Reliability Scores From Saliency Map Clusters for Improved Image-Based Harvest-Readiness Prediction in Cauliflower

Jana Kierdorf and Ribana Roscher, *Member, IEEE*

***Abstract*— Cauliflower is a hand-harvested crop that must fulfill high-quality standards in sales making the timing of harvest important. However, accurately determining harvest-readiness can be challenging due to the cauliflower head being covered by its canopy. While deep learning enables automated harvest-readiness estimation, errors can occur due to field-variability and limited training data. In this letter, we analyze the reliability of a harvest-readiness classifier with interpretable machine learning. By identifying clusters of saliency maps, we derive reliability scores for each classification result using knowledge about the domain and the image properties. For unseen data, the reliability can be used to 1) inform farmers to improve their decision-making and 2) increase the model prediction accuracy. Using RGB images of single cauliflower plants at different developmental stages from the GrowliFlower dataset, we investigate various saliency mapping approaches and find that they result in different quality of reliability scores. With the most suitable interpretation tool, we adjust the classification result and achieve a 15.72% improvement of the overall accuracy to 88.14% and a 15.44% improvement of the average class accuracy to 88.52% for the GrowliFlower dataset.**

***Index Terms*— Harvest prediction, interpretability, reliability, saliency mapping, spectral clustering (SC).**

I. Introduction

ACCURATE harvest time forecasts are crucial for crop quantity and profitability in agriculture. For cauliflower, high-quality requirements for sale further complicate this process. To meet these standards, harvesting must be precisely timed within a short window. Since cauliflower growth is highly affected by climate, fields planted at different times may be ready for harvest simultaneously, and plants may develop differently within a single field. Therefore, it is a common agricultural practice that workers harvest plants individually by hand at different times. As the cauliflower head is covered by its canopy, the workers touch the head inside the plant and estimate the size, making harvesting highly time-consuming.

In digital agriculture, field monitoring is supported by satellite or UAV imagery [2] to observe plant development throughout the entire growth period. Machine learning methods increasingly form the basis for analyzing the acquired data, for example, to classify crop ripeness on a large scale [8] or to provide detailed predictions about harvest ripeness, the amount of harvest, or the date of harvest-readiness [6]. Predicting crop traits related to harvest is of economic benefit to farmers, so the model must be reliable, and the farmer should be able to have confidence in the model's decision.

We address the task of harvest-readiness estimation of single cauliflower plants and aim to derive a reliability score for the model's output that can be used to support the farmer in their decision-making process. To reach our goal, we use saliency mapping to identify image regions that have distinctive characteristics important for the model decision [1], [11]. We extend the clustering approach of saliency maps by [7] and combine the maps with knowledge about our application domain and the image properties to derive reliability scores of the model's output. Similar to our approach, some previous works also aim to improve the model through the integration of interpretations and explanations [5], [12], [14]. However, these works present ad hoc frameworks where the interaction between model and explanation is either learned during training or integrated through human interactions via retraining. Our work differs in that we propose a framework for deriving a reliability score for classification predictions that operates post-hoc during inference time without human interaction. Thus, the system can be applied to already trained models without changing the model architecture and without the need for re-training.

The main contributions of this letter are as follows.

1) A versatile post-hoc approach to derive intuitive reliability scores without time-consuming human interaction;
2) A use case where the reliability scores are used to improve harvest-readiness predictions on the GrowliFlower dataset by 15.73% to an overall accuracy of 88.14% and by 15.44% to an average class accuracy of 88.52%.

II. Materials and Methods

A. Framework

We solve the task of estimating the harvest-readiness of single cauliflower plants with deep learning-based image

Manuscript received 1 March 2023; revised 26 May 2023; accepted 27 June 2023. Date of publication 10 July 2023; date of current version 27 July 2023. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant RO 4839/7-1 | STO 1087/2-1, in part by the European Agriculture Fund for Rural Development with contributions from North-Rhine Westphalia under Grant 17-02.12.01—10/16—EP-0004617925-19-001, and in part by the DFG under Germany's Excellence Strategy, under Grant EXC 2070—390732324. *(Corresponding author: Jana Kierdorf.)*

Jana Kierdorf is with the Remote Sensing Group, Institute of Geodesy and Geoinformation, University of Bonn, 53113 Bonn, Germany (e-mail: jkierdorf@uni-bonn.de).

Ribana Roscher is with the Data Science for Crop Systems, Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany, and also with the Remote Sensing Group, Institute of Geodesy and Geoinformation, University of Bonn, 53113 Bonn, Germany.

Digital Object Identifier 10.1109/LGRS.2023.3293802

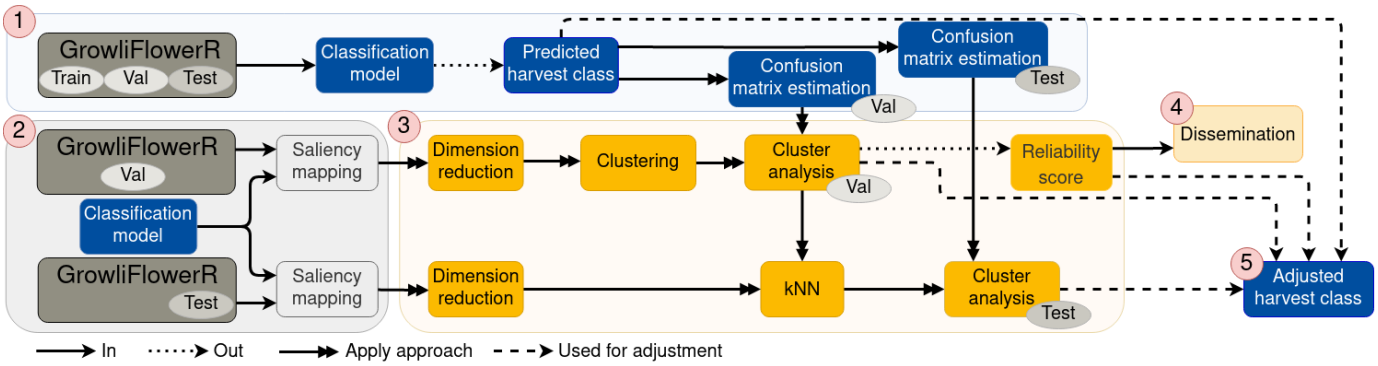


Fig. 1. Our framework. The different numbers represent (1) classification step, (2) saliency mapping step, and (3) clustering step of saliency maps with the assignment of reliability to the clusters by relating the confidence scores of the model to the corresponding saliency maps. (4) Dissemination to the farmer how reliable the model is while (5) adjustment step, where the predictions of (1) are improved using the reliability score of (3).

classification and combine it with an estimation of the reliability of the classification through clustering of saliency maps. Fig. 1 shows an overview of the five-step framework.

1) *Classification:* In the first step, images are classified into the classes `Ready` and `Not ready` for harvest within three days. We use a ResNet18 network [3]; however, the framework is flexible regarding the classifier.
2) *Saliency Mapping:* In the second step, we compute saliency maps for validation and test data post-hoc using the learned classifier. We consider gradient-weighted class activation mapping (Grad-CAM) [13], occlusion sensitivity mapping (OSM) [15], and local interpretable model-agnostic explanations (LIME) [10].
3) *Clustering:* We employ spectral clustering (SC) to identify groups of saliency maps computed on the validation data and derive reliability scores. The mean saliency map per cluster, denoted as prototype, is further analyzed. Test data can be assigned a reliability score by assigning its saliency map to the nearest cluster.
4) *Dissemination:* The reliability score is intuitively usable due to its value range between 0 and 1 and is communicated to the user together with the classification result.
5) *Adjustment:* In our use case, the classification results are adjusted based on the cluster assignments of the saliency maps to determine the final predicted classes. The decision depends on the summed percentage of false positives (FPs) and false negatives (FNs) per cluster. The evaluation of the classification step provides the assignments to FP and FN.

The framework does not require human interaction and can be applied to different models. However, human interaction is possible to further improve the classification results and reliability measures by analyzing and evaluating the human-understandable clusters of saliency maps.

B. Data

We use the images of the GrowliFlowerR dataset [4] of Field 2 from the dates 2021-08-23, 2021-08-25, 2021-08-30, and 2021-09-03 with given information about the harvest-readiness within the next three days. Three days is a compromise between harvest-readiness prediction accuracy and practicability of data acquisition. We divide the data into the classes `Not ready` and `Ready`. The plants representing both classes show a high similarity within the same day of acquisition but also between different days. The size of its head determines the ripeness; however, in most images, the canopy covers the head. The plant's stem is centered within the image, but depending on the plant's growth, the center of the cauliflower head can vary up to 20 cm from the stem.

We use the training, validation, and test set as described in [4]. If the plant shown in an image is already harvested, we exclude the image from the dataset. This results in a preliminary training set of 541 images, a validation set of 196 images, and a test set of 194 images. We apply standard augmentations like flipping and rotation on the training data. For images of class `Not ready`, we apply augmentations 50% more often than for images of class `Ready` to get a more balanced data distribution. After data augmentation, the training set contains 6224 images, 2432 of class `Not ready`, and 3792 images of class `Ready`.

For each image, we compute corresponding saliency maps. The datasets result in pairs of image and map. Thus, all target information of the images is also valid for the corresponding saliency maps.

C. Classification

We use a ResNet18 [3] architecture with cross-entropy loss, softmax activation, and two classes as output. We compute the model over 25 epochs and use an Adam optimizer with a weight decay of 0.0001. The learning rate starts at 0.0001 and is reduced with a learning rate scheduler with a step size of 5 and factor γ of 0.1.

D. Saliency Maps

Saliency maps aim to explain a model's decision by identifying important regions in the image. In our case, saliency maps highlight which image regions are important for predicting the classes `Ready` and `Not ready`, allowing conclusions about the reliability using the prior knowledge that the center of the image is important for the decision and the background should not play a role in the harvest-readiness estimation. We consider three well-established local approaches, as baseline approaches for saliency mapping,

namely a gradient-based approach, Grad-CAM, and two permutation-based approaches, OSM and LIME, where LIME differs in that it uses surrogate modeling, as our focus is not on the used methods.

Grad-CAM is a gradient-based model-specific method developed by Selveraju et al. [13] that uses gradient information to determine from which image regions the convolutional layer takes the information for prediction. The resulting map depends on the employed layer, where we follow the suggestions of Selveraju et al. [13] to use the last convolutional layer as it highlights object-level regions in the image, which are also easier to interpret. Grad-CAM provides information about the class of interest but no information about other classes.

The second approach, OSM, is a perturbation-based model-agnostic method developed by Zeiler and Fergus [15]. This method evaluates sensitivity toward occlusion. It uses a sliding window approach with patchsize p and stride s to permute the input by masking patches and, thus, determine the influence of the occlusion on the predicted model score. A blue pixel in the map indicates that the score after occlusion is lower than the original score, i.e., this pixel indicates the presence of the examined class. A red pixel indicates that the score after occlusion is higher than the original score, indicating a different class. Note that the smaller s, the finer the map's resolution. In our experiments, we chose $s = 2$ and $p = 11$.

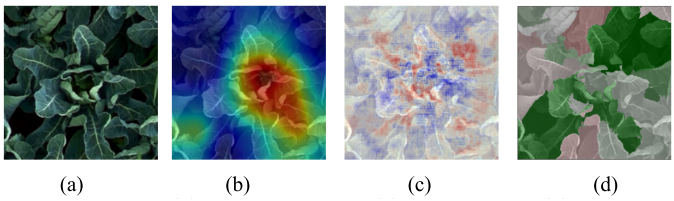
Like OSM, the third approach, LIME, is a perturbation-based model-agnostic method developed by Ribeiro et al. [10]. LIME perturbs the input and computes the prediction for these perturbed samples with the original model. Perturbation is applied by changing components in images that are meaningful to humans, such as superpixels. After perturbation, a local surrogate model is learned using the perturbed samples. In our work, we use a least squares linear regression model.

E. Spectral Clustering

We follow the idea of Lapuschkin et al. [7] using SC introduced by Ng et al. [9] to cluster the resulting saliency maps, which provides a better understanding of the model decision by taking into account image features other than RGB. SC involves clustering data based on a similarity measure derived from a new representation of the data. As similarity, we chose Gaussian similarity function with a kernel scale of 0.2 based on the Euclidean distance. Before we apply SC on our saliency maps, we perform principal component analysis on the vectorized data to reduce the dimensions of the data from 65 536 to 50. We decided on a dimension of 50 to obtain 95% of the variance because there is no unique eigenvalue difference, i.e., successive eigenvalues have no significant difference. We apply SC to the validation set and assign the closest cluster IDs to test data using kNN with $k = 5$. For our approach, we set the number of clusters $q = 8$ to be representative and generalizable to other datasets.

F. Evaluation Metrics

To evaluate the adjustment step, the summed percentage of FP and FN is considered in the calculated clusters q. We define



(a) (b) (c) (d)

Fig. 2. Resulting saliency maps or the used approaches (b) Grad-CAM, (c) OSM, and (d) LIME for a (a) RGB input image which is visualized in the maps' background.

$r_q = 1 - (\mathrm{FP}_q + \mathrm{FN}_q)$ as reliability score. The higher the reliability score the more reliable a prediction is in a specific cluster. If r_q falls below a threshold t in cluster q of the validation set, we swap the predicted class for all samples in cluster q and update the confusion matrix. We choose $t = 25\%$ based on a significant improvement in the validation set's accuracy. Threshold t is variable and selectable based on experience. Based on the updated confusion matrix, we adjust the overall and average class accuracy. We store the identified clusters for swapping and apply the same to the test data, followed by updating the test confusion matrix and accuracies.

III. Results and Discussion

We run our experiments on an AMD EPYC 7742 64-Core processor and an NVIDIA A100 for PCIe graphic card with 40 GB hBM2 RAM.

A. General Discussion

Our experiments find that clusters and harvest-readiness classes do not correlate. This is expected in the case of binary decision-making, where both classes may end up in the same cluster since they ideally use the same features. Instead, we focus on whether data within a cluster are correctly classified or not, which allows conclusions to be drawn about the reliability of the result. We use the confusion matrix for analysis. To assist the farmer in making harvesting decisions, we exploit the fact that the saliency maps of plant images end up in clusters whose classification result is primarily on the main diagonal of the confusion matrix (TP or TN) and maps that are associated with incorrect classification results (FP or FN) tend to end up in separate clusters.

B. Classification of Harvest-Readiness

On the validation set, we achieve an overall accuracy of 76.32% and an average class accuracy of 77.21%. For inference, we achieve an overall accuracy in classification of 72.41% and an average class accuracy of 73.08%. That means we are able to predict the harvest-readiness of approx. 3 out of 4 plants correctly.

C. Local Analysis: Saliency Maps of Single Sample Inputs

In some of the resulting Grad-CAM maps, a hotspot near the center is highlighted in the image as shown in Fig. 2(b). In other maps, the highlighted regions are located near the edges or scattered in the image. It is easy to analyze which

regions have an influence on the model's decisions since compact regions are highlighted.

A considerable amount of the OSM results resemble noise regardless of stride and patchsize for occlusion. Only a minor portion of the results show larger connected regions that are important for decision, as shown in Fig. 2(c). These are located in the area of the image that shows, among other things, the hidden cauliflower head or highlighted leaf regions. Many maps show several smaller highlighted regions which are difficult to explain because they do not indicate a unique plant trait. The ability of a simple explanation of the results varies more than for Grad-CAM.

In LIME maps, we see that the computed superpixels are not able to summarize pixels to semantically meaningful regions. This could be caused by the structure or the strong overlap of neighboring plants. Due to this, LIME saliency maps are difficult to connect to general statements about the reliability of classification outputs. An example of a sample analyzed by LIME is shown in Fig. 2(d). We consider LIME not suitable for our application.

Based on the assessment of single saliency maps, we consider Grad-CAM and OSM to be the most suitable approaches in our framework.

D. Global Analysis: Clustering of Saliency Maps and Reliability Derivation

Fig. 3 show the absolute number of Grad-CAM map assignments of the clustering results for eight clusters. A distinction is made between the validation and test set. The confusion matrix entries are differentiated by color. Our experiments have shown that eight clusters produce a good separability between false and correct predictions. Furthermore, depending on the amount of data, there are enough data points per cluster to make a reliable statement. Based on the distribution of validation data in Fig. 3(a), it becomes evident that cluster 5 contains about 95% false predictions, which are equally divided between FP and FN. This means that over 70% of all FN and FP belong to cluster 5. The cluster with the second highest proportion of false predictions is cluster 6. It is worth mentioning that the percentage is only 30%, which corresponds to only six images. The other clusters contain less than 20% false predictions. The clustering analysis allows saying with high confidence that samples assigned to cluster 5 are equivalent to a false prediction and should be adjusted. The reliability of the classification results of the saliency maps assigned to this cluster is, therefore, low and should be disseminated to the farmer. This is underlined in particular by the cluster assignments of the test data [Fig. 3(b)]. We observe that 80% of the false predicted test data are assigned to cluster 5. The proportion of false predictions in the other clusters is comparable to those within the validation data.

The prototypes of Grad-CAM are shown in Fig. 4. Half of the prototypes (2, 3, 7, 8) highlight the region in the center of the image. This is the location in the RGB input images of cauliflower heads covered by leaves, which are the indicators of cauliflower harvest-readiness. Even though the cauliflower

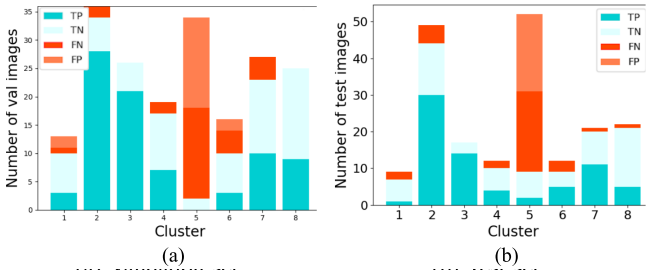


Fig. 3. Clustering of Grad-CAM results. Absolute number of (a) validation (val) images and (b) test images per cluster.

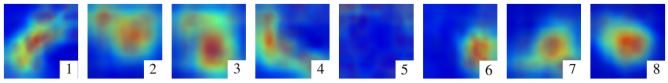


Fig. 4. Grad-CAM prototypes computed by mean saliency map per cluster (1)–(8).

head is not directly visible in the images, the model identifies the center of the plant as an essential feature for the classifier to determine the harvest-readiness. The interpretation of the classification results is straightforward and understandable for these clusters. The previously noticed cluster 5 also varies in this representation to the other clusters. In the image data assigned to the cluster, the classification model finds no distinctive features for determining the harvest-readiness. The visualization of the prototypes thus supports the model's reliability in addition to the cluster assignment since the visual representation is easier for the user to understand and interpret.

The clustering of the OSM maps shows a uniform distribution of false predictions in all clusters [Fig. 5(a)]. The percentage ranges from 10% to 30%. Based on the OSM cluster results, no statement can be made about the reliability of the results. The probability that a false prediction occurs in one of the clusters is similar for all clusters. The cluster assignment of the test data shows a similar distribution [Fig. 5(b)]. Only cluster 7 stands out. It should be noted that the assignment to this cluster corresponds to a single image only.

The prototypes also suggest no clear trend in terms of what the model uses as an informative feature in the RGB images (Fig. 6). Clusters 1 and 5 show a hotspot near the center, which, just like Grad-CAM, suggests that the model is paying partial attention to the canopy covering the head. Clusters 4, 6, and 8 give a hint of this. Comparing the prototypes of the OSM approach with those of the Grad-CAM approach, we see that for our scenario, the Grad-CAM approach results in more interpretable maps than the ones of OSM. Since no clear differentiation between false and correct prediction can be made in the data for OSM, the adjustment step introduced in this work is only applied to the Grad-CAM results. Adjusting the classification results based on the clustering results would worsen rather than improve the model results.

In summary, the combination of saliency map analysis and clustering provides information about the reliability of classification results. Nevertheless, some thought should be given to the saliency mapping approach to be used.

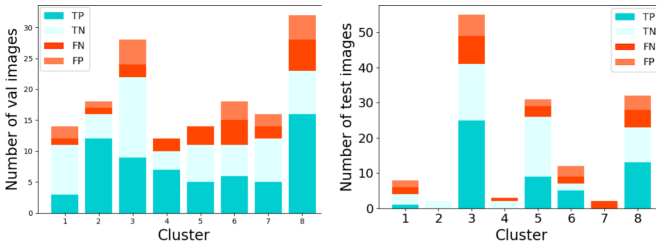


Fig. 5. Clustering of OSM results. Absolute number of (a) validation (val) images and (b) test images per cluster.

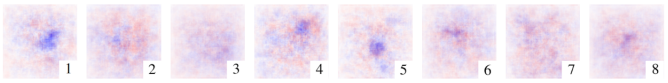


Fig. 6. OSM prototypes computed by mean saliency map per cluster (1)–(8).

E. Adjustment of Model Predictions

With regard to applying the adjustment step to Grad-CAM maps as explained in Section II-F, we achieve a 13.99% improvement in overall accuracy to 90.31% and a 13.39% improvement in average class accuracy to 90.60% for classification on the validation set. For inference, we achieve a 15.73% improvement in overall accuracy to 88.14% and a 15.44% improvement in average class accuracy to 88.52%.

IV. Conclusion and Future Directions

This work proposes a framework to derive a reliability score for cauliflower harvest-readiness estimations that operates post-hoc during inference time without the need for human interaction. Our work combines a ResNet18 classification model with an unsupervised SC approach of saliency maps to derive a reliability score for classification predictions. Since the reliability value is in a fixed range between 0 and 1, it is intuitive and can be provided to the farmer as a decision support. In addition, the classification predictions can be adjusted, and the accuracy can be improved. We compare three saliency mapping approaches: Grad-CAM, OSM, and LIMEs. Grad-CAM proves to be the most useful in our scenario.

For our use case, our approach enables the correct harvest-readiness estimation on GrowliFlowerR, a subset of the GrowliFlower dataset, of approx. 4 out of 5 cauliflowers compared to the state-of-the-art approach ResNet18 which achieves only approx. 3 out of 4 correct predictions. Our framework offers the advantage of not requiring any interaction with the training process and it can be applied to already trained models without accessing or modifying the model architecture. We provide interpretable visualizations and a reliability score for the model's decision. Since we only consider false predictions in our framework, the approach can also be used for reliability dissemination in multiclass tasks.

References

[1] M. Brahimi, M. Arsenovic, S. Laraba, S. Sladojevic, K. Boukhalfa, and A. Moussaoui, "Deep learning for plant diseases: Detection and saliency map visualisation," in *Human and Machine Learning*. Cham, Switzerland: Springer, pp. 93–117, 2018.

[2] S. Fountas, B. Espejo-García, A. Kasimati, N. Mylonas, and N. Darra, "The future of digital agriculture: Technologies and opportunities," *IT Prof.*, vol. 22, no. 1, pp. 24–28, Jan. 2020.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[4] J. Kierdorf et al., "GrowliFlower: An image time-series dataset for GROWth analysis of cauLIFLOWER," *J. Field Robot.*, vol. 40, no. 2, pp. 173–192, Mar. 2023.

[5] T. Kim, H. Kim, K. Baik, and Y. Choi, "Instance-aware plant disease detection by utilizing saliency map and self-supervised pre-training," *Agriculture*, vol. 12, no. 8, p. 1084, Jul. 2022.

[6] T. V. Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agric.*, vol. 177, Oct. 2020, Art. no. 105709.

[7] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever Hans predictors and assessing what machines really learn," *Nature Commun.*, vol. 10, no. 1, pp. 1–8, Mar. 2019.

[8] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," *Geosci. Frontiers*, vol. 7, no. 1, pp. 3–10, Jan. 2016.

[9] A. Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Adv. Neural. Inf. Process. Syst.*, 2002, pp. 849–856.

[10] M. T. Ribeiro, D. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[11] 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 Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 817–824, Aug. 2020.

[12] P. Schramowski et al., "Making deep neural networks right for the right scientific reasons by interacting with their explanations," *Nature Mach. Intell.*, vol. 2, no. 8, pp. 476–486, Aug. 2020.

[13] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[14] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, "Beyond explaining: Opportunities and challenges of XAI-based model improvement," *Inf. Fusion*, vol. 92, pp. 154–176, Apr. 2023.

[15] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, pp. 818–833, 2014.