Supplementary Materials

Strategy analysis of RGB scientific dataset

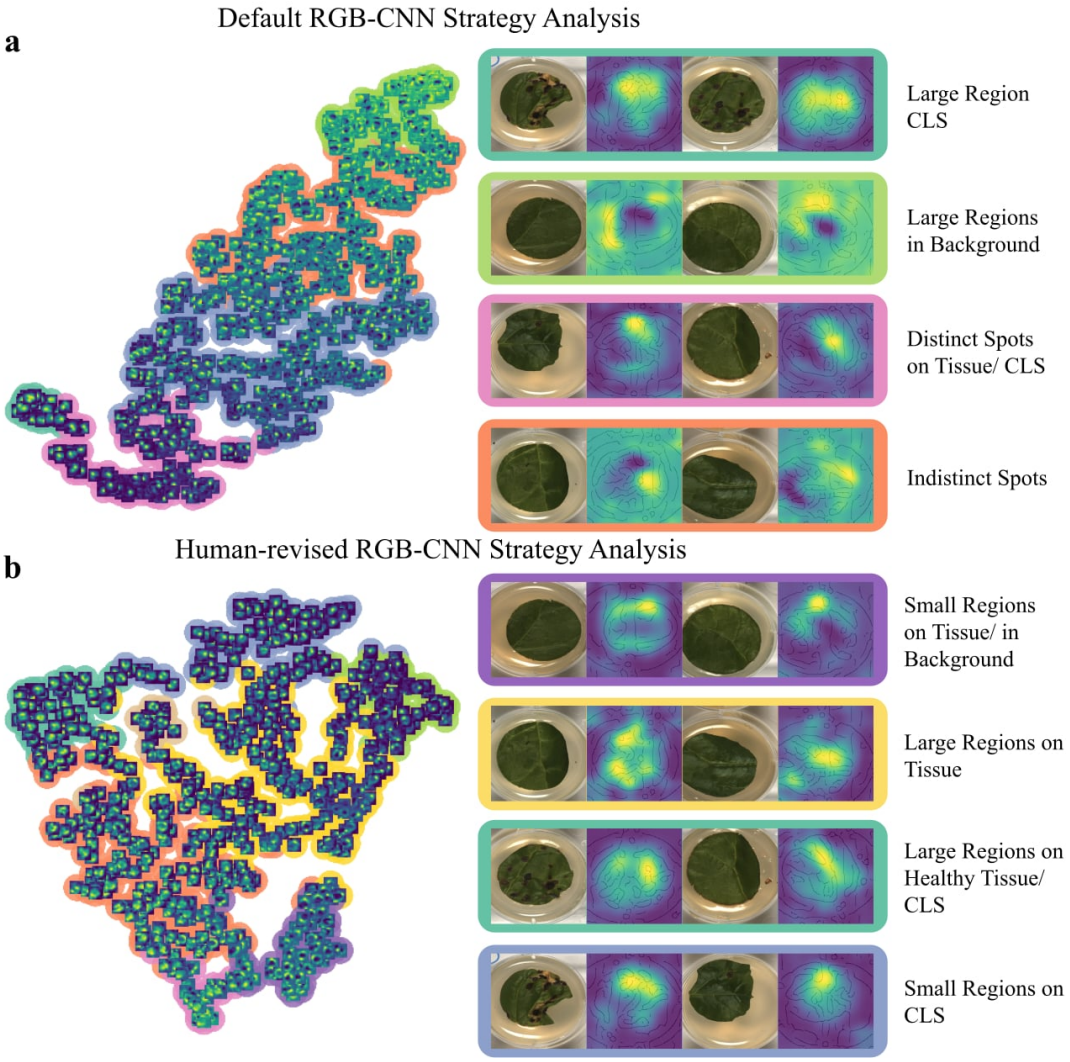

Figure 1: Cluster analysis of the different decision strategies after training CNNs on the RGB data with the cross-entropy loss (Default) in (a) and with the RRR loss in (b). The images are visualized in a two-dimensional t-SNE embedding and colored by the spectral clustering assignments.

Fig. 1(a) shows the strategies of the CNN trained on the RGB data for data points only in the test set. When CLS were visible the RGB-CNN correctly identifies these as relevant features for classifying the samples as inoculated. However, for many inoculated samples, for which no spots are visible the CNN surprisingly focuses on regions in the background, specifically often on the nutrition solution (agar), which the tissue was embedded in. Also for healthy samples the RGB-CNN focuses on the background.

One can identify different decision strategies, after training with RRR as illustrated in Fig. 1(b). However, even using different hyper-parameters for RRR, we were not able to reach a converged state such that the RGB-CNN fully ignores the background.

Faithfulness of learned explanations

Investigating the faithfulness of an explanation method is a very valid and relevant topic of research ([1], [2], [3]). From our experiments the objection can be made that the revised models have merely learned to produce acceptable explanations, but still focus on wrong features.

Although we have shown the generalization performance in previous experiments using non-confounded test sets, we ran further experiments to investigate the faithfulness of explanations that have been revised using the XIL framework. The questions we wanted to answer were the following: (Q1) are the features learned interactively using XIL more relevant for the original task than the identified features of the default model? We note that an important underlying assumption here is that the user feedback is correct and faithful. (Q2) Is the XIL revised model more strongly influenced by it's learned explanations in comparison to the default model with it's unrevised explanations?

We focused on using a more widely used dataset of the ML and computer vision community: MSCOCO 2014 [4]. This dataset presents a multi-label image classification problem of commonly found objects and is completed with a masked segmentation for each class of each sample. For the following experiments we used a subset of the COCO14 classes, focussing on the five classes: elephant, giraffe, cat, dog and truck.

As the MSCOCO dataset is a non-confounded dataset, the task when using XIL with this dataset is therefore to mainly improve the model to focus on right reasons, rather than penalizing it when focusing on wrong reasons. A characteristic of CE and RRR, is that both methods revise an ML model when it is using wrong features for a right prediction, but not when it is *not* using a feature for a right prediction. More specifically, a user might want to direct a model's attention to features that she finds very relevant (similar to Selvaraju *et al.*'s HINT method [5]). For this reason we adapted the RRR loss of Eq. 1 in the main text to the following, resulting in a HINT-like extension:

$$L(\theta,\ X,\ y,\ A) = \underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K} -c_k y_{nk}\log(\hat{y}_{nk})}_{\text{Right answers}} + \underbrace{\lambda_1 \sum_{n=1}^{N}\sum_{d=1}^{D}\left(A_{nd} - expl(y_n)\right)^2}_{\text{Learning explanations}} + \underbrace{\lambda_2 \sum_{i}\theta_i^2}_{\text{Weight regularization}} \quad . \tag{1}$$

The notations are largely the same as for Eq. 1 of the main text. $expl(y_n)$ represents the specific differentiable explanation method that produces an explanation in the dimensions of the input space given the true class label y_n. In comparison to the RRR loss used in previous experiments the matrix A now denotes a value of 1 for features that the user finds relevant, -1 for irrelevant and 0 for features, where the user is indifferent. In our experiments we only used values of 1 and -1. Additionally, to properly compute the difference between the user and model explanation we rescaled the model explanation to the range $[0, 1]$, thus additionally enforcing the model to ignore irrelevant regions.

For the following experiments we again focused on the GRAD-CAM explanation method. In this case we downscaled the user annotations to the original output dimensions of the GRAD-CAM prior to computing Eq. 1. The user interaction was again simulated, whereby the user annotations corresponded to the ground truth class segmentations provided with the COCO dataset.

Fig. 2 shows several example images (Fig. 2(a)) for which the default explanations are partly correct (Fig. 2(b)). However, it would be valid for a user to be unsatisfied with these explanations, given that only small regions of the to-be-predicted objects are highlighted. These results highlight that even the GRAD-CAM method produces explanations that a human user might not fully accept. With XIL in the form of Eq. 1 these explanations could be refined to coincide more with the user's explanations (Fig. 2(c)). we note that the default model was trained for as many iterations as the XIL model.

To answer Q1 we applied the method of [6], termed "Remove and Retrain" (ROAR). The idea here is to investigate indeed how relevant the features are that different explanation methods have deemed as important. This is done by removing a certain percentage of relevant features that an explanation method has identified, set these features to the mean of the training data and retrain a model from an initial parameter setting. If the model produces a low prediction performance this is an indication that the features of the explanation method are indeed relevant for the task. If the performance is high this is an indication that there are equally or more relevant features available for the task.

Fig. 3 shows the results of ROAR where the initial on ImageNet-pretrained VGG-16 [7] was retrained until convergence using the modified datasets. This was repeated for random explanations as a baseline, the default trained model GRAD-CAM explanations and the XIL revised GRAD-CAM explanations. One can indeed observe that given the assumption of relevant and faithful user feedback, with XIL it is possible for a differentiable model to improve its explanations to focus on more relevant features.

With the previous experiment we could show that with a human in the loop a model can be revised to focus on more relevant features, which accord more strongly with the user's explanations, even if the model's original explanations were not considered as entirely wrong. ROAR, however, was developed to test explanation methods which were not explicitly trained to improve their explanations. This is different in the XIL setting. Due to that, for ROAR, the same model is retrained over all conditions, we have not yet shown, that the revised model actually

Figure 2: Several examples from MSCOCO 2014. The left column (a) presents the original images, the middle column (b) presents the explanations (GRAD-CAM) after training without user feedback (default), the right column (c) presents the explanations after training with user feedback (XIL). Also here, as in the main text, light regions represent relevant regions for the model's decision, dark regions represent irrelevant regions. Here the user annotations were the complete class segmentations to illustrate that XIL can also aid in improving the explanations for non-confounded data.

focuses more strongly on it's learned explanations in comparison to the default model which had not optimized it's explanations. In other words it remains open to show that the explanations of the XIL revised model are more faithful to the model's decisions than the explanations of the default model are to the default model.

We therefore evaluated both models, the default and XIL revised model, on the test set where, similar to the ROAR experiment, we replaced a certain percentage of relevant features with the per channel mean. Particularly each model (default and XIL revised) was evaluated on the test set, where features were removed based on their explanations. Importantly this was set in comparison to evaluation on the test set, where random features had been removed.

The results can be found in Fig. 4, where also here a lower accuracy indicates a feature's importance for the specific model. One can observe that there is little difference between the evaluations of both models on the random-explanation-modified test set (baseline). There are however strong differences between evaluations on the test sets modified by their respective explanations, where the accuracy strongly drops for the XIL revised model based on its explanations, than the default model, even when taking the difference between baseline evaluations into account. Thus indicating that the learned explanations of the XIL revised model are more faithful to the model's decisions.

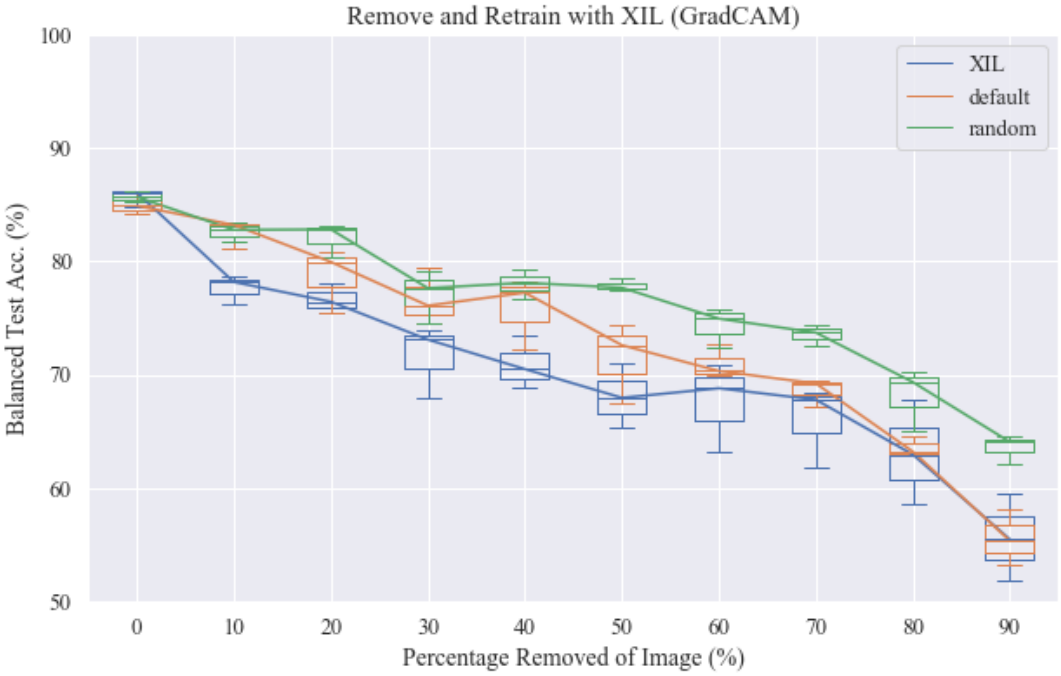

Figure 3: Three-fold cross-validation ROAR results for the MSCOCO 2014 dataset. A model was trained from an initial parameter setting with altered datasets where a certain percentage of the most relevant features were removed in the training and test set. The relevance of each feature is indicated by the explanations of the XIL trained and default trained model as well as random explanations. The lower the accuracy, the more likely it is that the removed features are informative for the original model.

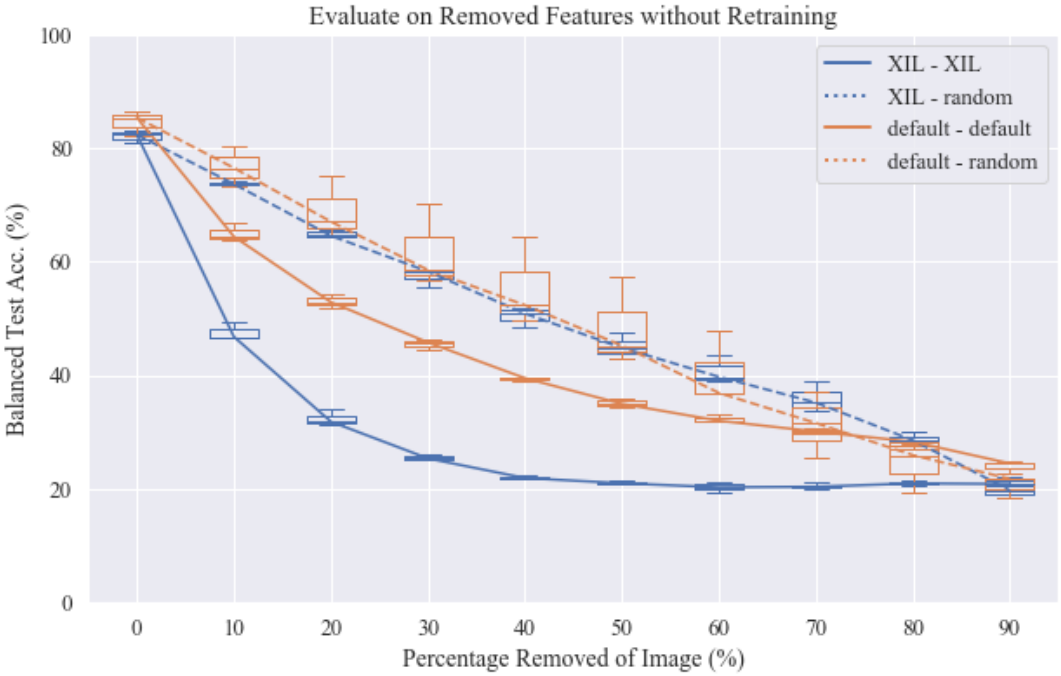

Figure 4: Three-fold cross validation of default and XIL trained models evaluated on the test set. Varying percentages of relevant pixels are removed from the test set, whereby the relevance is indicated by the explanations from the trained models (default, or XIL) or random pixel assignments.

Data augmentation and training based on prior knowledge

To compare to the setting of removing the confounders in the datasets based on prior knowledge before training, we train a model based on variants without confounders and test the resulting model on both variants, the original dataset (with confounders) and the dataset with removed confounders (w/o confounders). We here focused on the

HS data setting.

To remove confounders in the decoy fashion MNIST dataset we reverted to the original dataset (w/o confounders). Regarding the scientific plant phenotyping dataset, no variant without confounders is available. Therefore, we removed everything that is not belonging to the user annotated tissue and replaced it with the average reflectance of each hyperspectral channel. The trained models were tested on both dataset variants (w/o confounders and with confounders).

For both datasets, one can observe a similar behavior. Training with the XIL framework and training the model on the dataset variant w/o confounders results in a similar score when testing the model on the dataset without confounders. However, when testing on the original dataset one can observe that the model trained without confounders has an extreme accuracy drop. Instead, the models trained with the XIL framework generalized to not using confounders and perform well on both variants. The accuracy drop of XIL applied to the dataset without confounders can be explained by the change of data distributions. By a stronger weight of the right reasons one might be able to correct this even further.

	Fashion-MNIST		**Scientific Dataset (HS)**	
	trained w/o confounders	XIL	trained w/o of confounders	XIL
w/o confounders	85%	84%	86%	82%
with confounders	76%	85%	56%	95%

Table 1: Comparison of training on augmented data based on prior knowledge and XIL. The trained models were tested on both dataset variants (w/o confounders and with confounders).

Questionnaire and example online survey used in User Study

The competence of a classifier can be assessed by monitoring its behavior and beliefs over time, directability can be achieved by allowing the user to actively teach the model how to act and what to believe, while understandability can be approached by explaining the models decisions. To investigate how interaction with a machine and augmenting this interaction with explanations influences the trust of the user into the model we designed a questionnaire based on a binary classification toy problem. The Questionnaire document, as well as an example of the online survey for TC2, can be found at `https://github.com/ml-research/XIL`

Strategy analysis classification errors

Fig. 5 shows the class prediction versus underlying (ground truth) class for each sample of all four CNN training versions (from upper left to bottom right: default RGB-CNN, revised RGB-CNN, default HS-CNN, revised HS-CNN). Each sample is plotted in the embedding of Fig. 3 and Fig. 4 (main article). Particularly the HS-CNN, regardless of the training configuration, shows strong differences in the decision strategies for the two classes.

Explanations along hyperspectral data dimension

Below are several detailed examples of the explanations from the HS-CNNs. Figures 2-9 present the explanations from the default HS-CNN, showing samples from varying stages of disease progression (see captions for details). Figures 10-17 present the explanations from the revised HS-CNN, showing samples from varying stages of disease progression (see captions for details). Each Figure 2-17 depicts the sample in the leftmost panel, followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

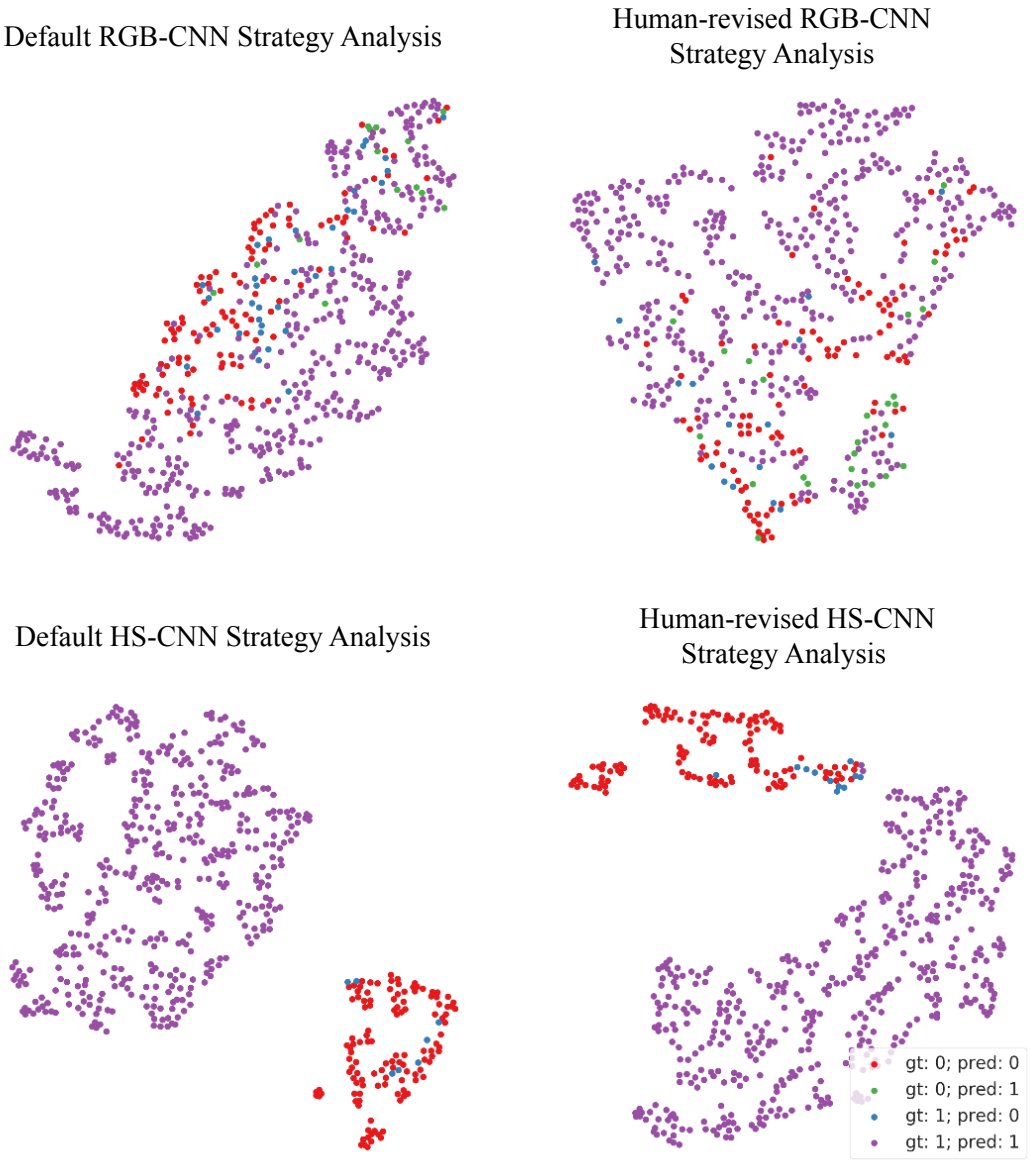

Figure 5: T-SNE embeddings of the different image type and training loss configurations colored by the ground truth (gt) and prediction (pred) labels of each sample. The top row depicts the results of training with RGB images, the bottom row with hyperspectral images. The left column shows the results of training only with the cross-entropy loss, whereas the right column shows the results of training with the right for the RRRLoss.

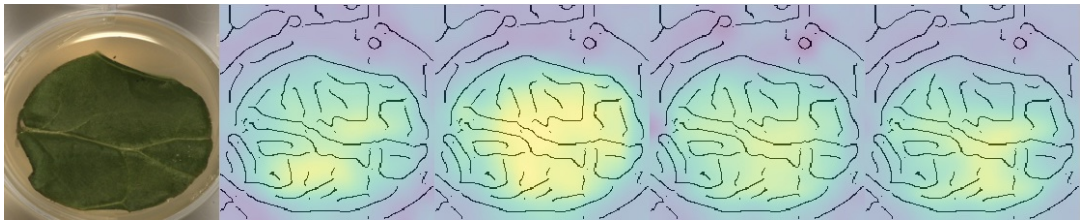

Figure 6: GRAD-CAMS with spatial and spectral activations from healthy sample of unrevised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

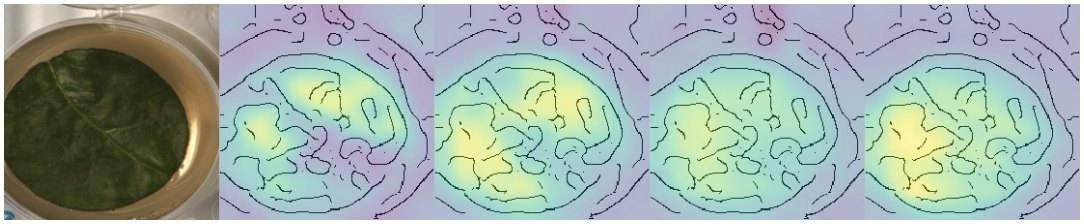

Figure 7: GRAD-CAMS with spatial and spectral activations from healthy sample of unrevised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

Figure 8: GRAD-CAMS with spatial and spectral activations from healthy sample of unrevised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

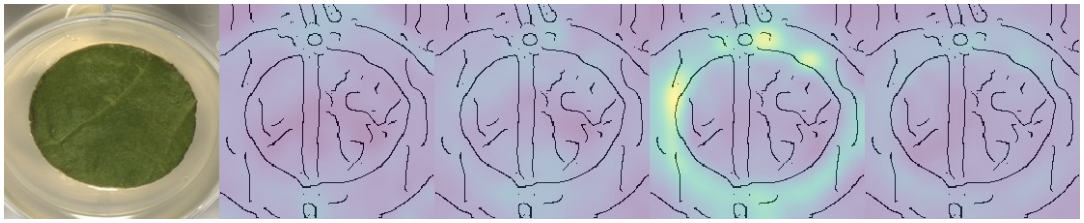

Figure 9: GRAD-CAMS with spatial and spectral activations from healthy sample of not regularized network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

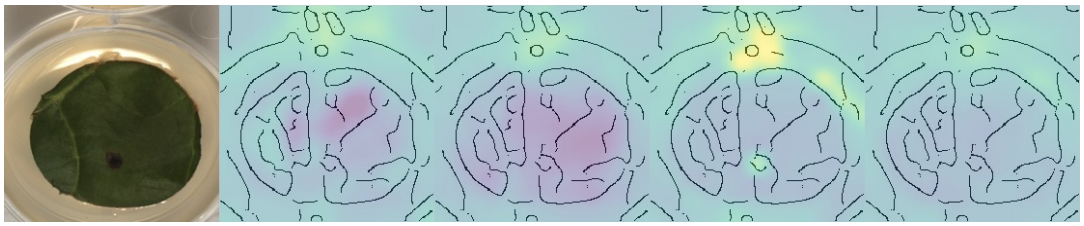

Figure 10: GRAD-CAMS with spatial and spectral activations from inoculated sample with single visible symptoms of unrevised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

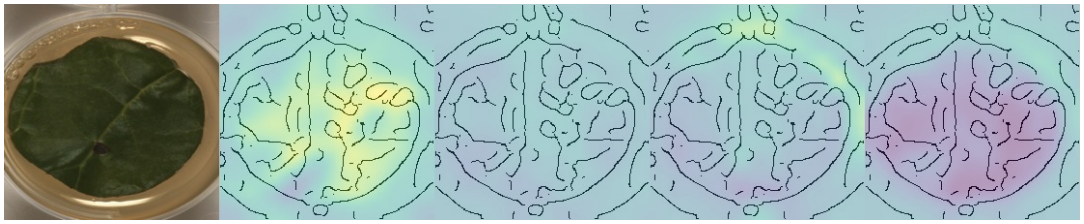

Figure 11: GRAD-CAMS with spatial and spectral activations from inoculated sample with single visible symptoms of unrevised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

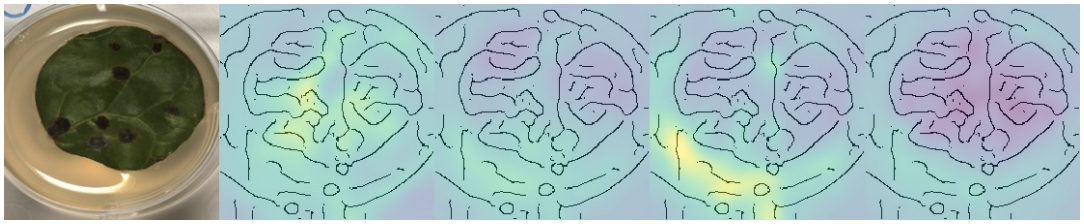

Figure 12: GRAD-CAMS with spatial and spectral activations from inoculated sample with multiple visible symptoms of not regularized network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

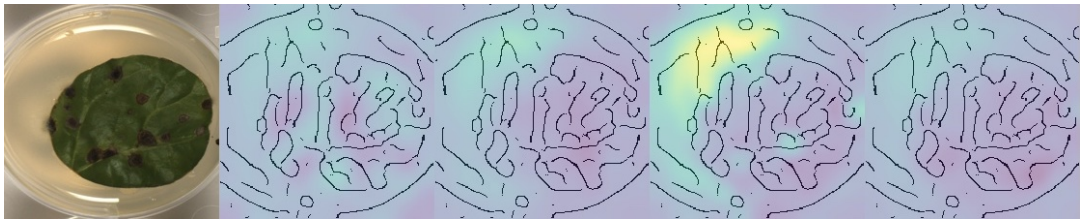

Figure 13: GRAD-CAMS with spatial and spectral activations from inoculated sample with multiple visible symptoms of unrevised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

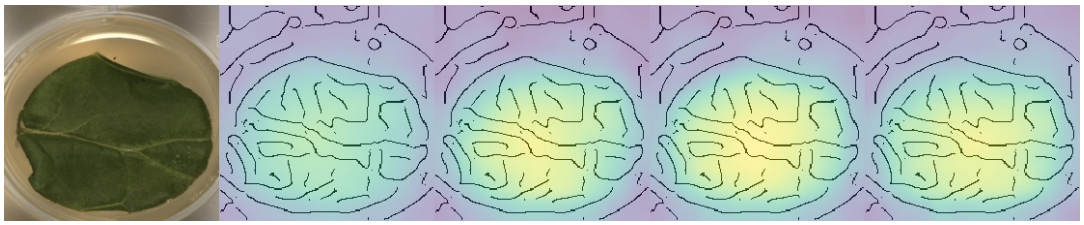

Figure 14: GRAD-CAMS with spatial and spectral activations from healthy sample of revised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

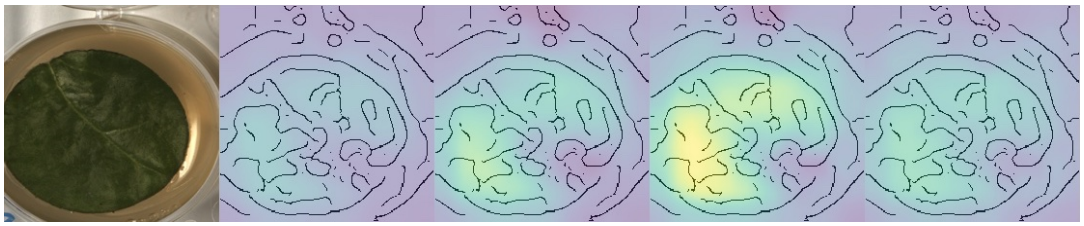

Figure 15: GRAD-CAMS with spatial and spectral activations from healthy sample of revised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

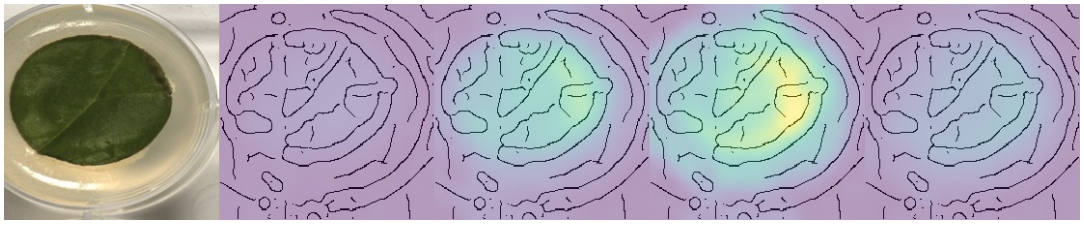

Figure 16: GRAD-CAMS with spatial and spectral activations from inoculated sample without visible symptoms of revised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

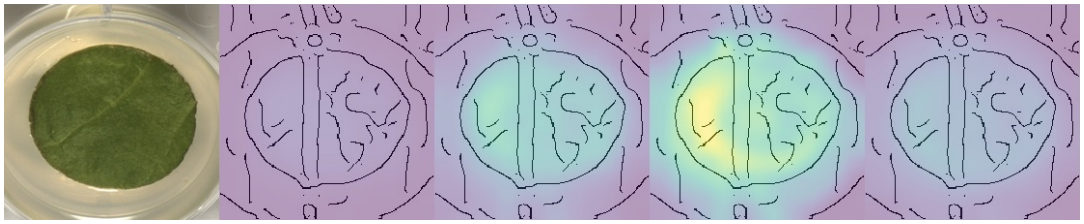

Figure 17: GRAD-CAMS with spatial and spectral activations from inoculated sample without visible symptoms of revised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

Figure 18: GRAD-CAMS with spatial and spectral activations from inoculated sample with single visible symptoms of revised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

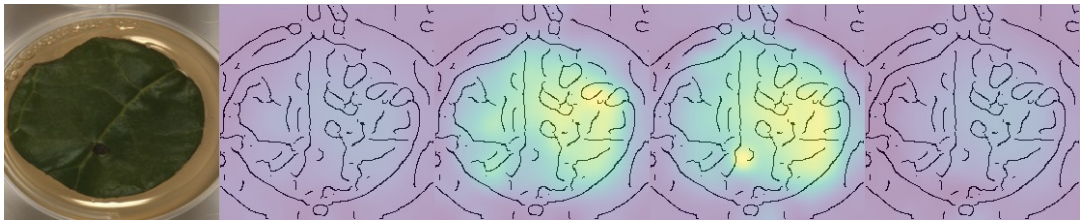

Figure 19: GRAD-CAMS with spatial and spectral activations from inoculated sample with single visible symptoms of revised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

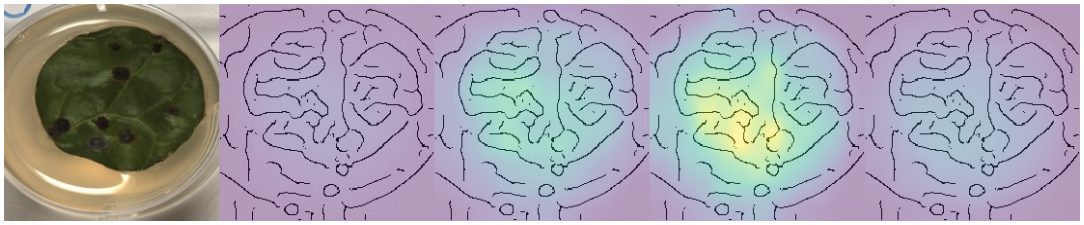

Figure 20: GRAD-CAMS with spatial and spectral activations from inoculated sample with multiple visible symptoms of revised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

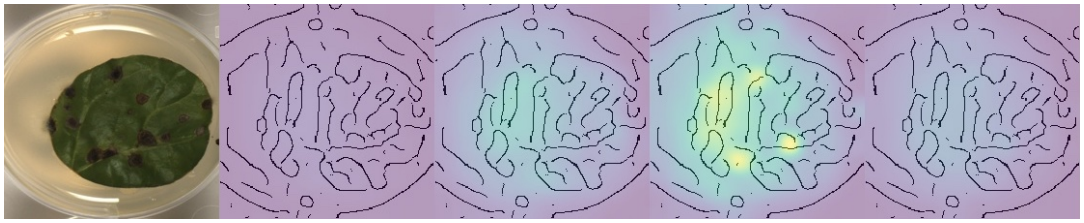

Figure 21: GRAD-CAMS with spatial and spectral activations from inoculated sample with multiple visible symptoms of revised network. Leftmost image shows the sample followed by the corresponding spatial activations maps mapped to four different hyperspectral areas. The areas are 380-537 nm, 538-695 nm, 696-853 nm, and 854-1010 nm.

References

[1] Dombrowski, A. *et al.* Explanations can be manipulated and geometry is to blame. In Wallach, H. M. *et al.* (eds.) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 13567–13578 (2019).

[2] Adebayo, J. *et al.* Sanity checks for saliency maps. In *Proceedings of Advances in Neural Information Processing Systems*, 9505–9515 (2018).

[3] Sixt, L., Granz, M. & Landgraf, T. When explanations lie: Why modified BP attribution fails. *CoRR* **abs/1912.09818** (2019). URL `http://arxiv.org/abs/1912.09818`.

[4] Lin, T. *et al.* Microsoft COCO: common objects in context. In *Proceedings of European Conference on Computer Vision*, 740–755 (2014).

[5] Selvaraju, R. R. *et al.* Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE International Conference on Computer Vision*, 2591–2600 (2019).

[6] Hooker, S., Erhan, D., Kindermans, P. & Kim, B. A benchmark for interpretability methods in deep neural networks. In *Proceedings of Advances in Neural Information Processing Systems*, 9734–9745 (2019).

[7] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proceedings of International Conference on Learning Representations* (2015).