
min	-6.272	-4.244
max	3.521	2.799

Table 1 Image statistics used to investigate the influence of different occlusion types on the relevance score.

To investigate how the relevance score changes for smaller and larger occlusion windows, we systematically replace image parts with occlusion patches of size 128×128, 32×32, and 8×8. Within an occluded window, we test six different occlusion types. In five out of six cases, we set all pixels inside the occluded window to a single baseline value according to the image statistics presented in Table 1. In an additional case, we randomly shuffle the pixels within a window using ten different seeds and average the estimated relevance maps. This way the original values are preserved, the image statistics remain unchanged, and the image appears overall more organic. Yet, rearranging the pixels disrupts structures that are possibly relevant for the prediction.

3.2 Results and Discussion

Figure 2 presents a collection of relevance maps obtained from the experiments investigating the influence of the occlusion parameters, i.e. the window size and occlusion type, on the relevance estimation. The results indicate a strong dependence of the relevance score estimation on both parameters.

As can be expected, the relevance score of an image patch decreases with its size. Creating stronger distortions by blocking out larger parts of the original signal naturally has a greater impact on the prediction compared to small-scale modifications, as potentially relevant structures not only might be disrupted but entirely replaced. Note that an occlusion window of size 128×128 already covers a quarter of the entire image 256×256 patch. Despite that larger occlusion patches have a higher impact on the prediction, they yield rather coarse relevance maps. Obtaining a single relevance score, individual contributions within a large occlusion patch cannot be separated from each other and are averaged out. The influence of larger patches cannot be attributed to any particular feature or structure present in the image. Thus, such coarse relevance maps have little informative value and do not contribute significantly to the explanation of the prediction.

Smaller occlusion windows on the other hand yield fine-grained relevance maps with an overall appearance that mimics the underlying image structures. The higher resolution allows to attribute individual relevance scores to specific image features providing reasonable explanations of the prediction. In particular, the relevance map resulting from 8×8 occlusion patches with a replacement value representing the channel-wise mean of water pixels accentuates the disturbing influence of border regions between water and non-water areas. At the same time, the influence smaller image patches have on the prediction is particularly small, such that they might be neglected or that the network simply interpolates over them.

While the size of the occlusion window indicates a trade-off between the resolution of the relevance map and the influence of individual patches on the prediction, the occlusion value determines the estimated relevance scores and thus the overall appearance of the relevance map. The relative importance of water and non-water areas as well as the boundary region between them strongly depends on the selected occlusion value and how well it represents the original signal. We obtain similar relevance maps when replacing pixel values within an occlusion window by the overall mean and the mean representing non-water areas. As non-water areas dominate the overall image, this similarity is to be expected. Both the medium and high-resolution relevance maps indicate a slightly disturbing influence of non-water areas on the prediction, while water areas appear to be relevant. This can be attributed to the difference between the occlusion value and the average of water areas. When replacing water pixels inside an occlusion patch with a single value that is significantly higher than the water mean, closed water bodies can hardly be identified as those. Particularly small water structures are disrupted and simply lack the information needed to classify them as water deteriorating the prediction. Non-water areas on the other hand are represented well by both average values. The variance of pixel values inside an occluded patch over non-water areas is thereby reduced, yielding slightly better predictions - but with only little impact.

We observe a similar behavior when using the maximum of the respective channel as the occlusion value. The strong difference between the occlusion value and the average of water bodies completely deteriorates their prediction. However, the maximum channel value does not seem to be an appropriate representation of non-water areas as well. Whether the occlusion patch covers water or non-water areas, it appears like an outlier to the network, which decreases the prediction of the water class. While the network seems to be able to distinguish between water and non-water pixels, it struggles to identify the occluded patch as one of the classes.

Occluding non-water patches by the minimum or the average of the water value as well results in a probability decrease for class water. Again, the occlusion patch does not represent non-water areas well enough and rather interrupts their structures - which even for small occlusion windows seemingly are important for the prediction. However, occluding water areas barely impacts the prediction. When using the channel minimum, this is quite surprising, given that the occlusion value is significantly lower than the average water value. It appears like it is more important for the network to not have outliers in non-water patches rather than in water patches and thus to clearly distinguish between the two classes when forming the prediction. This becomes particularly visible in the high-resolution relevance map estimated from occlusion values that represent the water average, which highlights the disturbing influence of border regions between water and no-water areas. Replacing border patches with patches representing water

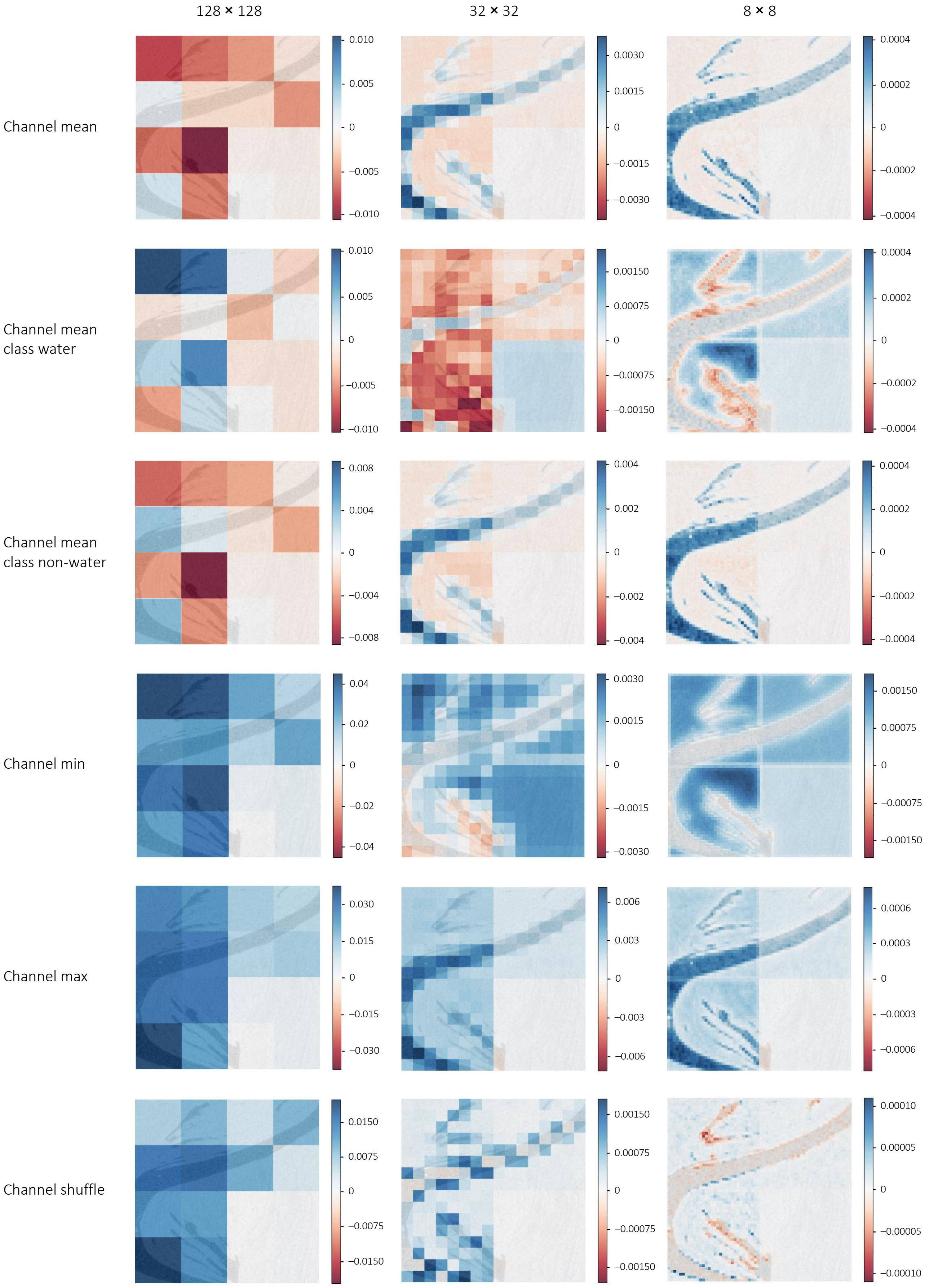

Figure 2 Experimental results showing the influence of two occlusion parameters, i.e. the size of the occlusion window (columns) and the occlusion type (rows), on the relevance estimation. Blue patches (positive relevance scores) are important for the prediction of water areas in the underlying sample image. Red patches (negative relevance scores) have a disturbing influence on the prediction.

areas in a way enlarges the water bodies and thus increases the probability of a pixel belonging to the water class. Even for small occlusion windows of size 8×8, there seems to be an impact on the way the network forms its prediction. We do not observe this effect when using the average of non-water pixels as the occlusion baseline. We assume that the border regions contain both water and non-water pixels and are therefore well represented by the non-water mean. Thus, we believe the relative importance of border regions is attributed to the strong cut between the average water and non-water values. This in turn again indicates that the occlusion value affects the network's ability to discriminate water from non-water areas. For the estimation of relevance scores, this appears to be more important than preserving the structure inside closed shapes.

The relevance maps estimated by randomly shuffling the pixels inside an occlusion window show the strongest differences from the relevance maps obtained in the previous cases. Its overall appearance is rather noisy, which can be attributed to the difference between neighboring occlusion windows. Although the occlusion patch has a rather organic structure, it becomes more difficult to interpret the relevance map. It indicates that shuffling pixels inside closed-water areas slightly improves the prediction while shuffling pixels in non-water areas yields a slight decrease. However, both effects are rather small compared to the more pronounced relevance scores for border regions. Shuffling pixels in these patches distorts the separation line between water and non-water areas. Water pixels potentially appear in non-water areas and vice versa. The randomness between occlusion patches thereby explains why some border regions appear relevant and others disturbing. Once more, this confirms that a clear distinction between water and non-water areas is more important than preserving the structure within closed areas.

Overall the obtained results indicate that the relevance score of an image patch heavily depends on the occlusion value and on how well it represents the occluded patch. Based on that, either closed water, non-water, or border areas determine the prediction and are identified to have a favoring or disturbing influence on it.

4 Concluding Remarks

Explainable machine learning aims to provide human-interpretable explanations of predictions made by machine learning models. We are particularly interested in methods that indicate which parts of an input image have a positive or negative influence on the estimation. To this aim, we test several variations of a perturbation approach which is widely used for its simplicity and direct meaning of the obtained relevance maps. We extend this approach from image classification to the prediction of semantic maps and evaluate its usefulness in the context of flood prediction from SAR images.

We demonstrate that using the occlusion method to understand whether an image patch is relevant for the prediction of water bodies or not strongly depends on the size of the occluded area as well as on the value that is used to cover that area. On a more general term, the qualitative results clearly show that the obtained relevance scores highly depend on the selected occlusion parameters and are often not consistent when these are varied. Thus, different occlusion parameters yield different explanations for the same prediction. The lack of consistency makes it hard to trust the explanation method and turn also the prediction. Moreover, while the occlusion method indicates whether certain areas are relevant for the prediction, it remains unclear if the prediction changes because the original values are missing or because new information is inserted.

While a quantitative evaluation of relevance maps is difficult, there are a few measures available that at least provide a quality estimate in certain aspects. Future work will include such performance metrics in the evaluation of the perturbation-based approach.

5 Literature

[1] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explain it to me – facing remote sensing challenges in the bio- and geosciences with explainable machine learning," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-3-2020, pp. 817–824, 2020. [Online]. Available: https://isprs-annals.copernicus.org/articles/V-3-2020/817/2020/

[2] W. Samek and K.-R. Müller, *Towards Explainable Artificial Intelligence*. Cham: Springer International Publishing, 2019, pp. 5–22.

[3] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42 200–42 216, 2020.

[4] W. Samek, T. Wiegand, and K. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *CoRR*, vol. abs/1708.08296, 2017. [Online]. Available: http://arxiv.org/abs/1708.08296

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

[6] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 3319–3328. [Online]. Available: https://proceedings.mlr.press/v70/sundararajan17a.html

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

[8] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th*

International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 3145–3153. [Online]. Available: https://proceedings.mlr.press/v70/shrikumar17a.html

[9] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[10] M. T. Ribeiro, S. Singh, and 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*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[12] V. Petsiuk, A. Das, and K. Saenko, "RISE: randomized input sampling for explanation of black-box models," *CoRR*, vol. abs/1806.07421, 2018. [Online]. Available: http://arxiv.org/abs/1806.07421

[13] M. Ancona, E. Ceolini, A. C. Öztireli, and M. H. Gross, "A unified view of gradient-based attribution methods for deep neural networks," *CoRR*, vol. abs/1711.06104, 2017.

[14] M. Ivanovs, R. Kadikis, and K. Ozols, "Perturbation-based methods for explaining deep neural networks: A survey," *Pattern Recognition Letters*, vol. 150, pp. 228–234, 2021.

[15] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg, "Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[16] L. Chen, X. Cai, J. Xing, Z. Li, W. Zhu, Z. Yuan, and Z. Fang, "Towards transparent deep learning for surface water detection from sar imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103287, 2023.

[17] B. Pradhan, S. Lee, A. Dikshit, and H. Kim, "Spatial flood susceptibility mapping using an explainable artificial intelligence (xai) model," *Geoscience Frontiers*, vol. 14, no. 6, p. 101625, 2023.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.