Received: 5 June 2020 | Accepted: 26 March 2021

DOI: 10.1002/ppj2.20020

The Plant Phenome Journal OPEN ACCESS

ORIGINAL RESEARCH

How useful is active learning for image-based plant phenotyping?

Koushik Nagasubramanian[1] | Talukder Jubery[2] | Fateme Fotouhi Ardakani[2] | Seyed Vahid Mirnezami[2] | Asheesh K Singh[3] | Arti Singh[3] | Soumik Sarkar[2] | Baskar Ganapathysubramanian[2]

[1] Department of Electrical Engineering, Iowa State University, Ames, IA, USA

[2] Department of Mechanical Engineering, Iowa State University, Ames, IA, USA

[3] Department of Agronomy, Iowa State University, Ames, IA, USA

Correspondence
Baskar Ganapathysubramanian and Soumik Sarkar, Department of Mechanical Engineering, Iowa State University, Ames, Iowa, USA.
Email: baskarg@iastate.edu; soumiks@iastate.edu

Assigned to Associate Editor Michael Gore.

Abstract

Deep learning models have been successfully deployed for a diverse array of image-based plant phenotyping applications including disease detection and classification. However, successful deployment of supervised deep learning models requires large amount of labeled data, which is a significant challenge in plant sciences (and most biological) domain due to the inherent complexities. Specifically, data annotation is costly, laborious, time consuming and needs domain expertise for phenotyping tasks, especially for diseases. To overcome this challenge, active learning algorithms have been proposed to reduce the amount of labeling needed by deep learning models to achieve good predictive performance. Active learning methods work by adaptively suggesting samples to annotate using an acquisition function to achieve maximum (classification) performance under a fixed labeling budget. We report the performance of four different active learning methods, (1) Deep Bayesian Active Learning (DBAL), (2) Entropy, (3) Least Confidence, and (4) core-set, with conventional random sampling-based annotation for two vastly different image-based classification datasets. The first image dataset consists of soybean [*Glycine max* L. (Merr.)] leaves belonging to eight different soybean stresses and a healthy class, and the second consists of nine different weed species from the field. For a fixed labeling budget, we observed that the classification performance of deep learning models using active learning based acquisition strategies is better than random sampling-based acquisition for both datasets. The integration of active learning strategies for data annotation can help mitigate labelling challenges in the plant sciences applications particularly where resources dedicated to annotations are limited.

1 | INTRODUCTION

With the advent of high throughput phenotyping in plant sciences (Araus et al., 2018; Singh et al., 2016; Singh et al., 2020; Singh et al., 2021), we are now able to collect copious amounts of image data. Deep learning (DL) architectures have advanced the state-of-the-art performance for image-based classification tasks (Krizhevsky et al., 2012), and have been successfully deployed for a diverse array of image-based plant phenotyping applications including disease detection, classification, and quantification (Singh et al., 2018).

Abbreviations: AL, active learning; BALD, Bayesian Active Learning by Disagreement; CNN, convolutional neural networks; DBAL, Deep Bayesian Active Learning; DL, deep learning; LC, Least Confidence; ML, machine learning.

However, one of the critical drawbacks of DL models is its necessity to have a large amount of labeled data to achieve good model accuracy. This is especially true for plant science applications, where annotating data can be costly, laborious, and time consuming to obtain, and generally need domain expertise (for instance, for plant disease image labeling that requires trained plant pathologists). To overcome this drawback, one effective and practical strategy is to use Active Learning (AL) based image annotation (Cohn et al., 1996). Weak supervision (Ghosal et al., 2019), synthetic dataset creation (Valerio Giuffrida & Scharr Forschungszentrum Jülich, 2017), and transfer learning (Tapas, 2016) are some of the other methods available to reduce the amount of labeling needed. However, when large amounts of unlabeled data are available, but the task of labeling is hard or infeasible, AL methods are very useful. Active learning methods adaptively select the most informative samples for labeling for the highest improvement in test accuracy. The goal of AL is to achieve maximum predictive performance under a fixed labeling budget, which makes it desirable for plant science applications.

Many AL methods have been proposed with different heuristics (Settles, 2009) to reduce the amount of labeling needed for training machine learning (ML) models for classification tasks. A small amount of data is randomly chosen initially for labeling; this labeled dataset is used to train a neural network model. Then, a batch of data from the remaining unlabeled data set is adaptively selected using an *acquisition function* for labeling by human domain experts. The *acquisition function* serves to select the most useful samples in the unlabeled dataset for improving neural network model performance. This process of choosing limited samples from unlabeled data sets, having the human expert annotate/label these limited samples, adding them to the labeled set, and retraining the model continues until one of two termination criteria is met – a desired performance threshold of the model is achieved, or the labeling budget is exhausted.

Recently, in non-plant sciences problems, AL methods have been successfully applied for improving the performance of DL models, for example, DL-based image classification (Wang et al., 2016), biomedical image segmentation (Yang et al., 2017), text classification (Zhang, Lease, & Wallace, 2017), and object detection (Kao et al., 2018). In the field of plant phenotyping, uncertainty-based sampling method was used to select samples for training a Faster R-CNN model for panicle detection in cereal crops (Chandra et al., 2020).

The continual improvement of AL strategies in the ML community can be leveraged to significantly augment plant phenotyping efforts through state-of-the-art AL techniques. As a first step, there is a need to perform a comparative evaluation of the available sophisticated AL strategies in the context of canonical plant phenotyping applications. We compare four active learning methods defined by different *acquisition functions*: ***least confidence*** (Culotta & McCallum, 2005)***, entropy*** (Shannon, 1948)***, Deep Bayesian Active Learning*** (Gal et al., 2017), and ***core-set*** (Sener & Savarese, 2017) on two disparate plant phenotyping problems – soybean stress identification (Ghosal et al., 2018) and weed species classification (Olsen et al., 2019).

Core Ideas

- Active learning methods reduce the amount of expert annotation needed in challenging image-based plant classification tasks.
- Most *acquisition functions* built on uncertainty-based sampling perform better than simple random sampling.
- However, random sampling is a good baseline for easy (for example, images under constant illumination conditions, i.e., less noisy data) classification tasks.

2 | MATERIALS AND METHODS

2.1 | Datasets

2.1.1 | Soybean stress dataset

The dataset consists of 16,573 RGB images of soybean [*Glycine max* L. (Merr.)] leaves across nine different classes (i.e., eight different soybean stresses, and the ninth class containing healthy soybean leaf). Details on the dataset can be found in (Ghosal et al., 2018). Briefly, these classes cover a diverse spectrum of biotic and abiotic foliar stresses in soybean. Figure 1 illustrates the nine different soybean leaf classes used in this study. The entire data set of 16,573 images consisted of bacterial blight (caused by *Pseudomonas syringae* pv. *glycinea*; number of images = 1,524), Septoria brown spot (caused by *Septoria glycines*; number of images = 1,358), Frogeye leaf spot (caused by *Cercospora sojina*; number of images = 1,122), Healthy (number of images = 4,223), Herbicide injury (number of images = 1,395), iron deficiency chlorosis (number of images = 1844), Potassium deficiency (number of images = 2,186), bacterial pustule (caused by *Xanthomonas axonopodis* pv. *glycines*; number of images = 1,674), and sudden death syndrome (caused by *Fusarium virguliforme*; number of images = 1,247).

2.1.2 | Weed species dataset

The data set consists of 17,509 RGB images of weed species across nine different classes (eight weed classes and one non-weed class). Figure 2 illustrates the nine

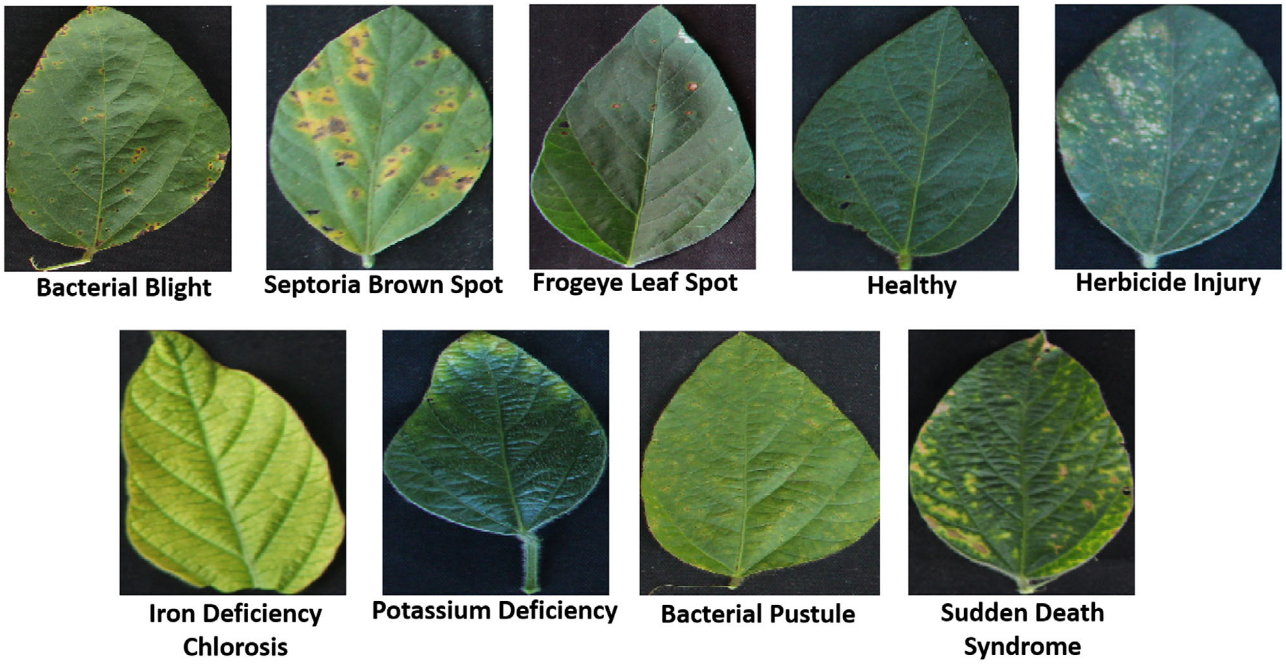

FIGURE 1 The nine classes of data (eight stress, and one healthy) collected on soybean leaflets, which comprised the first data set

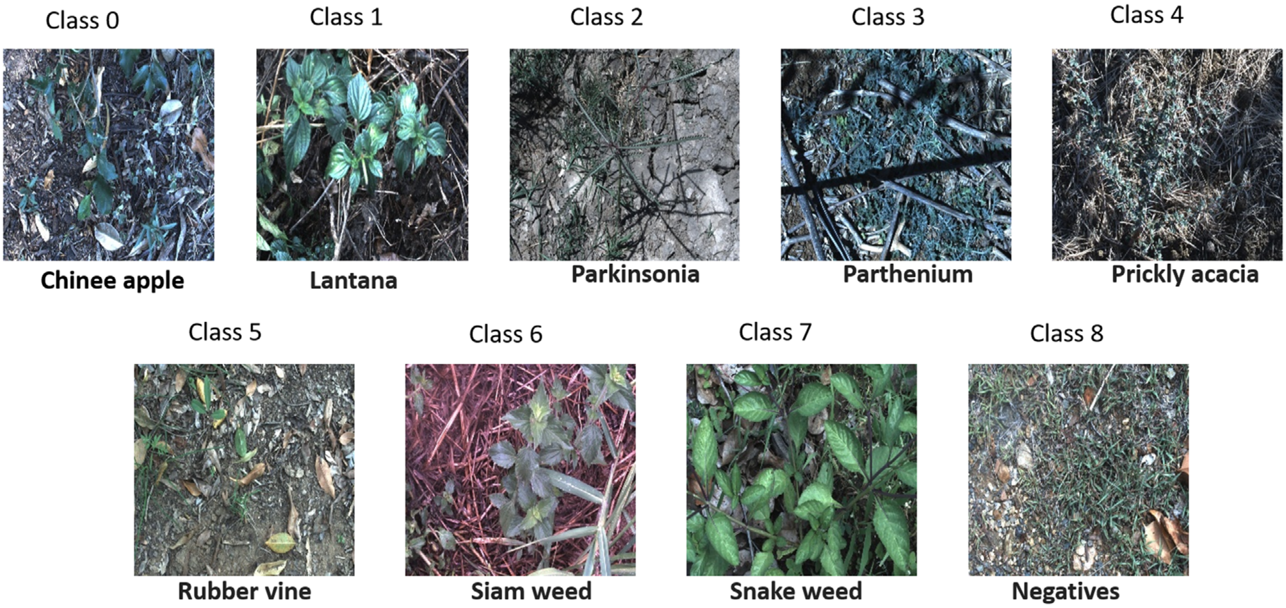

FIGURE 2 The nine classes of second data set consisting of eight weed species, and one weed free class (labelled as Negatives). Images taken from publicly available dataset associated with (Olsen et al., 2019)

different classes used in this study, and the full description can be found in Olsen et al., 2019). The entire data set of 17,509 images consisted of Chinee apple (*Ziziphus mauritiana*; number of images = 1,125), *Lantana camara* (number of images = 1,064), *Parkinsonia aculeata* (number of images = 1,031), *Parthenium hysterophorus* (number of images = 1,022), Prickly acacia (*Acacia nilotica*; number of images = 1,062), Rubber vine (*Cryptostegia grandiflora*; number of images = 1,009), Siam weed (*Chromolaena odorata*; number of images = 1,074), Snake weed (*Stachytarpheta*; number of images = 1,016), and weed free (number of images = 9,106).

2.2 | Experimental setup

We trained a neural network-based classification model for identifying the class labels for input images in the two data sets, with the goal to achieve ***maximum classification performance for a fixed labeling budget***. We evaluated each of the four active learning strategies (core-set, DBAL, entropy, and least confidence) based on how well the neural network performed – i.e., using the classification accuracy on the complete dataset. We used MobileNetV2 (Sandler et al., 2018) architecture for data set #1 (soybean stress classification) and ResNet-50 (He et al., 2016) architecture for dataset #2 (weed species classification). These networks were specifically chosen because of their well-documented popularity and consistent performance, as well as to test the capability of AL on two distinct and well-used networks. MobileNetV2 is a smaller, more compact network, while ResNet-50 is a large network, and were appropriate for dataset #1 (controlled condition imaging) and data set #2 (field-based imaging), respectively.

Each data set was analyzed separately. The AL approach was repeated 10 times for each dataset. While 10 runs are excessive due to the time taken for execution; however, they provide statistical robustness in comparison metrics that is useful for other practitioners. We randomly selected and labeled 5% of the samples from the complete data to create a fixed size validation set before starting the active learning experiment (ideally, this initial random sampling should be well-balanced in terms of samples per class. However, in practice, this is tough to guarantee, as no label information is available initially). The validation dataset (829 images for soybean dataset, and 876 images for the weed dataset) remained the same for the different labeling budgets and was kept fixed for all the 10 repetitions of the experiment. Each run starts with an initial random batch of 1,000 samples spread across different classes, which was used for the evaluation of all four active learning methods. After training the neural network model for 100 epochs, we used the best performing model on the validation datasets to query a batch of 1,000 samples from the remaining unlabeled dataset. This selection was performed using the *acquisition function* of each of the four active learning algorithms (so each AL approach will potentially select distinct set of 1,000 samples to next annotate). These 1,000 samples were added to the labeled dataset, to retrain the neural network model. This process was repeated until the labeling budget was exhausted (labeling budget was 9,000 samples for the soybean stress classification, and 10,000 samples for the weed species classification). We saved the model with best validation accuracy. The model was retrained from scratch after every selection of new labeled samples for 100 epochs with a batch size of 16. We used dropout for regularization with probability value of 0.001 for the MobileNetV2 model. A dropout layer with probability 0.5 was added after the last fully connected layer for the ResNet-50 model. We optimized the model using the Adam (Kingma & Ba, 2014) optimizer with the default learning rate of 0.001. We used Keras (Chollet, 2015) with a Tensorflow (Abadi et al., 2016) backend for the implementation. A schematic of the approach is shown in Figure 3.

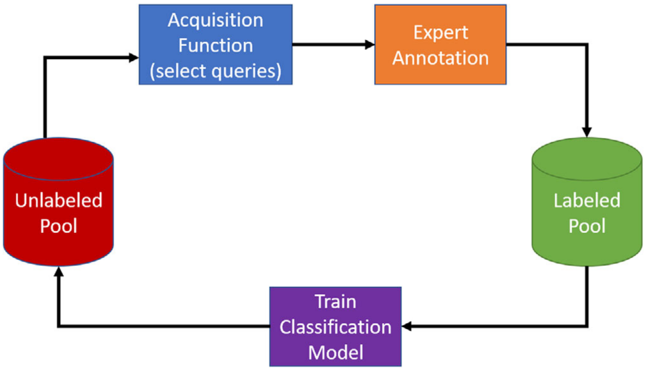
Acquisition Function (select queries)
Expert Annotation
Unlabeled Pool
Labeled Pool
Train Classification Model

FIGURE 3 Illustration of the pool-based active learning cycle for the soybean stress and weed stress classification datasets. The four method of active learning included: Least Confidence, Entropy, Deep Bayesian Active Learning, and Core-set

2.3 | Evaluated methods

Formally, let x_i be the input and y ε (1, …, N) be the output of the classification model in the active learning setup. The neural network was trained using labelled set L_{pool}. The active learning methods selects a batch of b points $[x_1^*, \ldots, x_b^*]$ from the unlabeled pool U_{pool} for expert annotation according to an acquisition criterion and add these b points to the labeled set L_{pool}. The four active learning methods are described below:

2.3.1 | Random

Principle behind the acquisition function

: The samples are chosen at random from the unlabeled data. This represents the baseline if AL methods are not used.

2.3.2 | Least confidence

Principle behind the acquisition function

This is an uncertainty sampling-based active learning method. The samples that have the lowest value for its most confident output label have the highest uncertainty in prediction. The unlabeled samples are sorted in ascending order according to maximum predicted classification probability, and the samples with the lower rank are chosen for labeling (Culotta & McCallum, 2005).

$$[\mathrm{x}_1^*, \dots, \mathrm{x}_\mathrm{b}^*] = \underset{[x_1,\dots,x_b] \subseteq \boldsymbol{U}_{pool}}{\text{argmin}} \quad \max_{k=1\dots N} p\left(\frac{y=k}{x_i}\right). \qquad (1)$$

2.3.3 | Entropy

Principle behind the acquisition function
This is an uncertainty sampling-based active learning method. The entropy of the soft-max score measures how much the predicted probabilities for each class differ from each other. The samples that have the highest output label entropy have the highest uncertainty in prediction. The unlabeled samples with highest entropy H of the predicted classification probability distribution p are chosen for labeling (Shannon, 1948).

$$H\left(\frac{y}{x}, \boldsymbol{L}_{pool}\right) = -\sum_{k=1}^{N} p\left(\frac{y=k}{x}\right) log\left(p\left(\frac{y=k}{x}\right)\right). \qquad (2)$$

$$[\mathrm{x}_1^*, \dots, \mathrm{x}_\mathrm{b}^*] = \underset{[x_1,\dots,x_b]\subseteq U_{pool}}{\text{argmax}} \quad H\left(\frac{y}{x}, \boldsymbol{L}_{pool}\right) \qquad (3)$$

2.3.4 | Deep bayesian active learning (DBAL)

Principle behind the acquisition function
This is an uncertainty sampling-based active learning method. The Monte Carlo dropout (MC-dropout) (Gal & Ghahramani, 2016) based uncertainty estimation is combined with Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2011) acquisition framework for selecting the samples in DBAL (Gal et al., 2017). The MC-dropout based uncertainty estimates are computed by averaging the outputs of T different forward stochastic passes of the input through the trained neural network model with weights w_t for the pass t during the test time. A new dropout mask with probability value of 0.5 is applied to the fully connected layer before the output soft-max layer during each of the T forward passes. During each forward pass, the dropout layer in the model turns off the output of the neurons with probability 0.5. In the other three uncertainty-based active learning methods we assumed the output of the soft-max layer as the classification probability, whereas in DBAL the classification probability is calculated as the average of T different soft-max scores of the model for a given input. The BALD acquisition function calculates the mutual information between the data samples and the model weights. Unlabeled data samples with larger mutual information between the predicted label and model weights were selected for labeling. The uncertainty estimate p is:

$$p\left(\frac{y\ =\ k}{x}, L_{pool}\right) = \frac{1}{T}\sum_{t\ =\ 1}^{T} p\left(\left(\frac{y\ =\ k}{x}\right), w_t\right) \qquad (4)$$

The acquisition criterion I is:

$$\begin{aligned} I\left(y; \frac{w}{x}, L_{pool}\right) = \ & H\left(\frac{y}{x}, L_{pool}\right) - \frac{1}{T}\sum_{t\ =\ 1}^{T}\sum_{k\ =\ 1}^{N} \\ & -p\left(\left(\frac{y\ =\ k}{x}\right),\ w_t\right) \\ & \times \log\left(p\left(\left(\frac{y\ =\ k}{x}\right),\ w_t\right)\right) \qquad (5)\end{aligned}$$

$$[\mathrm{x}_1^*, \dots, \mathrm{x}_\mathrm{b}^*] = \underset{[x_1,\ \dots,x_b]\subseteq U_{pool}}{\text{argmax}}\ I\left(Y; \frac{w}{x}, L_{pool}\right) \qquad (6)$$

2.3.5 | Core-set

Principle behind the acquisition function
A set of diverse samples that best represents the distribution of the entire dataset in the representation space learned by the neural network model are chosen for labeling. We used the output of the layer (convolution layer for MobileNetv2 model and global average pooling layer for ResNet-50 model) before the soft-max layer as the representation vector. The greedy approximation method was used to implement the core-set selection (Sener & Savarese, 2017).

3 | RESULTS AND DISCUSSION

3.1 | AL methods performance

Mean accuracy for different active learning methods for the two canonical problems on soybean stress classification and weed species classification are presented in Figures 4 and 5, respectively.

For the soybean stress classification dataset, we clearly observe that all the uncertainty sampling-based active learning methods outperform the diversity (core-set) and random sampling. Whereas, for the weed species classification dataset, all four active learning algorithms (uncertainty and diversity sampling) outperform random sampling. The performance gain due to AL methods over random sampling for plant domain datasets is similar to the improvement observed

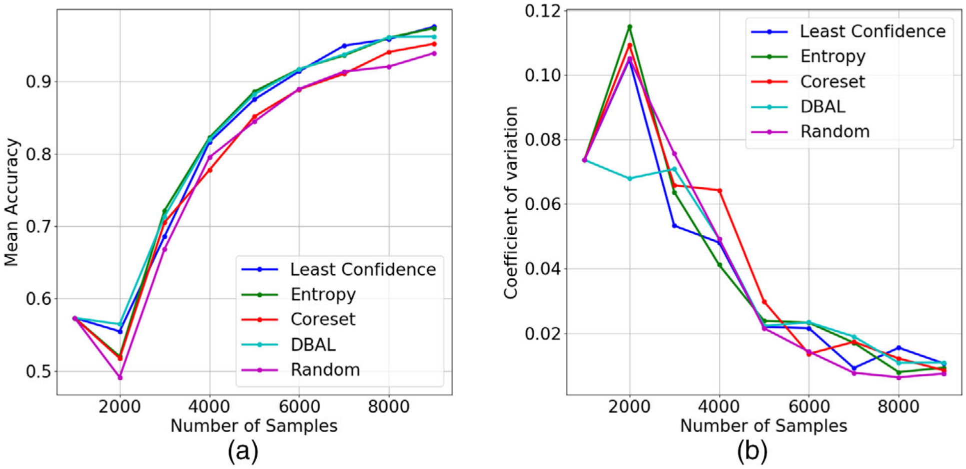

FIGURE 4 (a) MobileNetV2 accuracy plots of different active learning algorithms for soybean stress classification dataset. The results were averaged over 10 experiments. (b) We show the coefficient of variation for the classification accuracy (mean /std) from 10 repetitions

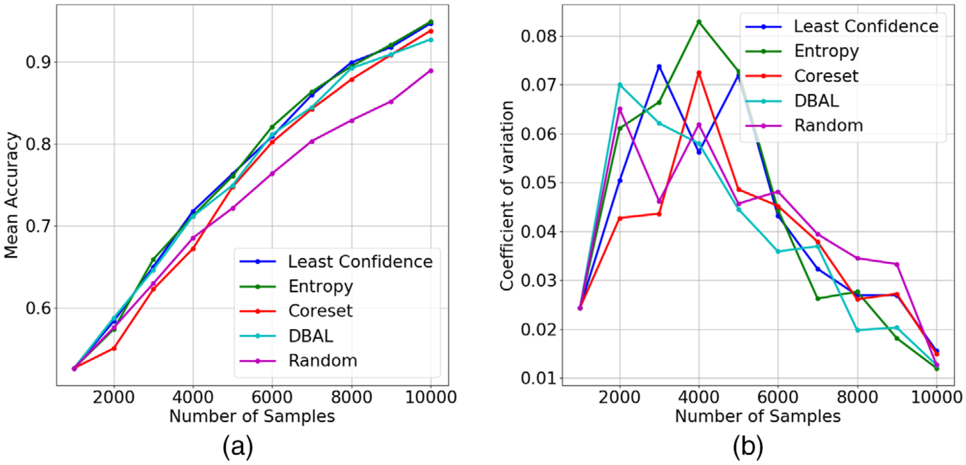

FIGURE 5 (a) ResNet-50 accuracy plots of different active learning algorithms for weed species classification dataset. The results were averaged over 10 experiments. (b) We show the coefficient of variation for the classification accuracy (mean / std) from 10 repetitions

in other domain datasets like MNIST and CIFAR10 (Beluch et al., 2018). The overall performance gains of active learning algorithms were higher for the weed species dataset than the soybean stress dataset. One reason for this could be the challenging nature of the weed species dataset, which was collected under diverse field conditions whereas the soybean dataset was collected under indoor conditions with primarily constant illumination. Additionally, the field images for the weed data set had more background objects and obscurity compared to the soybean dataset, which consisted of images under more controlled conditions (Ghosal et al., 2018). Hence, the random sampling-based annotation method provides a stronger baseline for the soybean stress dataset. The dip in accuracy at 2,000 samples for the soybean dataset was due to high class-imbalance in the expert annotated dataset after sample selection. The coefficient of variation of the classification accuracy for the soybean dataset followed a similar decreasing trend for all the active learning methods as shown in Figure 4b. For the weed dataset, the coefficient of variation was initially high (until annotation 4,000 samples) and then followed a decreasing trend for all the active learning methods as shown in Figure 5b. The challenging nature of the weed dataset might have caused higher fluctuations in model accuracy (high coefficient of variation) when only a small number of labelled samples (up to 4,000 samples) are used for training.

To identify the best active learning method for each dataset, we rank ordered the active learning methods-based on the classification accuracy from 1 to 5 (1 for the highest accuracy and 5 for the lowest accuracy) for each repetition of the experiment. The mean of the ranks from 10 different repetitions is shown in Figure 6 for the soybean and weed

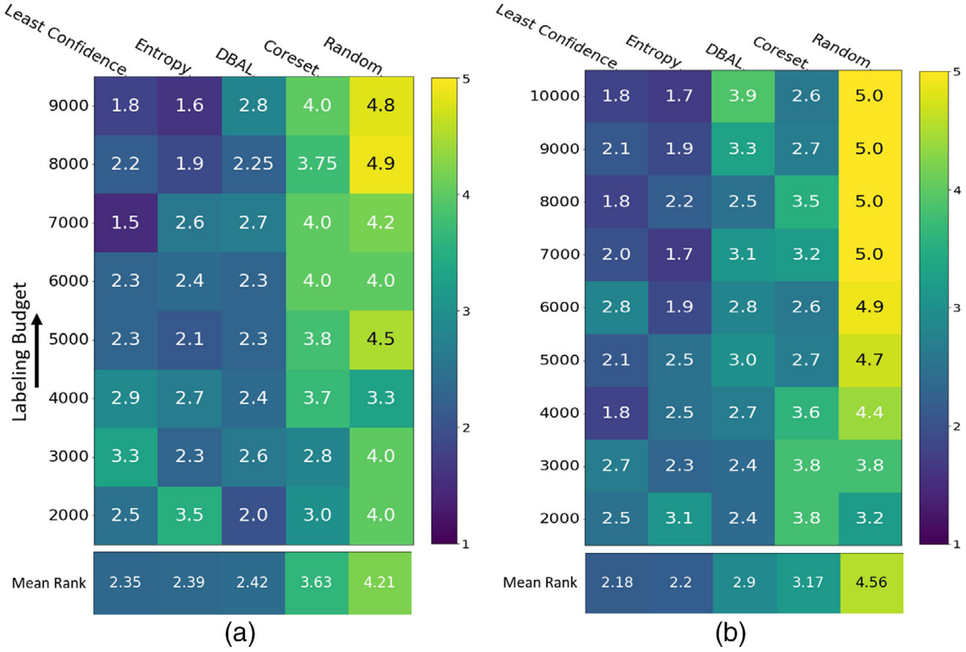

FIGURE 6 We rank ordered the active learning methods-based on the classification accuracy from 1 to 5 (1 – highest accuracy, 5- lowest accuracy) obtained in each repetition of the experiment. The average of the ranks from 10 repetitions are shown in (a) soybean dataset, and (b) weed dataset. The overall mean rank performance of active learning methods across different labeling budgets is shown in the bottom row of (a) and (b)

datasets. The mean rank of the active learning methods across *all labeling budgets* is shown in the last row of Figures 6a and 6b for the soybean and weed dataset, respectively. We observed that Least Confidence sampling was the best performing active learning method for both the datasets considering all the different labeling budgets. However, DBAL was the best performing method when the labeling budget was small (up to 4,000 samples for the soybean dataset and 2,000 samples for the weed dataset). Although the computational time of Least Confidence is negligible compared to DBAL, its overall performance was better when considering performance across all labeling budgets. The overall rank order of the active learning methods followed similar trend for these two datasets. The uncertainty sampling-based methods (Least Confidence, Entropy and DBAL) performed better than diversity sampling-based core-set method. All the active learning methods performed better than the baseline random sampling for the soybean and weed datasets.

3.1.1 | Which samples are selected?

A random selection strategy is expected to blindly pick new samples for annotation, therefore the distribution of selected points is expected to be uniform (Figure 7 last row). In contrast, we anticipate the AL-based methods to pick fewer samples from classes that are well predicted, and instead pick more points from classes that are not well predicted. We visualize this expected behavior in Figures 7a (for soybean stress classification) and 7b (for weed classification). The per-class classification accuracy of different active learning methods and random sampling is shown in the first column of Figures 7a and 7b.

We also plot how many additional samples are selectively chosen to achieve this differential per-class accuracy. This is shown in the second column in Figures 7a and 7b. The per class sample selection percentage, i.e., how many samples are used per class (calculated as the number of samples selected from a class/total number of samples available in a class) of different active learning algorithms are presented in the second column of Figures 7a and 7b. The accuracy plot of the random method indicates the classes that are hard and easy to predict. The clear inverse relationship between the performance of individual classes shown in the accuracy plot of random and the number of samples chosen is apparent across all uncertainty-based AL methods. This is in stark contrast to a naive random sampling.

3.1.2 | Soybean dataset sample selection

Classes ‘0’ and ‘7’ have low per-class classification accuracy from random sampling-based annotation (Figure 7a). Least Confidence, Entropy and DBAL methods chose more samples from the classes ‘0’ and ‘7’ and obtained better per-class accuracy than random sampling. These results are consistent with previous work (Ghosal, et al, 2018), where class ‘0’, Bacterial blight and class ‘7’ bacterial pustule were reported

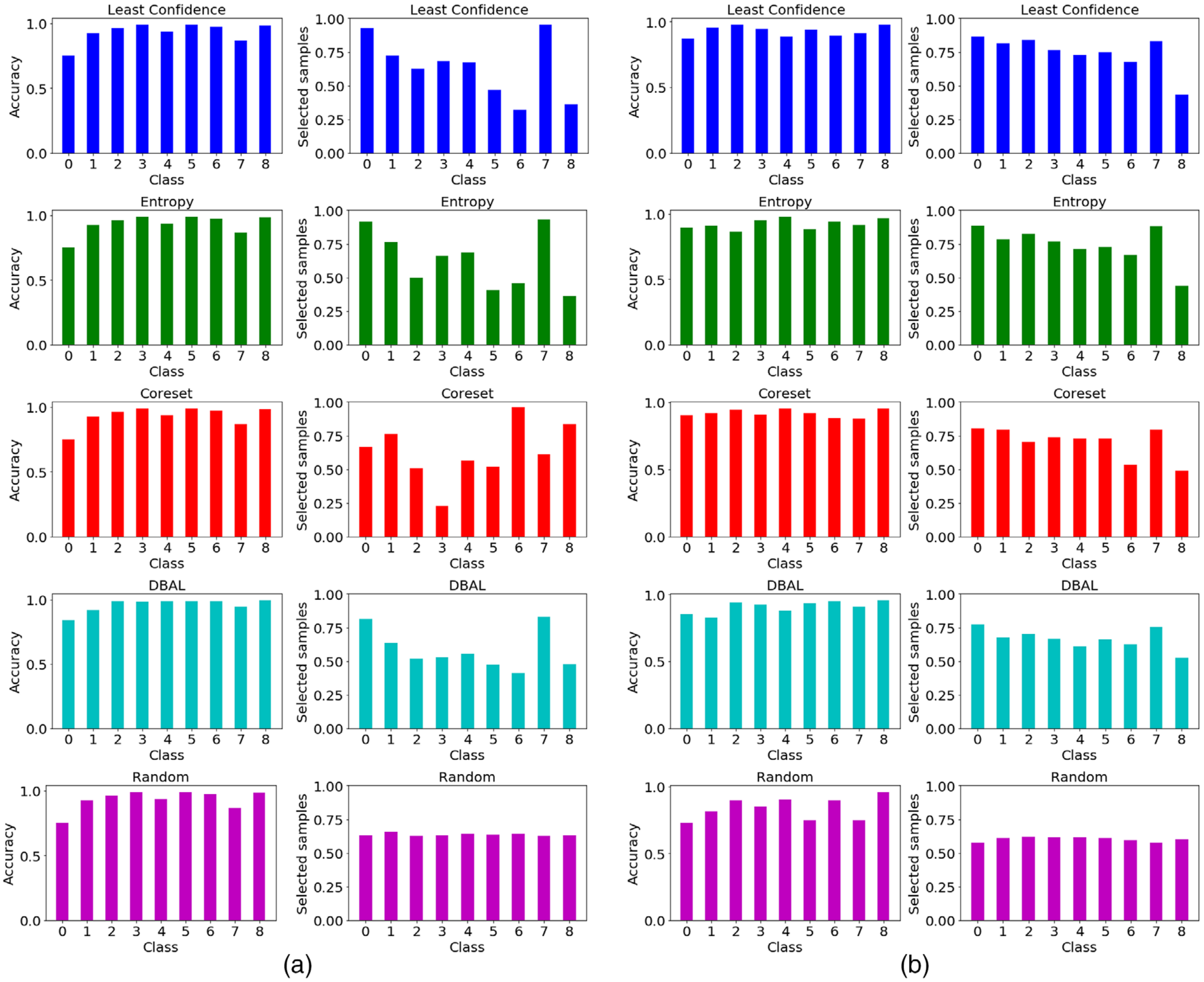

FIGURE 7 (a) An example of per class classification accuracy of MobileNetV2 model on the soybean stress classification dataset using different active learning algorithms from a single experiment for a labeling budget of 9,000 samples. (b) Per class sample selection percentage (Number of sample selected from a class/Total number of samples available in a class) of different active learning algorithms for the results shown in (a). The nine classes are as following: 0 = bacterial blight, 1 = Septoria Brown Spot, 2 = Frogeye Leaf Spot, 3 = Healthy, 4 = Herbicide Injury, 5 = Iron Deficiency Chlorosis, 6 = Potassium Deficiency, 7 = Bacterial Pustule, 8 = Sudden Death Syndrome. (b) An example of per class classification accuracy of ResNet50 model on the weed dataset using different active learning algorithms from a single experiment for a labeling budget of 10,000 samples. (b) Per class sample selection percentage (Number of samples selected from a class/Total number of samples available in a class) of different active learning algorithms for the results shown in (a). The nine classes are as following: 0 = Chinee apple, 1 = *Lantana*, 2 = *Parkinsonia*, 3 = *Parthenium*, 4 = Prickly acacia, 5 = Rubber vine, 6 = Siam Weed, 7 = Snake Weed, 8 = Negatives. The best way to look at this plot is bottom up, i.e., first look at that random sampling approach (which is the baseline) and compare how the accuracy of specific classes are improved via AL approaches

to be the most confusing (i.e., least discriminative between the two) bacterial diseases among the 9 classes, causing even expert raters rating challenges during manual classifying due to similarity of disease symptoms. Least confidence, Entropy and DBAL do an excellent job in choosing more samples from stresses that are highly confusing when compared to less confusing stresses. This is very promising from a domain perspective because confounding symptoms classes are more extensively sampled by these three AL methods. These uncertainty-based methods sampled only a small percentage of samples from classes that have high per-class accuracy for random sampling (classes ‘3’, ‘5’, ‘6’, and ‘8’) method. Uncertainty-based AL algorithms adaptively sampled more from the low accuracy classes of the random sampling method (classes ‘0’, ‘1’, and ‘7’) and sampled less from the high accuracy classes of random sampling methods (classes ‘6’ and ‘8’), contrasting it with the diversity-based core-set method. In contrast to the uncertainty-based acquisition functions of

LC, Entropy, and DBAL, core-set uses a diversity based sampling. Its comparatively poor performance can be explained by the fact that it chooses less samples from classes exhibiting less diversity, even if that class is difficult to classify.

3.1.3 | Weed dataset sample selection

Classes ‘0’ and ‘7’ have low per-class classification accuracy from random sampling-based annotation (Figure 7b). Least Confidence, Entropy and DBAL methods chose more samples from the classes ‘0’ and ‘7’ compared to random sampling method and obtained better per-class accuracy than random sampling method. Classes ‘6’ and ‘8’ have high per-class classification accuracy from random sampling-based annotation. Uncertainty-based AL algorithms adaptively sampled less from the high accuracy classes of random sampling method (classes ‘6’ and ‘8’).

The AL methods show promising results for plant sciences problems where extensive data are needed to train useable models. These include diverse applications including complex phenotype extracting workflows like the cluttered image problem for soybean cyst nematode egg detection (Akintayo et al., 2018), hyperspectral imaging (Nagasubramanian et al., 2019; Roscher et al., 2016), abiotic stress disease rating (Naik et al., 2017; Zhang, Naik, et al., 2017), and root imaging (Falk, Jubery, Mirnezami, et al., 2020; Falk, Jubery, O’Rourke, et al., 2020).

4 | CONCLUSIONS

In this work, we explore the usefulness of active learning methods for reducing the labeling needed for two different classification tasks. We observed that uncertainty-based active learning methods consistently outperformed random sampling-based annotation for both the soybean stress classification and weed classification task. Least confidence sampling method was the best performing active learning method for both the datasets. We believe that active learning methods can be quite helpful in reducing the amount of labeling needed for image-based plant phenotyping tasks like classification, detection, and segmentation. We note that recent theoretical developments place active learning methods on firmer grounds [especially considerations of how discrepancy in training distribution vs original distribution can be fixed, see Farquhar et al. (2021)], cementing their utility. There are several promising avenues in which the power of AL methods can be combined with methods with complementary strengths. Specifically, we advocate for integrating AL methods with transfer learning, semi-supervised and unsupervised representation learning methods to further increase the labeling efficiency for the challenging phenotyping tasks.

ACKNOWLEDGMENTS

We thank Olsen et al. (2019) for making the weed dataset publicly available. We thank J. Brungardt, B. Scott, J. Zhang, and graduate and undergraduate students (A.K.S. laboratory) for help in imaging and data collection; A. Lofquist and J. Stimes (B.G. laboratory) for developing the marking app and data storage back-end.

AUTHOR CONTRIBUTIONS

Koushik Nagasubramanian: Conceptualization; Data curation; Methodology; Software; Validation; Writing-original draft; Writing-review & editing. Talukder Jubery: Conceptualization; Methodology; Software; Writing-review & editing. Fateme Fotouhi Ardakani: Software; Validation; Writing-review & editing. Seyed Vahid Mirnezami: Software; Writing-review & editing. Asheesh Singh: Conceptualization; Funding acquisition; Supervision; Writing-original draft; Writing-review & editing. Arti Singh: Conceptualization; Data curation; Funding acquisition; Supervision; Writing-review & editing. Soumik Sarkar: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing-review & editing. Baskar Ganapathysubramanian: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing-original draft; Writing-review & editing

FUNDING INFORMATION

This work was funded by the Iowa Soybean Association (A.S.), an Iowa State University (ISU) internal grant (to all authors), NSF/USDA National Institute of Food and Agriculture grants 2017-67007-26151 (to all authors), 2020-68013-30934 (BG, ZJ), 2020-67021-31528 (BG), a Bayer Chair in Soybean Breeding at Iowa State University (A.K.S.), Raymond F. Baker Center for Plant Breeding at Iowa State University (A.K.S.), an ISU Plant Sciences Institute fellowship (to B.G., A.K.S., and S.S.), and USDA IOW04714 (to A.S. and A.K.S.) and NSF 1954556 (all authors).

CODE AVAILABILITY

All the codes for the active learning approaches described in this work are available for the community at https://github.com/koushik-n/Active-Learning-Plant-Phenotyping.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Koushik Nagasubramanian https://orcid.org/0000-0003-2708-3945
Arti Singh https://orcid.org/0000-0001-6191-9238

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., … Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Osdi*, *16*, 265–283. https://doi.org/10.1038/nn.3331

Akintayo, A., Tylka, G. L., Singh, A. K., Ganapathysubramanian, B., Singh, A., & Sarkar, S. (2018). A deep learning framework to discern and count microscopic nematode eggs. *Scientific Reports*, *8*, 9145. https://doi.org/10.1038/s41598-018-27272-w

Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., & Cairns, J. E. (2018). Translating high-throughput phenotyping into genetic gain. *Trends in Plant Science*, *23*, 451–466. https://doi.org/10.1016/j.tplants.2018.02.001

Beluch, W. H., Genewein, T., Nürnberger, A., & Köhler, J. M. (2018). The power of ensembles for active learning in image classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9368–9377.

Chandra, A. L., Desai, S. V., Balasubramanian, V. N., Ninomiya, S., & Guo, W. (2020). Active learning with point supervision for cost-effective panicle detection in cereal crops. *Plant Methods*, *16*(1), 1–16. https://doi.org/10.1186/s13007-020-00575-8

Chollet, F. (2015). Keras. https://keras.io

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active Learning with Statistical Models. In *Journal of Artiicial Intelligence Research, Vol. 4*. http://www.jair.org/papers/paper295.html

Culotta, A., & McCallum, A. (2005). Reducing labeling effort for structured prediction tasks. *AAAI*, *5*, 746–751.

Falk, K. G., Jubery, T. Z., Mirnezami, S. V., Parmley, K. A., Sarkar, S., Singh, A., Ganapathysubramanian, B., & Singh, A. K. (2020). Computer vision and machine learning enabled soybean root phenotyping pipeline. *Plant Methods*, *16*, 5. https://doi.org/10.1186/s13007-019-0550-5

Falk, K. G., Jubery, T. Z., O'Rourke, J. A., Singh, A., Sarkar, S., Ganapathysubramanian, B., & Singh, A. K. (2020). Soybean root system architecture trait study through genotypic, phenotypic, and shape-based clusters. *Plant Phenomics*, *2020*, 1925495. https://doi.org/10.34133/2020/1925495

Farquhar, S., Gal, Y., & Rainforth, T. (2021). On statistical bias in active learning: How and when to fix it. *International Conference on Learning Representations*. https://openreview.net/forum?id=JiYq3eqTKY

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 1050–1059.

Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep bayesian active learning with image data. *Proceedings of the 34th International Conference on Machine Learning-Volume*, *70*, 1183–1192.

Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., & Sarkar, S. (2018). An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences*, *115*, 4613–4618. https://doi.org/10.1073/pnas.1716999115

Ghosal, S., Zheng, B., Chapman, S. C., Potgieter, A. B., Jordan, D. R., Wang, X., Singh, A. K., Singh, A., Hirafuji, M., Ninomiya, S., Ganapathysubramanian, B., Sarkar, S., & Guo, W. (2019). A weakly supervised deep learning framework for sorghum head detection and counting. *Plant Phenomics*, *2019*, 1525874. https://doi.org/10.34133/2019/1525874

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *ArXiv Preprint ArXiv:1112.5745*.

Kao, C. - C., Lee, T. - Y., Sen, P., & Liu, M. - Y. (2018). Localization-aware active learning for object detection. *Asian Conference on Computer Vision*, 506–522.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Nagasubramanian, K., Jones, S., Singh, A. K., Sarkar, S., Singh, A., & Ganapathysubramanian, B. (2019). Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods*, *15*(1), 98. https://doi.org/10.1186/s13007-019-0479-8

Naik, H. S., Zhang, J., Lofquist, A., Assefa, T., Sarkar, S., Ackerman, D., Singh, A., Singh, A. K. & Ganapathysubramanian, B. (2017). A real-time phenotyping framework using machine learning for plant stress severity rating in soybean. *Plant Methods*, *13*(1), 23. https://doi.org/10.1186/s13007-017-0173-7

Olsen, A., Konovalov, D. A., Philippa, B., Ridd, P., Wood, J. C., Johns, J., Banks, W., Girgenti, B., Kenny, O., Whinney, J., Calvert, B., Azghadi, M. R., & White, R. D. (2019). DeepWeeds: A multiclass weed species image dataset for deep learning. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-018-38343-3

Roscher, R., Behmann, J., Mahlein, A. - K., Dupuis, J., Kuhlmann, H., & Plümer, L. (2016). Detection of disease symptoms on hyperspectral 3D plant models. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, *3*(7), 88–96.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. - C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Sener, O., & Savarese, S. (2017). Active learning for convolutional neural networks: A core-set approach. *ArXiv Preprint ArXiv:1708.00489*.

Settles, B. (2009). Active learning literature survey. http://burrsettles.com/pub/settles.activelearning.pdf

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science*, *21*, 110. https://doi.org/10.1016/j.tplants.2015.10.015

Singh, A., Jones, S., Ganapathysubramanian, B., Sarkar, S., Mueller, D., Sandhu, K., & Nagasubramanian, K. (2020). Challenges and opportunities in machine-augmented plant stress phenotyping. *Trends in Plant Science*.

Singh, A. K., Ganapathysubramanian, B., Sarkar, S., & Singh, A. (2018). Deep learning for plant stress phenotyping: Trends and future perspectives. *Trends in Plant Science*, *23*, 883–898. https://doi.org/10.1016/j.tplants.2018.07.004

Singh, D. P., Singh, A. K., & Singh, A. (2021). *Plant Breeding and Cultivar Development*. 1st ed. Academic Press. p 662.

Tapas, A. (2016). Transfer learning for image classification and plant phenotyping. *International Journal of Advanced Research* *in Computer Engineering and Technology (IJARCET)*, *5*, 2664–2669.

Valerio Giuffrida, M., & Scharr Forschungszentrum Jülich, H. (2017). *ARIGAN: Synthetic Arabidopsis Plants using Generative Adversarial Network*. https://openaccess.thecvf.com/content_ICCV_2017_workshops/papers/w29/Giuffrida_ARIGAN_Synthetic_Arabidopsis_ICCV_2017_paper.pdf

Wang, K., Zhang, D., Li, Y., Zhang, R., & Lin, L. (2016). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*, 2591–2600. https://doi.org/10.1109/TCSVT.2016.2589879

Yang, L., Zhang, Y., Chen, J., Zhang, S., & Chen, D. Z. (2017). Suggestive annotation: A deep active learning framework for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 399–407.

Zhang, J., Naik, H. S., Assefa, T., Sarkar, S., Reddy, R. V. C., Singh, A., Ganapathysubramanian, B., & Singh, A. K. (2017). Computer vision and machine learning for robust phenotyping in genome-wide studies. *Scientific Reports*, *7*, 44048. https://doi.org/10.1038/srep44048

Zhang, Y., Lease, M., & Wallace, B. C. (2017). Active discriminative text representation learning. *Thirty-First AAAI Conference on Artificial Intelligence*.

How to cite this article: Nagasubramanian K, Jubery T, Fotouhi Ardakani F, et al. How useful is active learning for image-based plant phenotyping? *Plant Phenome J*. 2021;*4:*e20020. https://doi.org/10.1002/ppj2.20020