Explainable AI for deep learning based potato leaf disease detection

1st Siwar Bengamra *†, 2nd Ezzeddine Zagrouba * and 3rd André Bigand †
* LIMTIC Laboratory
University of Tunis El Manar
Ariana, Tunisia
† LISIC Laboratory
University of the Littoral Opal Coast
Calais Cedex, France
Email: bengamra.siwar@gmail.com, e.zagrouba@gmail.com and bigand@univ-littoral.fr

2023 IEEE International Conference on Fuzzy Systems (FUZZ) | 979-8-3503-3228-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/FUZZ52849.2023.10309803

***Abstract*—The field of agriculture research has been transformed by deep learning, which has demonstrated impressive capabilities in detecting and classifying plant disease from leaf images. Although several deep learning-based models are proposed, they are considered as blackbox, lack transparency and are notoriously difficult to interpret. Recently, eXplainable Artificial Intelligence (XAI) methods has demonstrated the potential to interpret the model decision-making process through saliency explanations that highlight the most relevant parts of the input image deemed important for predictions. In this work, we proposed a new XAI saliency method for explanation of potato disease detector based on particular perturbations driven by intermediate object detection results. In order to compare our proposed method with the state of the art, qualitative and quantitative experiments are performed for potato leaf disease detector models on PlantDoc dataset.**

***Index Terms*—Potato leaf disease detection, Deep Learning, Explainable AI, Perturbation-based methods**

I. Introduction

Potato leaf diseases are the primary cause of crop losses, jeopardizing food security with broad impacts on society and economy. Traditional visual detection of potato plant diseases is often tedious requiring human expertise to check if plants are affected or suitable for human consumption. Furthermore, the traditional process takes a lot of time, and is expensive especially when the farm is wide with a lot of plants. So automatic early detection is crucial to reduce (if possible prevent) the disease transmission from unhealthy to healthy plants.

With recent advances in deep Convolutional Neural Networks (CNN), the early detection of potato leaf diseases has been made much easier, very quickly and cheaper in comparison to the traditional process. The trained models could even be suitable for consumer applications on smartphones. In recent years, a lot of research has been carried out in potato leaf diseases detection (i.e. localization and classification of disease) based deep learning models [1]–[3].

Despite the remarkable performance, deep models are still considered as blackbox and explanation of its decisions remains difficult and not intuitive for human users. There is an ever-growing demand for eXplainable Artificial Intelligence (XAI) to ensure trust on predictions performed by the deep models. With the explainability methods, it became possible to aid scientists, or end users, in analyzing the reasons for high performances of detection and possible failures in certain cases. A particular class of XAI methods, named saliency or attributions methods, provide saliency maps (or heatmap) that highlight which parts of the input image deemed relevant for the prediction (i.e. classification or object detection result) performed by the learned model. There are two main categories of attribution methods, backpropagation-based methods which compute attributions by back-propagating the output of the network back to the input image space, and perturbation-based methods which perturb (or mask) the input image and measure the effect this perturbation has on the model's output. The major advantage of perturbation-based methods is that are applicable to any trained model (even the more complex ones) regardless of its architecture unlike the backpropagation-based category which requires access to the internal working (i.e. architecture, gradients, etc.) of the model to generate explanations.

Although several works have investigated explainability methods in the agricultural field, it should be emphasized that existing studies are made for explainability of leaf disease classification. But, there is a difference between outputs of classifiers and object detectors. So, for classification we are asked to explain a class probability unlike the object detection problem which requires explanation of classified bounding box enclosing potato leave disease and identified by the coordinates of its corners with class probability. This motivate us to propose a perturbation-based method for explaining both localization and classification aspects of potato leaf diseases.

Our method is inspired by the Detector Randomized Input Sampling for Explanation method (D-RISE) [4], the first perturbation-based method proposed to explain the predictions of object detectors, which extended the randomized input masking idea originally proposed for the explainability of classifier [5] to explain object detectors. According to recent previous works [6], [7], random perturbations for saliency map generation seems to produce coarse grained results and requires an excessive computing time when attempting to get

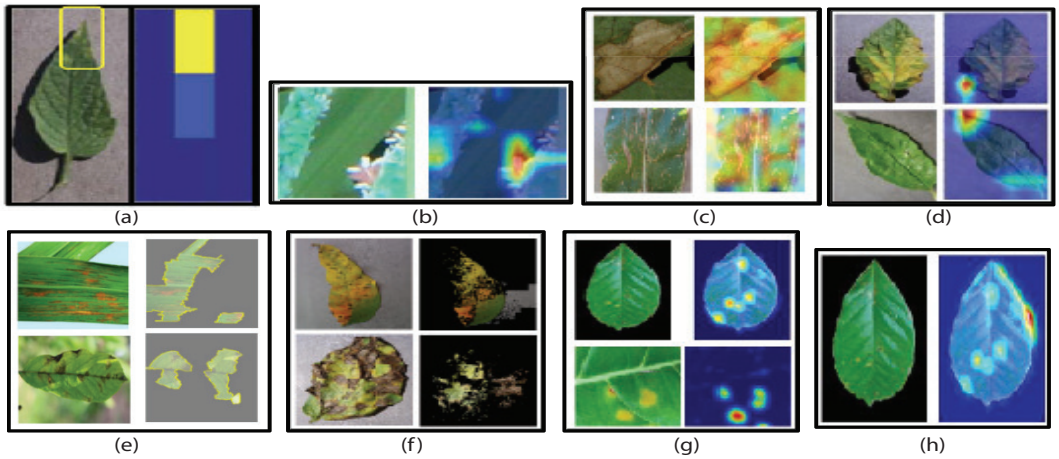

Fig. 1. Examples of available visual explanation results (right images for each block) produced for classifier models on plant leaves images (left images in each block) with different attribution methods.(a) Occlusion method, (b,c,d) GradCAM method, (e,f) LIME method, (g) GradCam++ and (h) ScoreCAM. The saliency maps are taken from papers cited in section II.

generalized results. To address such limitations, we propose to iteratively perform particular perturbations spatially driven by intermediate predictions results in order to produce refined results with minimal number of perturbations.

Our main contributions are as follows: (1) we propose a novel XAI method for object detectors, which analyzes changes in intermediate prediction results obtained with masked version of input image to generate new particular perturbation for the next iteration while stable state is not yet reached. (2) The proposed method could be applied to any object detector model without access to internal information of the architecture. (3) Qualitative and quantitative evaluations on potato leaves images and disease detection models are performed.

The rest of the paper is organized as follow: first, an overview of previous related works is presented in section II. Then, the proposed method is put forward in section III. In section IV, experimental results are presented and discussed. Finally, the conclusion of this research paper is shown in section V.

II. Related work

In agriculture field, many related works focused in providing explainability to the predictions made by the trained deep models using attribution methods (perturbation-based or backpropagation-based).

As a backpropagation-based method, Gradient-weighted Class Activation Mapping (Grad-CAM) [8] has been applied in recent works, such as in [9] to highlight regions responsible for the classification of the image into Flower or Non Flower Class (Fig. 1-(b)), and in [10] to show the disease portions focused by the classifier model on leaf images (Fig. 1-(c)). It should be noted that the GradCAM method in [11] has shown the need of the classifier model to learn more discriminative features since relevant pixels in the saliency map do not represent the disease areas (see Fig. 1-(d)). More CAM variants have been recently used to explain how classifier models operate in order to trust them. In [12], a comparison between GradCAM++ [13] and ScoreCAM [14] has been performed to exhibit the differences in the identification of regions responsible for classifying coffee leaf image as abnormal (Fig. 1-(g,h)).

On the other side, other works have explored perturbation-based methods for visual explanation of plant disease classifications. In [15], authors investigate the use of occlusion map method [16] to understand how the used deep model perform disease classification of tomato plant. The heatmaps obtained in [15] roughly showed the symptoms considered in the classification decision, as illustrated in Fig. 1-(a). Additionally, the popular perturbation-based method Local Interpretable Model-Agnostic Explanations (LIME) [17] is investigated in recent works to interpret the decisions made by classifier of plant diseases. In fact, the authors in [18] used LIME to check if the model focuses on the affected regions of the leaves to make its predictions (see Fig. 1-(f)). LIME is also used in the research work [10] to confirm that the proposed classifier focuses on the disease portion, as shown in Fig. 1-(e).

We remark that related previous works have investigated existing visual explainability methods to understand classifier model behavior for plant leaf disease recognition. Unfortunately, no quantitative evaluation is performed for the used explainability methods, limited only to visual evaluation. Additionally, there is a lack of works on the explainability of plant disease (i.e object) detection which consider both the localization and classification aspects of the detection. It should be noted that no proper XAI saliency map method is proposed in potato leaf disease detection. After this study, one can say that D-RISE is the only perturbation-based method founded to explain both model's classification and localization [19]. The other perturbation-based methods designed for image classifiers, such as LIME and RISE, require research efforts to extend them to be used for object detectors [20]. That is why we take D-RISE as a state-of-the-art in order to improve explainability accuracy.

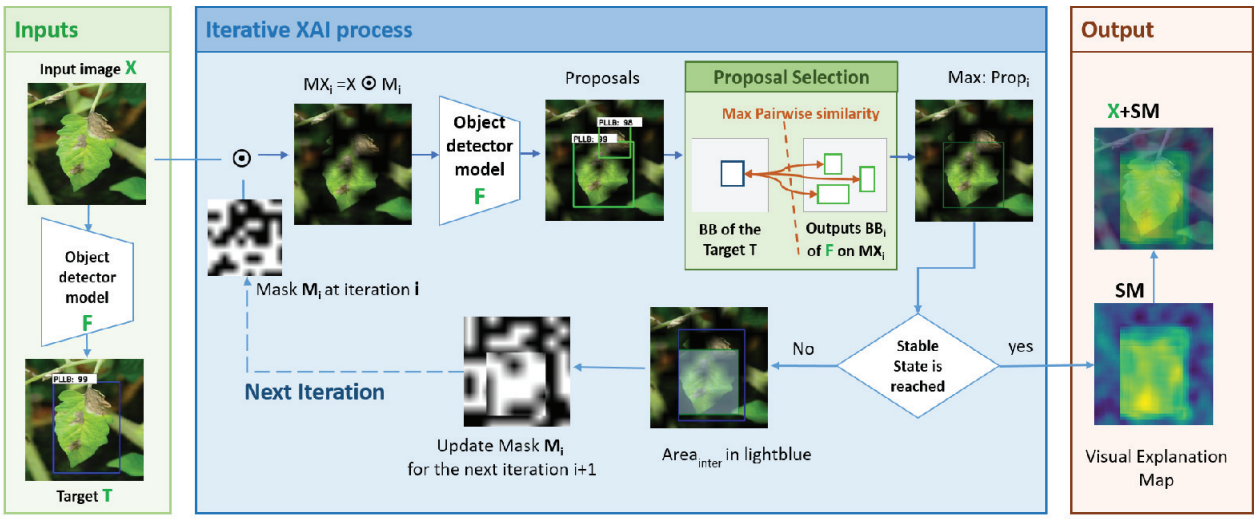

Fig. 2. Overview of the proposed explainability method workflow.

III. Proposed Method

The proposed method gets an object detector model F, an input image X and a target prediction T to explain identified by the coordinates of the bounding box BB (x_{min}, y_{min}, x_{max}, y_{max}) with associated probability vector $P = [P_{PLEB}, P_{PLLB}]$, representing the probability that BB belongs to each of the classes: PLEB (Potato Leaf Early Blight) or PLLB (Potato Leaf Late Blight). For example, if the target to explain is classified as potato leaf late blight with a probability equals to 0.84, the probability vector will be $P = [0, 0.84]$. The output of our proposed method is a saliency map SM having the same size of X and highlighting the most important pixels in X that contribute to both classification and localization of T by F. Fig 2 gives an overview of the proposed method workflow.

The key idea is to perturb the image X with a mask M_i, then fed the masked version MX_i of X ($MX_i = X \odot M_i$) to the model F to have proposals of bounding boxes. We select the proposal $Prop_i$ that has maximum similarity with the target T. $Prop_i$ is defined by it bounding box BB_i (x^i_{min}, y^i_{min}, x^i_{max}, y^i_{max}) with probability vector $P^i = [p^i_{PLEB}, p^i_{PLLB}]$. The similarity between two detections is computed as follow:

$$Sim_i(T, Prop_i) = IoU(BB, BB_i).P^i.cosine_{sim}(P, P^i) \quad (1)$$

where IoU is the intersection over union between two bounding boxes computed as a ratio of the area of overlap to the area of the union.

$$IoU(BB, BB_i) = \frac{BB \cap BB_i}{BB \cup BB_i} \quad (2)$$

and $cosine_{sim}$ is the cosine similarity of probability vectors to evaluate how similar the two bounding boxes BB and BB_i look to the network.

$$cosine_{sim}(P, P^i) = \frac{P.P^i}{\|P\| \, \|P^i\|} \quad (3)$$

For masking the input image, we adopt the same mask generation approach from [5]. This method consists of sampling a binary mask with a size smaller than image size by setting each element independently to 1 with probability p and to 0 with probability $(1-p)$, then upsampling it to the image size using bilinear interpolation. Instead of masking random regions as in D-RISE method, we propose to use the detector's output $Prop_i$ obtained at each iteration i, to differently perturb the image addressing spatial analysis. For this, we put out two interpretation regions comparing the target T to the obtained result $Prop_i$ as follow:

- $Area_{inter}$ is the intersection area between T and $Prop_i$,
- $OutArea_{inter}$ is the rest of the image outside the intersection area.

So we propose to perturb the interpretation regions separately. The $Area_{inter}$ are outcome where the model correctly detect part of the target T. So we aim to keep these pixels for the next iterations considering them important for the prediction of T. For this, we slightly perturb $Area_{inter}$ with its 10x10 neighborhood, by setting elements to 1 with probability $p = 0.8$ in the binary mask. The pixels of the intersection area $Area_{inter}$ will thus be softly masked to precisely investigate the differences between the importance of the pixels inside the $Area_{inter}$. For the rest of the pixels $OutArea_{inter}$, we investigate their importance values with new random perturbations.

All steps described above are repeated until having a high degree of similarity Sim_i between detection result $Prop_i$ and target T. Hence the stable state is reached when Sim_i converges to 1 and absolute difference between similarities from consecutive iterations is less than or equal to ε. So, the saliency map SM will be a weighted sum of masks M_i across all iterations $\{1, .., N\}$, where the weights are the similarity between T and $Prop_i$ at each iteration i.

$$SM_{F,X,T} = \sum_{1}^{N} sim_i(T, F(X \odot M_i)).M_i \quad (4)$$

IV. Experiments

In this section, we quantitatively and qualitatively evaluate the performance of the proposed explainability method on deep models used for potato disease detection.

A. Experimental setup

The value of ε was experimentally determined to be 10^{-3}.

a) Dataset: We have used the PlantDoc dataset recently published in [21] with instance-level annotations (i.e. bounding boxes enclosing leaf disease object). PlantDoc is an open access repository of plant leaf images in a field environment published online at [22] which contains 2568 images of leaves across 13 plant species and 27 classes. From this dataset, we extract only images of potato leaves splitted into two diseases: Potato leaf early blight and Potato leaf late blight. Sample images with annotated bounding boxes from PlantDoc dataset are shown in Fig. 3.

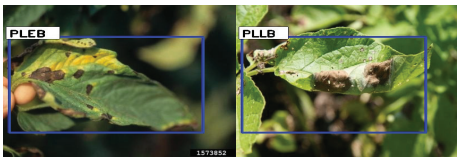

Fig. 3. Sample images from dataset with annotated bounding boxes for two disease classes: Potato Leaf Early Blight (left) and Potato Leaf Late Blight (right).

b) Deep learning models: For the training and detection of potato plant diseases, we used Faster RCNN [23], that offers high performances in plant leaf disease detection. So the first deep model $Model1$ used in this research work is Faster RCNN based on Resnet50 architecture, pre-trained on ImageNet dataset [24] and finetuning on the target potato leaves images to detect diseases. Since data augmentation techniques are widely used during training to enhance the dataset's diversity and avoid overfitting, we train the same setup of $Model1$ with data augmentation techniques, namely horizontal flipping and vertical flipping, to put out $Model2$ for potato leaf disease detection. The architecture of $Model1$ and $Model2$ is illustrated in Fig. 4. The used data augmentation

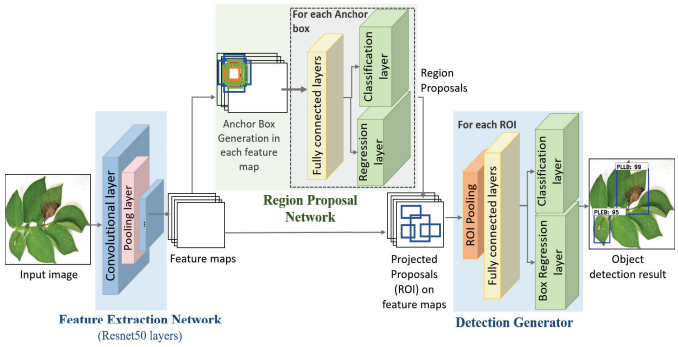

Fig. 4. Graphical illustration of the architecture of $Model1$ and $Model2$.

techniques are applied on-the-fly for each batch since this can generate more unique training images than offline augmentation, and can hence improve generalization capability [25]. The metric used to evaluate object detector performances is the mean Average Precision (mAP) [26]. $Model2$ yields a 46.22% of mean average precision (mAP) across two classes, slightly outperforming $Model1$ which achieves 39.37%. Experiments were conducted on a desktop computer with 3.9GHz Intel i7-9750H CPU and an NVidia GeForce GTX 1650 GPU.

B. Evaluation metrics

We used the deletion and insertion metric proposed in [5] to quantitatively evaluate the proposed explanation method. The idea behind the deletion metric is that the gradual removal the top N salient pixels (i.e. mask them with blurred ones) from the input image by following the order suggested by the obtained saliency map will force the model to change its decision. If the pixels that were highlighted in the saliency map are truly important, we would expect a rapid decrease in detection performance as the model's output deviates from the original prediction rapidly with increasing N [27]. On the other hand, starting with a blurred image, the insertion metric consists in progressively showing the most important pixels to the model and measuring how fast the object detection approaches the target. We would expect a sharp increase in object detection performance as more and more important pixels are introduced. We follow [28] to blur input images by using Gaussian Blur with kernel size = 51 and sigma = 50.

C. Quantitative evaluation

In this section, we provide a comparative quantitative evaluation of explainability methods based on deletion and insertion curves generated for $Model2$. Fig. 5 shows the deletion curve, which measures the drop in similarity by iteratively removing important pixels from the input image. The importance of pixels are given by the obtained saliency maps. We see that our method drop faster than D-RISE. This implies that the relevant pixels highlighted with our saliency maps are more faithful to the model compared to those obtained with D-RISE. Fig. 6,

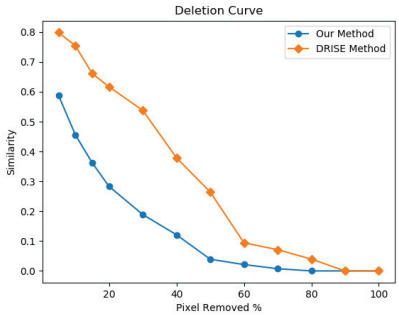

Fig. 5. Deletion plot.

on the other hand, illustrates the increase in similarity as more and more important pixel are added back to the image. As can be seen, the explanation based on our method increases faster than that based on D-RISE. This sharp rise during insertion

supports that our method better identifies the contribution of the image pixels for the prediction. The Area Under the

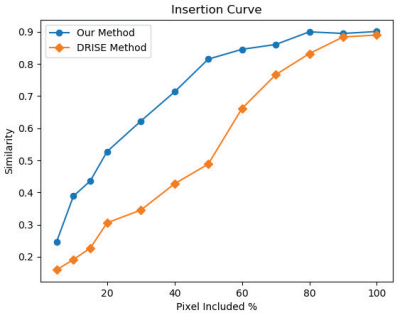

Fig. 6. Insertion plot.

deletion and insertion Curves (AUC) represent the deletion and insertion scores, respectively. Table I reports the obtained deletion and insertion scores over test images. For both potato disease detector models ($Model1$ and $Model2$), our proposed method achieves better performance on both metrics compared to D-RISE method. The obtained deletion scores indicate that removing salient pixels based on our saliency maps forced the models to change its decisions, so the network's output will quickly deviate from the original prediction, and hence the detection performance will rapidly diverge. Analyzing the insertion scores, we can deduce that our method converges better to the original prediction.

TABLE I
QUANTITATIVE EVALUATION IN TERMS OF DELETION (LOWER IS BETTER) AND INSERTION (HIGHER IS BETTER) SCORES

Model	Metric	Our Method	D-RISE method
$Model2$	Deletion	**0.114**	0.275
	Insertion	**0.697**	0.532
$Model1$	Deletion	**0.059**	0.154
	Insertion	**0.26**	0.19

D. Visualization results

In our method, performing separately and differently the perturbation of interpretation regions made it produce a more precise saliency map. Our method could hence more faithfully reveal the object detection process. As shown in Fig 7, our proposed method presents higher concentration at the relevant pixels in comparison with the D-RISE method. We qualitatively compare the saliency maps produced by our proposed explanation method for the used object detector models $Model1$ and $Model2$. Examples of visualization results are shown in Fig. 8. According to the obtained results, we observe that the important area, represented with green and yellow colors, is highlighted with a regular shape which can be explained by the fact that the model has learned from bounding box annotations (i.e. class and localization information) and the target to explain has also regular shape.

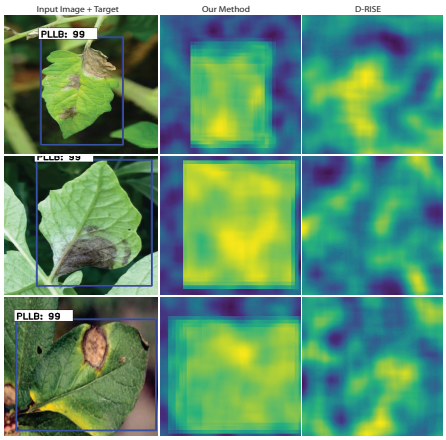

Fig. 7. Qualitative comparison of our proposed XAI method with D-RISE method for $Model2$.

Inside the important area, our method provides fine-grained interpretation by locating most relevant pixels (e.g. yellow pixels in the saliency maps at second row of Fig. 8) that correspond mainly to the affected parts of disease in leaves. Qualitative comparison between saliency maps produced for $Model1$ and $Model2$ shows that the saliency map become more focused on affected parts as the object detection performance increases with $Model2$. Our proposed explanation method can work as a hint to determine the limitations of deep models. For instance, the saliency maps produced for $Model1$ indicate that it did not learn properly discriminative features of potato diseases in the images recommending revising the learning configuration for training (e.g. data quality, hyper-parameters of models, etc.).

V. CONCLUSION

The presented work introduces, for the first time, a visual explanation method for deep learning model that detect and localize potato plant diseases. The explanation of object detection results is based on particular perturbation which is not random but driven by the intermediate model's output obtained for masked versions of input image. The experimental results have proved that our proposed XAI method achieves better results than the implementation of the D-RISE method. Thanks to the explanation results, we can control and track further performance improvement of other versions of detector models. As future work, it is interesting to prove the efficiency of our XAI method, not only, on other detector models, but also on other plant diseases.

REFERENCES

[1] J. Johnson, G. Sharma, S. Srinivasan, S. K. Masakapalli, S. Sharma, J. Sharma, and V. K. Dua, "Enhanced field-based detection of potato blight in complex backgrounds using deep learning," *Plant Phenomics*, vol. 2021, 2021.

[2] J. Rashid, I. Khan, G. Ali, S. H. Almotiri, M. A. AlGhamdi, and K. Masood, "Multi-level deep learning model for potato leaf disease recognition," *Electronics*, vol. 10, no. 17, p. 2064, 2021.

[3] B. Anjanadevi, I. Charmila, N. Akhil, and R. Anusha, "An improved deep learning model for plant disease detection," *International Journal of Recent Technology and Engineering*, vol. 8, no. 6, pp. 5389–5392, 2020.

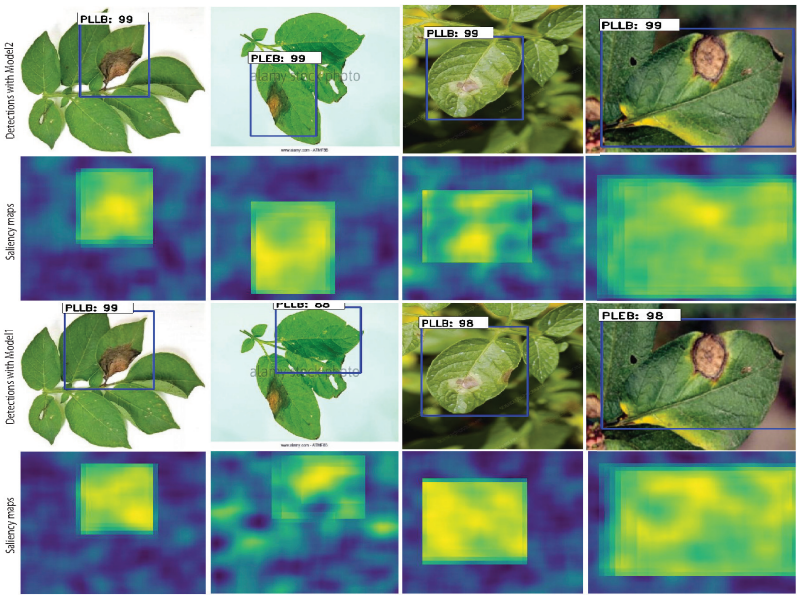

Fig. 8. Qualitative comparison between saliency maps produced by our proposed method for $Model2$ (second row) and $Model1$ (fourth row) on few example images where $Model2$ is more efficient than $Model1$. The importance in saliency maps increases from blue to yellow.

[4] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. Saenko, "Black-box explanation of object detectors via saliency maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 443–11 452.

[5] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *BMVC*, 2018.

[6] Q. Yang, X. Zhu, J.-K. Fwu, Y. Ye, G. You, and Y. Zhu, "Mfpp: Morphological fragmental perturbation pyramid for black-box model explanations," in *2020 25th International conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 1376–1383.

[7] S. Sattarzadeh, M. Sudhakar, A. Lem, S. Mehryar, K. N. Plataniotis, J. Jang, H. Kim, Y. Jeong, S. Lee, and K. Bae, "Explaining convolutional neural networks through attribution-based input sampling and block-wise feature aggregation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 11 639–11 647.

[8] 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.

[9] S. V. Desai, V. N. Balasubramanian, T. Fukatsu, S. Ninomiya, and W. Guo, "Automatic estimation of heading date of paddy rice using deep learning," *Plant Methods*, vol. 15, no. 1, pp. 1–11, 2019.

[10] P. S. Thakur, P. Khanna, T. Sheorey, and A. Ojha, "Explainable vision transformer enabled convolutional neural network for plant disease identification: Plantxvit," *arXiv preprint arXiv:2207.07919*, 2022.

[11] X. Zhang, H. Gao, and L. Wan, "Classification of fine-grained crop disease by dilated convolution and improved channel attention module," *Agriculture*, vol. 12, no. 10, p. 1727, 2022.

[12] M. Yebasse, B. Shimelis, H. Warku, J. Ko, and K. J. Cheoi, "Coffee disease visualization and classification," *Plants*, vol. 10, no. 6, p. 1257, 2021.

[13] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[14] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[15] M. Brahimi, K. Boukhalfa, and A. Moussaoui, "Deep learning for tomato diseases: classification and symptoms visualization," *Applied Artificial Intelligence*, vol. 31, no. 4, pp. 299–315, 2017.

[16] 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.

[17] 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*, 2016, pp. 1135–1144.

[18] M. H. K. Mehedi, A. S. Hosain, S. Ahmed, S. T. Promita, R. K. Muna, M. Hasan, and M. T. Reza, "Plant leaf disease detection using transfer learning and explainable ai," in *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2022, pp. 0166–0170.

[19] T. T. H. Nguyen, V. B. Truong, V. T. K. Nguyen, Q. H. Cao, and Q. K. Nguyen, "Towards trust of explainable ai in thyroid nodule diagnosis," *arXiv preprint arXiv:2303.04731*, 2023.

[20] C. Silva, A. Morais, and B. Ribeiro, "A generic approach to extend interpretability of deep networks," in *Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings*. Springer, 2022, pp. 488–499.

[21] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, and N. Batra, "Plantdoc: A dataset for visual plant disease detection," 2019.

[22] S. et. al. (2019) Plantdoc dataset. [Online]. Available: https://public.roboflow.com/object-detection/plantdoc/

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[25] S. O'Gara and K. McGuinness, "Comparing data augmentation strategies for deep image classification," in *IMVIP 2019: Irish Machine Vision and Image Processing Conference Proceedings*, 2019.

[26] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–308, 2009.

[27] J. Cooper, O. Arandjelovic, and D. Harrison, "Believe the hipe: Hierarchical perturbation for fast and robust explanation of black box models," *arXiv preprint arXiv:2103.05108*, 2021.

[28] Q. Zhang, L. Rao, and Y. Yang, "A novel visual interpretability for deep neural networks by optimizing activation maps with perturbation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3377–3384.