This ICCV workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Analyzing the Behavior of Cauliflower Harvest-Readiness Models by Investigating Feature Relevances

Niklas Penzel[1], Jana Kierdorf[2], Ribana Roscher[2,3], and Joachim Denzler[1]

[1]Computer Vision Group, Friedrich Schiller University Jena
[2]Remote Sensing Group, Institute of Geodesy and Geoinformation, University of Bonn
[3]Data Science for Crop Systems, Institute of Bio- and Geosciences, IBG-2: Plant Sciences,
Forschungszentrum Jülich GmbH
{*niklas.penzel,joachim.denzler*}*@uni-jena.de*, {*jkierdorf,ribana.roscher*}*@uni-bonn.de*

Abstract

The performance of a machine learning model is characterized by its ability to accurately represent the input-output relationship and its behavior on unseen data. A prerequisite for high performance is that causal relationships of features with the model outcome are correctly represented. This work analyses the causal relationships by investigating the relevance of features in machine learning models using conditional independence tests. For this, an attribution method based on Pearl's causality framework is employed. Our presented approach analyzes two data-driven models designed for the harvest-readiness prediction of cauliflower plants: one base model and one model where the decision process is adjusted based on local explanations. Additionally, we propose a visualization technique inspired by Partial Dependence Plots to gain further insights into the model behavior. The experiments presented in this paper find that both models learn task-relevant features during fine-tuning when compared to the ImageNet pre-trained weights. However, both models differ in their feature relevance, specifically in whether they utilize the image recording date. The experiments further show that our approach is able to reveal that the adjusted model is able to reduce the trends for the observed biases. Furthermore, the adjusted model maintains the desired behavior for the semantically meaningful feature of cauliflower head diameter, predicting higher harvest-readiness scores for higher feature realizations, which is consistent with existing domain knowledge. The proposed investigation approach can be applied to other domain-specific tasks to aid practitioners in evaluating model choices.

1. Introduction

Essential components of digital agriculture are reliable and well-generalizing machine learning models. The performance heavily relies on the model's behavior and whether the model relates the input features to the output targets in a causally correct way. One relevant application in digital agriculture is the accurate estimation of plant growth [19], harvest ripeness [2], the amount of harvest [6, 18], or the date the crop is ready to be harvested [21]. Predicting the optimal time to harvest not only maximizes crop yield but also ensures the quality and nutritional value of the produce. To analyze the model behavior regarding the causal relationships between specific features and the model outcome, methods that estimate the relevance of input features can be employed.

Well-known approaches are attribution methods, which have had a recent boost in the field of explainable machine learning [32]. Especially saliency mapping methods, which determine which areas in images are important for the decision of a machine learning model, are now widely used in various application areas [4, 16, 17, 36, 39]. However, since explainable machine learning methods present properties of a machine learning model such as the learned decision process between the input and the output in a human-understandable way, they build on correlation rather than causation.

This problem is also observed in the analysis done by Karimi et al. [14]. They investigate the causal relationship between model decision explanations (E) and model predictions (Y). More specifically, using Reichenbach's common cause principle [29], they study the treatment effect on E and Y using hyperparameters as interventions. In other words, how do predictions and explanations vary when hyperparameters change? They find that the hyperparameters have a high direct impact on E without going through Y, meaning other influences dominate. This observation is more pronounced for models with higher performance, further motivating the need to explain model decisions based on causal principles instead of misleading correlations.

Following these observations, recent works started investigating models and generating explanations based on causal principles [27, 30, 31]. We follow this approach and investigate causal explanations for a typical application from digital agriculture, namely harvest-readiness prediction. To be specific, we utilize the GrowliFlower data set [15] containing images of cauliflower plants in different growing stages and two image classification models that can estimate the corresponding harvest-readiness [16]. For our model investigation based on causal principles, we use the methodology by Reimers et al. [31]. Their method builds on Pearl's causality framework [26] and describes a structural causal model that encompasses supervised learning. The authors employ this structural model to investigate questions of whether trained classifiers use pre-defined semantically meaningful features to generate their predictions. To answer such questions, they apply Reichenbach's common cause principle [29] as in [14] and test for the conditional independence (CI) between features and predictions given reference annotations.

Additionally to our investigation of cauliflower harvest-readiness models, we extend the method of [31] by giving intuition about the model behavior for different feature settings once a feature is indicated as being used. We do this by combining conditional dependence insights from [31] together with ideas from partial dependence plots (PDPs) [8, 10].

The main contributions of this paper are:

- Analysis of cauliflower harvest-readiness models based on a causal feature attribution method.
- Investigating model behavior on the constrained test distribution by combining ideas from partial dependence plots (PDPs) [8, 10] with the conditioning on reference annotations from [31].
- Verification and confirmation that the model adjustments described in [16] reduce bias contained in the capturing situations of the training data.

2. Feature Analysis Methodology

In this work, we present an investigation approach to analyze and explain the model behavior of models classifying the harvest-readiness of cauliflower plants. To broadly investigate whether meaningful and informative semantic features are used, we employ the method proposed in [31]. This method builds on the framework of causality by Pearl [26]. To be specific, it frames supervised learning as a structural causal model (SCM) before employing conditional independence (CI) testing to detect whether a trained classifier with model predictions $\hat{Y}$ uses semantically meaningful features X. An existing connection means that the trained model utilizes information contained in X.

We, follow [31] and condition on the reference annotations Y to alleviate confounding factors. Confounding factors are a critical issue since they could lead to falsely detecting a connection between X and $\hat{Y}$ through the latent process that generates the task-specific data. Hence, the question of whether the connection from X to $\hat{Y}$ exists can be answered by testing for the null hypothesis (H_0)

$$H_0 : X \perp\!\!\!\perp \hat{Y} | Y \tag{1}$$

using a CI test.

If H_0 is discarded, then the investigated model utilizes information contained in the semantically meaningful feature X. In the following, we detail the selected CI tests for our analysis. Furthermore, we describe how we extend the approach described above to gain insight into not only whether a feature is used but also how model behavior changes for different feature realizations.

2.1. Conditional Independence Test Selection

The performance of the feature attribution method developed by Reimers et al. [31] hinges on one crucial decision: the selection of suitable CI tests. Many such tests exist [22] based on varying characterizations of CI. Previous work suggests employing multiple different CI testing methods based on varied but equivalent characterizations of CI [27, 30]. We specifically follow [27] and select a test based on mutual information estimation together with two tests estimating different kernel-based measures.

However, theoretic work by Shah and Peters [37] shows that there are no CI tests that reliably work for arbitrary joint distributions, which can result in false positives. Additionally, we cannot make valid assertions for how our variables of interest X, $\hat{Y}$, and Y are jointly distributed. Hence, we rely on a selection of nonparametric tests.

CMIknn: The first CI test we select is Conditional Mutual Information by k-nearest neighbor estimators (CMIknn) [34] utilizing the fact that two variables X and Y are conditionally independent given a third set of conditioning variables Z if and only if the CMI is zero.

CMIknn introduces two separate hyperparameters: k_{CMI}, defining the CMI estimator, and k_{perm}, determining the neighborhood size for the local permutation scheme necessary to keep dependencies between X or Y and Z. We follow [27] and [34] and set k_{perm} to five, and use ten percent of the available data to estimate CMI, i.e., $k_{\text{CMI}} = 0.1 \cdot n$.

cHSIC: The Hilbert-Schmidt Independence Criterion (HSIC) [12] and its conditional version (cHSIC) [9] are kernel-based tests that measure the dependence between variables. The cHSIC test is performed as a shuffle significance test between features X, predictions $\hat{Y}$, and reference annotations Y. With $N = 1000$ repetitions, we estimate the null distribution H_0 and derive a corresponding p-value. If the p-value is significantly small (we use $p < 0.01$), the null hypothesis H_0 is discarded, and conditional dependence is assumed.

To enable the detection of non-linear relationships, the observations are mapped into an infinite-dimensional reproducing kernel Hilbert space (RKHS) using the kernel trick [24]. The test statistics for HSIC and cHSIC are the Hilbert Schmidt norms of the cross-covariance and the conditional cross-covariance operator, respectively. These test statistics depend on the selected kernels. We follow previous work [9, 27] and use Gaussian radial basis function (RBF) kernels. The optimal kernel widths σ can differ between our variables X, $\hat{Y}$, and Y. Hence, we use the heuristic by Gretton et al. [11] to heuristically determine fitting kernel widths.

RCoT: The cHSIC test has known issues, such as the unknown null distribution of the test statistic and computationally expensive approximations. To address the first issue, Zhang et al. [40] proposed the Kernel Conditional Independence Test (KCIT) as an alternative. KCIT is based on the CI characterization by Daudin [7] and tests whether correlations between residual functions in an RKHS vanish.

Furthermore, Strobl et al. [38] proposed the Randomized Conditional Correlation Test (RCoT), which approximates KCIT in a computationally efficient way using random Fourier features [28]. Despite its name, it is a CI test and recommended over the related RCIT variant [38].

Again we use the heuristic by Gretton et al. [11] to determine the hyperparameters for the necessary RBF kernels. We generally employ the settings used in [27] but crucially follow the suggestion in [38] and use a slower shuffle significance test instead of estimating the null distribution directly with the Lindsay-Pilla-Basak method [23]. This is recommended for sample sizes less than 500, as is the case in our analysis.

2.2. Approximating Feature Influence

To investigate the influence of specific features on the model decisions, we utilize insights from partial dependence plots (PDPs) [8, 10], i.e., visualizing observed variables with respect to changing feature values. Specifically, we visualize the model behavior for the feature realizations contained in our test set. We do this by plotting the model output against the observed feature values. Furthermore, we follow the key insight from [31] and separate these visualizations according to the reference classes, i.e., condition on Y, to reduce confounding.

For categorical features, we then estimate Gaussian distributions and plot the means and standard deviation per category to gain insights into the general trend of the model behavior as well as corresponding uncertainties. For continuous features, we calculate the standard deviation of the feature values in our test data to estimate the distribution parameters in a sliding window approach resulting in similar plots.

The generated visualizations, extend the result for a specific feature beyond the binary decision of whether the feature is used or not used and enable investigating how the model behavior changes for specific feature realizations. In contrast, the testing methodology from [31] relies on the temporal order of variables in the SCM to circumvent the necessity for interventions to generate explanations. Hence, [31] does not produce counterfactual explanations that tell us how the model's predictions change when the feature value X changes.

However, the intuitions we gain from our visualizations, are limited. While we can generally get an idea of how the model behaves for a change in feature values as they appear in our test set, this is not a causal counterfactual explanation in the sense of Pearl's [26] do-calculus. Nevertheless, they provide an additional way to analyze model behavior under the constraints posed by a given test set.

3. Cauliflower Harvest-readiness Classification

3.1. Models

We analyze two, previously proposed, models, that determine the harvest-readiness of cauliflower from drone images [16]. Additionally, we compare our results to a pretrained model without task-specific finetuning. The underlying architecture is a binary ResNet18 classifier [13] that predicts `Ready` and `Not Ready` for harvest (72.41% accuracy). For further details about the hyperparameters, we refer to [16]. The authors expand their classification framework by calculating saliency maps of the input images, which they cluster using Spectral Clustering (SC) [25] to derive a reliability score for the predictions based on the cluster assignments. They use these scores to improve the predictions of their model by swapping class assignments of unreliable predictions (improves to 88.14% accuracy).

The selected models are as follows: First, the base ResNet18 model M_{Base} fine-tuned on cauliflower data without prediction adjustments; second, the adjusted model M_{Adjusted} derived from M_{Base} using the combination of saliency mapping and SC [16], and finally, M_{ImageNet}, i.e., a ResNet18 with ImageNet [35] pre-trained weights. We select the last model to ensure that features detected for M_{Base} and M_{Adjusted} are learned during the fine-tuning step on the

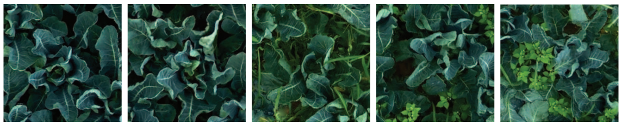


Figure 1: Examples of cauliflower plants in the GrowliFlowerR dataset [15]. Note the varying amount of weeds contained in the images.

cauliflower data.

3.2. Cauliflower Data

To train our models, we use images and in-situ measurements, called reference data, contained in the GrowliFlowerR dataset [15]. This dataset is composed of image timeseries displaying cauliflower plants over their entire growth. From the given reference data, we additionally use the *recording date*, the plant *position*, the *plant diameter*, and the *head diameter*.

We select images of the time-points 2021/08/23, 2021/08/25, 2021/08/30, 2021/09/03 about which information of the harvest-readiness within the next three days is known. Example images are shown in Fig. 1. We split the data into a training, validation, and test set. Standard augmentations like rotation and flipping are applied to the training set, resulting in 6244 images used for training, 196 images used for validation, and 194 images used for testing. All images contain different amounts of weeds.

In the following, we detail the features that we deem interesting for analyzing the behavior of the harvest-readiness models. To structure it, we categorize the features into coarse groups: Capturing circumstances, plant location, and actual semantic content, i.e., cauliflower properties and the amount of visible weed.

3.2.1 Capturing Circumstances

To analyze the influence of some capturing circumstances, we first select the *recording date* of the images as a feature. Additionally, we try to encapsulate some information about the exposure and, therefore, the weather during the capturing process. For this, we additionally analyze the *average image brightness* by transforming the image into a grayscale representation using [3]. We then calculate the average brightness as the mean of all pixel intensity values.

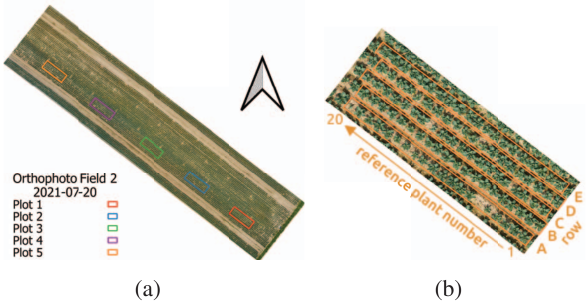
Fig. 5 in Appendix A.1 shows that the *recording date* and the *average brightness* features are related. The observed marginal distributions for each recording date clearly differ in their respective means. In other words, both of these features are not independent and share some information.

To further investigate this observation, we determine the mutual information (MI) between the *recording date* and the *average image brightness*. We select the estimator described in [33] as it was specifically developed to approximate the MI between two variables, where one is discrete (*recording date*), and one is continuous (*average image brightness*). We set $k = 3$ following the suggestions in [20, 33]. Using this setup, we estimate the MI to be ≈ 0.763.

A likely explanation is that the weather is a confounding factor for both features. Images that were taken during sunny weather (see Fig. 3, last row) result in a higher *average brightness*. Additionally, the *recording date* encodes some information about the weather during the recording process. Nevertheless, both features also encode disjoint information. The *recording date* implicitly encapsulates information about the growth status of the plants, while the *average brightness* encodes more weather information.

3.2.2 Plant Position

The plant position is given by three variables: the *plot number* as well as the specific *row* and *column* within the plot (Fig. 2). The plots are distributed along the whole field, meaning that plants in different plots may show different stages of development.



(a) (b)

Figure 2: Overview of (a) the distribution of reference plots in the field from the GrowliFlowerR dataset and (b) the plant positions, which are indicated by their row (A-E) and column (1-20). The source of the figures is [15].

3.2.3 Semantic Image Content

One set of features that is especially interesting for domain experts is semantic image content, e.g., properties of the individual plants. Here, we first describe the cauliflower properties annotated in the dataset we used. Afterward, we detail how we approximate the amount of weed in the images without relying on reference annotations.

Cauliflower Properties. Essential features that describe the development status of a plant are the *plant diameter* and *head diameter*. The plant diameter is easy to determine in the field and in images at earlier stages of development. As soon as neighboring plants start to overlap, the determination of the diameter from images is more complex because the boundaries of different plants have to be defined first. Depending on the cauliflower cultivar, the plants are more or less self-covering. Hence, the head is not visible from the outside, and its own leaves cover it to protect it from abiotic and biotic stresses like sunlight or animals. The size of the cauliflower head is the indicator of whether a cauliflower plant is ready for harvest. However, the *plant diameter* is not correlated, i.e., it is impossible to derive the head size from only the size of the plant.

Weeds Ratio. A visible difference in our data is the amount of weed growing next to the cauliflower plants. We are interested in whether the selected cauliflower models change their decisions based on the amount of visible weed. Toward this goal, we use simple linear iterative clustering (SLIC) [1] and a merge algorithm to segment the images in an unsupervised fashion. Hence, quantifying the relationship between cauliflower plants and weeds in our images.

We detail this approach in Appendix A.2. Fig. 3 displays some examples of the superpixel categorization for three images containing varying amounts of weed. The final extracted feature is the ratio between the amounts of weed and ground pixels divided by the number of cauliflower pixels. We term this feature *weeds ratio* and observe mostly values between zero and one, i.e., most pixels contain cauliflower.

However, note that our unsupervised approach makes mistakes, e.g., the first row in Fig. 3. This observed behavior has to do with the illumination and smaller fluctuations in color. These mistakes decrease the signal-to-noise ratio of the actual semantic feature and our measured feature values. Nevertheless, a visual examination of our test and validation images reveals that our approach generally works well, disregarding these smaller errors. Further, other features, especially metadata features, are also proxy features for some latent properties meaning they contain a similar amount of noise.

4. Feature Analysis Results

Table 1 summarizes the results for our three selected models for all eight features of interest described in Sec. 3.2. The first observation is that M_{ImageNet} uses none of our chosen features. This behavior is expected given the different problem domains. Hence, this result confirms that the detection of features for the other two models is learned during the tuning step.

No model uses any features encoding the plant location, which is desirable. Given the close proximity of the plants in one cauliflower field, a difference in prediction procedure would likely indicate a bias in our data or training process.

For the content-related features, we observe that both the M_{Base} and the M_{Adjusted} learn to utilize the *head diameter*. This result reaffirms that both models actually learn the important feature of the task at hand. The *head diameter* is the latent feature also used by cauliflower farmers to determine whether a cauliflower is `Ready` for harvest. It is worth noting that although the head is not visible in the image (see Sec. 3.2), the network generates a representation to infer this important feature. Furthermore, both models disregard the *plant diameter*, which is congruent with existing domain knowledge. Additionally, we observe that both models do not change their behavior depending on the number of visible weeds, i.e., the *weeds ratio* is not used.

Finally, the models differ in behavior for the two features related to the capturing circumstances. While both models use some information contained in the *average image brightness*, only the base model additionally uses the *recording date*. Given the relationship between these features (see Appendix A.1), we hypothesize that this behavior could be a result of the signal-to-noise ratio in the *recording date* feature. The adjusted model does not significantly change its predictions for different dates. We analyze this further in the following Sec. 4.1, where we investigate how the model behaviors differ for different realizations of the indicated features in our test set. However, we first state our expectations.

Expectations for Task-Relevant Features Our first expectation is later *recording dates* and higher *average brightness* lead to higher predictions of harvest-readiness for M_{Base} while the trend is less significant for M_{Adjusted}. We hold this expectation because class `Ready` for harvest and sunny weather are more likely for later dates but are, of course, not causally related. Furthermore, we expect a well-performing model to learn the relationship between the feature *head diameter* and the correct class. In other words, given that both M_{Base} and M_{Adjusted} use this feature significantly, according to Table1, we expect an upward trend. Finally, we expect that the predicted scores by M_{ImageNet} are nearly constant for different feature realizations.

4.1. Influence of Task-Relevant Features

To further analyze the features indicated in Table 1, we visualize the relationship between the model output and the respective feature realizations for all images in our test set. As described in Sec. 2.2, to reduce confounding, we follow the approach described in [31] and condition on the reference annotations.

Fig. 4a visualizes the *recording date* against the model outputs. We observe that predictions of M_{ImageNet} are, on average, almost constant for all recording dates. However, we see a difference between the base and the adjusted model. Both display an upwards trend, i.e., later recording dates indicate higher output scores for images of both classes. Nevertheless, looking at the standard deviations, we see that this

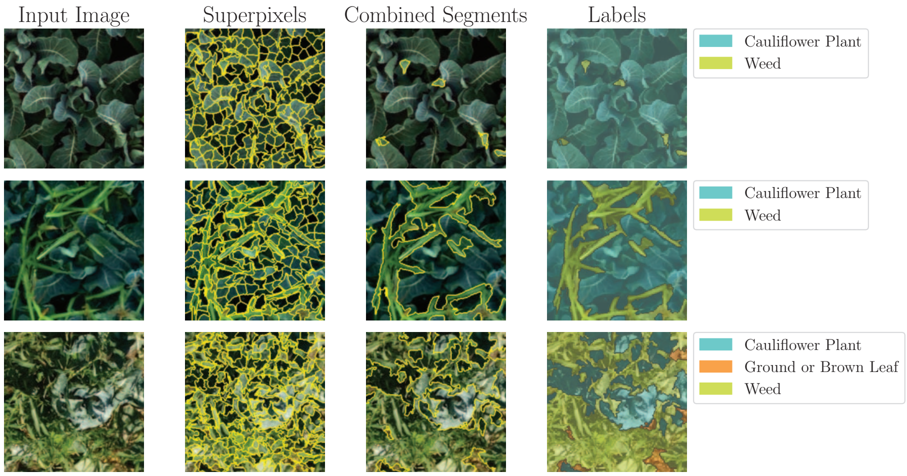


Figure 3: Examples of our unsupervised weeds segmentation into superpixels. The rows contain images with an increasing amount of weed. Note that the last row contains a sample taken under bright sunshine, leading to increased exposure.

Table 1: Feature relevance of the selected cauliflower models.

	Recording	Average	Position			Head	Plant	Weeds
Model	Date	Brightness	Plot	Col	Row	Diameter	Diameter	Ratio
M_{Base}	✓	✓	✗	✗	✗	✓	✗	✗
M_{Adjusted}	✗	✓	✗	✗	✗	✓	✗	✗
M_{ImageNet}	✗	✗	✗	✗	✗	✗	✗	✗

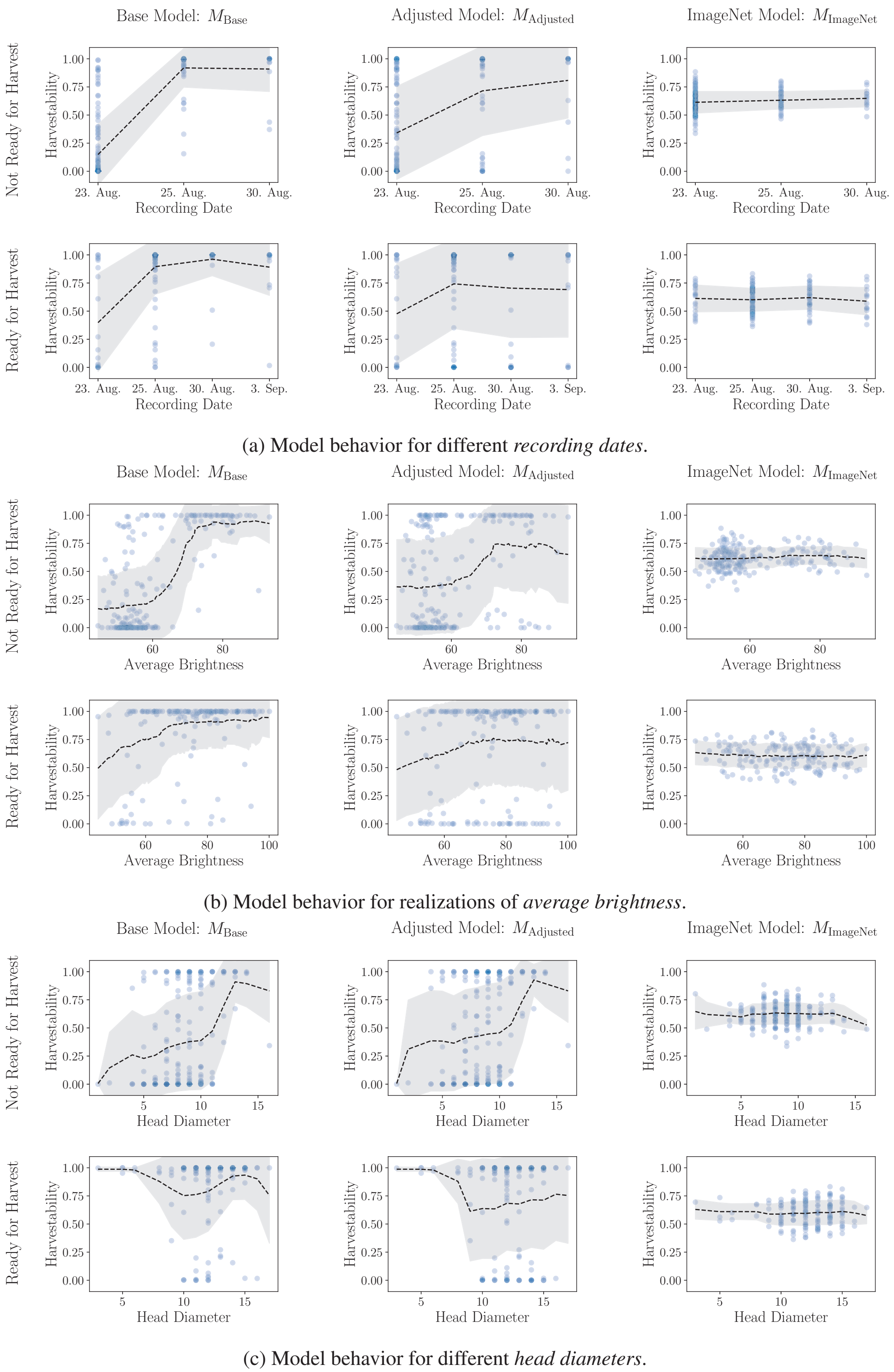
trend is more pronounced for M_{Base}. Further, Table 1 tells us that the usage by M_{Adjusted} is not significant.

However, Table 1 indicates utilization of the *average brightness* for both M_{Base} and M_{Adjusted}, while the M_{ImageNet} does not change behavior based on this feature. This observation is reflected in Fig. 4b. Note, there is only a slight upward trend visible for images of the class `Ready` to harvest. However, for the other class, we observe a clear change for an *average brightness* above an intensity of $\sim$ 70. Here, M_{Base} and M_{Adjusted} both predict higher harvest-readiness scores. In other words, if the unseen images are, on average, brighter, the models predict ready-to-harvest with a higher probability. We hypothesize that this could be a bias in the training data, given the weather conditions during the recording. This is supported by the slight upward trend for the *recording date* discussed above and the relationship between the recording date and average image brightness discussed in Appendix A.1. Nevertheless, even though M_{Adjusted} utilizes information contained in the average brightness and changes behavior accordingly, our visualization reveals that this change is less pronounced compared to M_{Base}. Hence, the adjustments made to the base model reduce the biases encoded in the capturing circumstances.

Expert knowledge of cauliflower farmers indicates that the *head diameter* encodes the information necessary to decide whether cauliflower is ready for harvest. Table 1 already indicated that both M_{Base} and M_{Adjusted} learn this relationship. Fig. 4c now visualizes the model behavior for specific realizations of this feature. Both models behave similarly. For images of the class `Ready`, Fig. 4c indicates an increasing uncertainty for larger head diameters. This observation could be a consequence of little data in this class with smaller feature values. However, interesting is the clear upward trend for images of the class `Not Ready`. This result indicates that the model implicitly extracts information related to the *head diameter* for unseen images and predicts higher harvest-readiness scores for larger diameters. A cauliflower harvest-readiness model that learns the causal relationship between the inputs and desired outputs should exhibit this behavior.

5. Conclusions

In this work, we analyze three models on a feature level where two models are specialized to predict the harvest-readiness of cauliflower plants. Toward this goal, we utilize a feature attribution method [31] built on the foundation of Pearl's causality framework [26]. We extend this method



(a) Model behavior for different *recording dates*.

(b) Model behavior for realizations of *average brightness*.

(c) Model behavior for different *head diameters*.

Figure 4: Model behaviors when presented with our test data. We plot the features indicated as being used in Table 1 against the predicted harvest-readiness score. We follow [31] and split our visualization between the two classes, `Ready` to harvest and `Not Ready` to harvest.

over the binary indication of whether a feature is used by visualizing and regressing the model outputs against feature realizations in our test set. This approach is inspired by PDPs [8, 10] and lets us investigate the model behavior for different feature values on the constrained test scenario.

We find that both analyzed cauliflower models learn task-specific features during the fine-tuning process and improve over a pre-trained ImageNet model on a feature level. For the base model, three features are indicated as being used: the *recording date*, the *average image brightness*, and the *head diameter*. In contrast, the adjusted model only utilizes the latter two to a significant level. Using our proposed visualization, we investigated this difference and found that the trends for both the *recording date* (to a nonsignificant level) and the *average image brightness* are reduced for the adjusted model. We conclude from this observation that the adjustments made to the base model [16] reduce the corresponding bias observed in the base model. Furthermore, the adjusted model keeps the observed behavior for the semantically meaningful feature: *head diameter*, by predicting, on average higher scores for higher feature realizations. This behavior is consistent with existing domain knowledge. Hence, our approach enables users to evaluate and compare competing models in terms of causal feature usage leading to increased robustness towards unseen data.

Acknowledgment

This project was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 491111487 and (RO 4839/7-1 | STO 1087/2-1) and in part by the DFG under Germany's Excellence Strategy – EXC 2070 – 390732324.

A. Appendix

A.1. Recording Date and Average Brightness

In Sec. 3.2.1, we describe our features of interest. Amongst these features are the recording date and the average image brightness. Fig. 5 shows that these two features are not independent from one another.

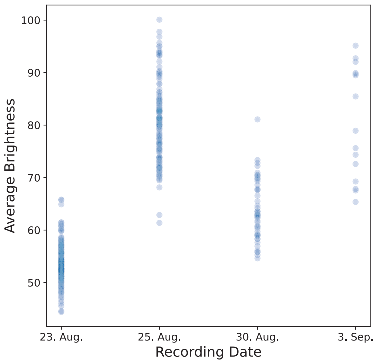


Figure 5: Relationship between recording date and average image illumination. We can clearly see the influence of the date on the distribution of the average image brightness.

A.2. Segmenting Cauliflower and Weeds

As briefly described in Sec. 3.2.3, we combine our detected superpixels into four broader sets of image regions. However, the first step is to separate the image into superpixels. For this, SLIC [1] uses linear clustering in a five-dimensional space. This space consists of the three color channels of the CIELAB color space combined with the x and y pixel coordinates. To ensure color similarity and pixel proximity with a distance measure that incorporates superpixel size. We use the Scikit-Learn [5] implementation and set the approximate number of segments to 250. For other hyperparameter settings, we follow the standard parameterization in [5]. Fig. 3 contains examples of the generated superpixels for some of our test images.

The second step now is to use superpixel statistics to categorize them into larger groups. An initial visual inspection of our validation data revealed three key components in our cauliflower images: the cauliflower plants, weeds, and the ground. Additionally, given the uniformity of our images, we observed that color information is enough for a coarse segmentation. The key idea is the turquoise tint of the cauliflower superpixels in comparison to the green color of the weeds. Additionally, we sometimes observe brown color for ground or brown leaves. Hence, we combine superpixels according to their mean colors. We give detailed instructions as pseudo-code in Algorithm 1. In Algorithm 1, we set the hyperparameter λ to 0.11 for our test data.

Algorithm 1 An algorithm to combining superpixels.

```
Require: list of superpixels $S$, cauliflower tolerance $\lambda$
Ensure: $cauli \leftarrow []$, $weed \leftarrow []$, $ground \leftarrow []$, and, $other \leftarrow []$
for $s$ in $S$ do
    $(r, g, b) \leftarrow \texttt{mean}(s)$    ▷ Average color in segment
    if $|g - b| < \lambda$ then    ▷ similar green and blue
        $cauli \leftarrow cauli + s$
    else if $g > r$ and $g > b$ then    ▷ Green dominant channel for weed
        $weed \leftarrow weed + s$
    else if $r > b$ then    ▷ Red dominant over blue
        $ground \leftarrow ground + s$
    else    ▷ Rejections
        $other \leftarrow other + s$
    end if
end for
return $cauli, weed, ground, other$
```

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012. 5, 8

[2] B. Arad, J. Balendonck, R. Barth, O. Ben-Shahar, Y. Edan, T. Hellström, J. Hemming, P. Kurtser, O. Ringdahl, T. Tielen, et al. Development of a sweet pepper harvesting robot. *J. Field Robot.*, 37(6):1027–1039, 2020. 1

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's J. Softw. Tools*, 2000. 4

[4] 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*, pages 93–117. Springer, 2018. 1

[5] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G.Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013. 8

[6] Y. Cai, K. Guan, D. Lobell, A. B. Potgieter, S. Wang, J. Peng, T. Xu, S. Asseng, Y. Zhang, L. You, et al. Integrating satellite and climate data to predict wheat yield in australia using machine learning approaches. *Agric. For. Meteorol.*, 274:144–159, 2019. 1

[7] J.J. Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980. 3

[8] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. of Stat.*, 29:1189–1232, 2001. 2, 3, 8

[9] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. *Adv. Neural. Inf. Process. Syst.*, 20, 2007. 3

[10] B.M. Greenwell, B.C. Boehmke, and A.J. McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018. 2, 3, 8

[11] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. *Adv. Neural. Inf. Process. Syst.*, 19, 2006. 3

[12] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, A.J. Smola, et al. A kernel statistical test of independence. In *Nips*, volume 20, pages 585–592. Citeseer, 2007. 3

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, 2016. 3

[14] A.-H. Karimi, K. Muandet, S. Kornblith, B. Schölkopf, and B. Kim. On the relationship between explanation and prediction: A causal view. In *40th Intern. Conf. on Machine Learning (ICML)*, July 2023. 1, 2

[15] J. Kierdorf, L.V. Junker-Frohn, M. Delaney, M.D. Olave, A. Burkart, H. Jaenicke, O. Muller, U. Rascher, and R. Roscher. Growliflower: An image time-series dataset for growth analysis of cauliflower. *J. Field Robot.*, 40(2):173–192, 2023. 2, 4

[16] J. Kierdorf and R. Roscher. Reliability scores from saliency map clusters for improved image-based harvest-readiness prediction in cauliflower. *IEEE Geosci. Remote.*, 20:1–5, 2023. 1, 2, 3, 8

[17] T. Kim, H. Kim, K. Baik, and Y. Choi. Instance-aware plant disease detection by utilizing saliency map and self-supervised pre-training. *Agriculture*, 12(8), 2022. 1

[18] T.V. Klompenburg, A. Kassahun, and C. Catal. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.*, 177:105709, 2020. 1

[19] A. Kocian, S. Massa, S. Cannazzaro, L. Incrocci, S. Di Lonardo, P. Milazzo, and S. Chessa. Dynamic bayesian network for crop growth prediction in greenhouses. *Comput. Electron. Agric.*, 169:105167, 2020. 1

[20] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):066138, 2004. 4

[21] D.J. Lary, A.H. Alavi, A.H. Gandomi, and A.L. Walker. Machine learning in geosciences and remote sensing. *Geosci. Front.*, 7(1):3–10, 2016. 1

[22] C. Li and X. Fan. On nonparametric conditional independence tests for continuous variables. *Wiley Interdiscip. Rev. Comput. Stat.*, 12(3):e1489, 2020. 2

[23] B.G. Lindsay, R.S. Pilla, and P. Basak. Moment-based approximations of distributions using mixtures: Theory and applications. *Ann. Inst. Stat. Math*, 52:215–230, 2000. 3

[24] J Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London*, 209:415–446, 1909. 3

[25] A.Y. Ng, M.I. Jordan Y., and Weiss. On spectral clustering: Analysis and an algorithm. In *Adv. Neural. Inf. Process. Syst.*, pages 849–856, 2002. 3

[26] Judea Pearl. *Causality*. Cambridge university press, 2009. 2, 3, 6

[27] N. Penzel, C. Reimers, P. Bodesheim, and J. Denzler. Investigating neural network training on a feature level using conditional independence. In *ECCV Workshop on Causality in Vision (ECCV-WS)*, pages 383–399, Cham, 2022. Springer Nature Switzerland. 2, 3

[28] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, 2007. 3

[29] H. Reichenbach. *The direction of time*. University of California Press, 1956. 1, 2

[30] C. Reimers, N. Penzel, P. Bodesheim, J. Runge, and J. Denzler. Conditional dependence tests reveal the usage of abcd rule features and bias variables in automatic skin lesion classification. In *CVPR ISIC Skin Image Analysis Workshop (CVPR-WS)*, pages 1810–1819, June 2021. 2

[31] C. Reimers, J. Runge, and J. Denzler. Determining the relevance of features for deep neural networks. In *ECCV*, pages 330–346. Springer, 2020. 2, 3, 5, 6, 7

[32] 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. Spat. Inf. Sci.*, 3:817–824, 2020. 1

[33] B.C. Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014. 4

[34] J. Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Int. Conf. Robot. Artif. Intell.* PMLR, 2018. 2

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115:211–252, 2015. 3

[36] P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.G. Luigs, A.K. Mahlein, and K. Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.*, 2(8):476–486, 2020. 1

[37] R.D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *Ann. Stat.*, 48(3):1514–1538, 2020. 2

[38] E.V. Strobl, K. Zhang, and S. Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *J. Causal Inference*, 2019. 3

[39] L. Weber, S. Lapuschkin, A. Binder, and W. Samek. Beyond explaining: Opportunities and challenges of xai-based model improvement. *Inf. Fusion*, 2022. 1

[40] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012. 3