A Reflection on AI Model Selection for Digital Agriculture Image Datasets

Seth Ockerman[1,2], **John Wu**[2], **Zichen Zhang**[2], **Christopher Stewart**[2]

Grand Valley State University[1], The Ohio State University[2]

Abstract

Cameras, sensors, and autonomous vehicles deployed in agricultural settings are producing large, complex, and highly multidimensional datasets. Artificial intelligence techniques can extract insights hidden within these datasets to automate crop management and develop better farming practices. In particular, recent studies have shown that neural networks can accurately characterize crop health conditions within digital agriculture datasets. However, choosing between neural network architectures is challenging; One must select from multiple architectures and hyper parameters. Benchmark datasets, i.e., datasets that represent a class of similar datasets, are often used to select models for digital agriculture datasets. However, if benchmark datasets are not faithful representatives for digital agriculture datasets, their use could lead to poor model selection. This paper demonstrates the danger of using standard vision benchmarks to inform model selection for digital agriculture datasets. We then propose a gradient-boosting prediction approach that would significantly reduce costs to benchmark digital agriculture datasets directly, which could improve the fit between model and dataset.

1 Introduction

Artificial intelligence (AI) models learned from digital agriculture (DA) datasets can improve crop health, yield, and sustainability (Mulla 2013). As DA datasets proliferate and expectations on their efficacy rise, model selection is a key challenge, i.e., for a given dataset, which machine learning approach and hyperparameter settings will produce models capable of improving crop health? Benchmark datasets are widely used in machine learning for model selection. Benchmarks serve as reference points on efficacy and performance. They are used to predict performance for similar but untested datasets, providing one approach to solve model selection problems. CIFAR (Krizhevsky 2009) and MNIST (Deng 2012) are widely used benchmark datasets for computer vision model selection. *If* these benchmark datasets provide model selections representative for DA vision tasks, then their results can be used to characterize efficacy and performance without training and testing models on each DA dataset.

Given the importance of data in the model design process and the relatively few datasets that are used for benchmarking (Koch et al. 2021), we ask, *are widely used benchmark datasets actually faithful references for new, emerging DA datasets?* And, *are they faithful representatives at every stage in the process of model selection, including architecture selection, CNN filter dimensions, and hyperparameter searching?*

Whether they are faithful representatives or not, benchmark datasets greatly reduce costs for model selection. For this work, we trained one state-of-the-art neural network architecture with a DA dataset on AWS cloud. The total cost was $36. Given per-acre profit on US corn fields is $148, training and testing multiple model architectures across more hyperparameters with multiple metrics of efficacy and performance is clearly cost prohibitive (Boubin et al. 2019; Foreman 2014).

The structure and composition of DA datasets differs from widely used benchmarks. First, DA datasets increasingly require high-definition images and lever domain-specific convolutions. In vision tasks, many DA datasets include infrared and thermal channels, going beyond RGB channels included in widely used benchmarks. Convolutions may capture concepts like leaf area (Fang and Liang 2003), going beyond classic edge detection. Additionally, the variation in pixel values between different classes in DA is subtle: A small number of pixel inversions can shift the label on aerial images from normal to severe leaf defoliation (Zhang et al. 2022). In prior work, Zhang et al. (Zhang et al. 2022) eschewed benchmark datasets and explored model selection through exhaustive training and testing across 8 machine learning algorithms, multiple neural network architectures, and a variety of hyperparameters. They found only one approach provided a practical, cost-effective solution to manage crop scouting for pesticide use.

The high dimensionality of DA images and the inherent complexity of the models increases the cost of testing every model on a given DA dataset. We propose an approach to reduce these costs. We explore a gradient-boosting approach that makes use of previous classification models' weights to predict a new model's final accuracy on a DA dataset after just a few epochs of training, extending recent efforts to predict accuracy (Unterthiner et al. 2020; Yamada and Morimura 2016). This would significantly cut down on the cost of model selection for classification tasks on DA datasets and enable more models to be directly tested against the intended DA dataset.

In this paper, we study the influence of popular benchmarking practices on model selection for DA. We examine the inherent differences in neural network performance on 8 different datasets. Finally, we explore a gradient-boosting approach that exploits neural network weights to reduce costs for model selection. This paper is formatted as follows: Section 2 will summarize past work that examined popular benchmarking datasets before examining past approaches for predicting final neural network accuracy. Section 3 will detail our experiments with classical and DA datasets. Finally, section 4 will explore what an early accuracy prediction mechanism for DA models might look like.

2 Related Work

This section is organized as follows. Section 2.1 will explore past work on dataset profiling and bias to demonstrate the danger of using standard benchmarking datasets for agriculture model selecton. Section 2.2 will describe new trends leading to more DA datasets and the challenges presented in DA model benchmarking. Finally, section 2.3 will provide an overview of past work predicting neural network accuracy from weights.

2.1 Bias in Benchmarking Datasets

It is standard design practice to test new models (trained weights, hyperparameter choices, and architecture itself) by running them against popular datasets like CIFAR (Krizhevsky 2009). This is in part due to the popularity of dataset competitions as a method to popularize learning architectures. However, this popularity comes with drawbacks. An analysis by (Everingham et al. 2010) found no statistical difference between the performance of the top 10 algorithms in the 2010 PASCAL Visual Object Classes competition. This suggests top algorithms are not fundamentally different from one another. Researchers worry that the lack of dataset diversity among popular datasets is causing models to learn from idiosyncrasies of the images rather than significant generalized characteristics (Ponce et al. 2007; Torralba and Efros 2011). (Torralba and Efros 2011) also found that models trained on one representative dataset tend to test poorly on other representative datasets of the same category (i.e types of cars). This is not surprising given models tend to favor their own test sets. However, it is concerning that supposedly representative datasets do not create models with high enough levels of generalization to transfer to other similarly representative datasets.

While the diversity of representative datasets has improved over time, they still suffer from limitations. Recent work found that neural networks were learning from noise in biomedical image datasets instead of the relevant medical features ((Dhar and Shamir 2021)). These datasets were popular image benchmarks that many new algorithms and networks were tested against. While these datasets certainly provide a good sanity check for new approaches, they also can determine the success or failure of a new approach. This bias can hamper novel approaches from widespread adoption.

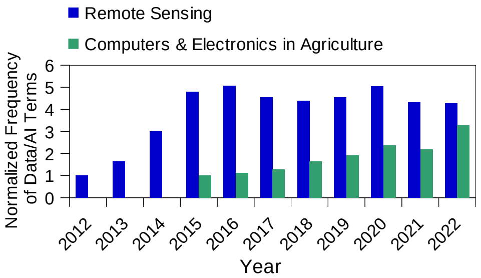

Fig. 1: Data and AI topics are surging in top digital agriculture journals.

Past works suggest current representative datasets have the potential to skew development toward a narrow solution space not representative of the complexity of real-world problems. In theory, the simple solution is to create a variety of datasets by sub-area which are perfectly representative. However, this is both impractical due to the black box nature of neural networks and impossible due to pure cost. A new benchmarking paradigm that enables lightweight low-cost model testing against specific datasets is needed.

2.2 A Surge in Datasets

In recent years, the velocity of agriculture dataset creation has surged (Lu and Young 2020). Figure 1 shows the frequency of data and AI topics in Remote Sensing and Computers and Electronics in Agriculture papers. These topics have surged 4X and 2X in each journal respectively. A recent survey of DA datasets analyzed a collection of the complex, use-case specific DA datasets and suggested that these datasets could be used for general DA benchmarking (Lu and Young 2020). However, while machine learning algorithms seek to perform classification and segmentation with near equal accuracy on all DA datasets, the specifics needed for success vary greatly depending on dataset and intended use-case. We contend that these different use-cases impact the efficacy of machine learning algorithms, in terms of accuracy, training time and computational cost. This necessities an approach which can take both dataset domain and use-case into account.

2.3 The Significance of Weights in Predicting Neural Network Accuracy

The primary way past works have attempted to predict the final accuracy of a network is through the use of early training curves (Domhan, Springenberg, and Hutter 2015). However, both concurrent and recent work has shown a strong relationship between a network's weights and its characteristics and performance. (Yamada and Morimura 2016) was able to use weights obtained early in a neural network's training process to predict its testing accuracy with higher accuracy than existing learning curve-based approaches. Very recently, (Unterthiner et al. 2020) found that using simple summary statistics based on network's fully trained weights could predict test set accuracy with an R^2 score of over 0.98. This presents a compelling case for more investigation into the use of weights to predict neural network performance.

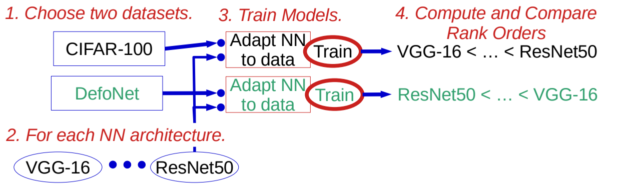

Fig. 2: We compared the rank order of NN architectures between *classic* image processing datasets and DA datasets.

(Yamada and Morimura 2016) uses a variety of weight features to predict eventual accuracy based on early epoch weights. (Unterthiner et al. 2020) creates a dataset of 32,000 small-scale neural network's fully trained weights mapped to their final test set accuracy. They then use that dataset (dubbed CNN Zoo) to train gradient boosting machines (GBMs) to predict the test set accuracy of large-scale models. While our early work (explored in 4) is inspired by both (Yamada and Morimura 2016) and (Unterthiner et al. 2020), it also builds on them in a significant way. (Unterthiner et al. 2020) uses a small 4-layer network with randomly initialized weights. To build on this, we extract features from a large deep neural network that incorporates pre-trained ImageNet weights. We theorize that transfer learning will enable us to reduce the number of hyperparameter configurations needed to create a representative solution space (i.e scale of 1000s vs less than 100). (Unterthiner et al. 2020)'s work treated a network's training cycle as one unit of data to be mapped to final accuracy. We instead record data at an epoch level. This enables the prediction of final accuracy based on only a few training epochs. In addition, both (Yamada and Morimura 2016) and (Unterthiner et al. 2020) focus on classical datasets. We instead focus on complex domain-specific datasets which better account for the potential bias of representative datsaets.

3 The Difference in AI model Performance Across Domain

We studied AI models trained and tested on classic vision benchmarks and DA datasets, comparing relative performance of the models on each dataset. We hypothesized that the rank order of AI models would be consistent, suggesting that widely used vision benchmarks can be used to for model selection on DA datasets.

3.1 Methodology

Figure 2 provides an overview of our methodology. To measure the distance between model rankings, we used 4 neural network architectures for image processing: InceptionV3 (I), VGG16 (V), EfficientNet (E), and ResNet50 (R) (He et al. 2016; Simonyan and Zisserman 2014; Szegedy et al. 2016; Tan and Le 2019). To support datasets with any number of classes, we appended a fully connected classification network atop each network. We trained models until validation accuracy stopped improving for 10 epochs.

We trained and tested these networks using two categories of datasets: classical datasets and DA datasets. We selected 4 classical datasets based on popularity: CIFAR-10, CIFAR-100, imagenette2 (a subset of Imagenet), and MNIST ((Deng et al. 2009; Krizhevsky 2009; Deng 2012)). The DA datasets selected are as follows: fruits-360, PlantVillage, weed seedlings, and leaf defoliation dataset ((Beck et al. 2020; Hughes and Salathe 2015; Mureșan and Oltean 2018; Zhang et al. 2022)). We selected these datasets because they each represent a fundamental task in DA (e.g. fruit classification, drone-based defoliation detection, diseased plant classification, etc.). More details on each dataset can be found in Figure 3.

Finally, we ranked each model's performance for each dataset. For example, on CIFAR-10 ranking by accuracy, we observed the following rank order: I, V, E, R. On the Leaf Defoliation dataset, we observed the following rank order: E, R, I, V. By focusing on rankings, we isolate and standardize the effectiveness of a given DNN relative to a specific dataset. This avoids the inherent bias of comparing network accuracy across different datasets. To measure the distance between datasets, we use Euclidean distances.

3.2 Results

Figure 3 shows the distance of each accuracy rank order from the baseline vector CIFAR-10. Figure 3 show that the AI models were ranked changed significantly between agriculture datasets and vision benchmarks. CIFAR-10, CIFAR-100, and Imagenette ranked models in the same order, suggesting these benchmarks can used in lieu of each other. However, all 4 agriculture datasets permuted rank order significantly. We repeated our tests and replaced accuracy with training loss as the rank order metric. Figure 3 shows similar results: Rank order for both CIFAR-100 and Imagenette were identical to CIFAR-10. Rank order on PlantVillage and Leaf Defoliation datasets differed greatly. However, MNIST, Weed Seedlings, and Fruits-360 displayed the same distance from CIFAR-10. We believe this is due to the simplicity of Fruits-360 and Weed Seedlings compared to the other datasets.

Our findings suggest that a model's performance on classical datasets is not representative of how a model will perform on a given DA dataset, making classical datasets a poor benchmarking choice for DA model selection. Additionally, our findings demonstrate that DA datasets are not faithful representations of how a model will perform on all DA datasets. A new benchmarking paradigm that goes beyond domain is needed.

4 Lightweight Model Profiling

Section 3 demonstrated that reusing DA learning architectures across different datasets, both classical and DA datasets, falsely assumes that a model's success is not dataset specific. This leads to subpar DA models being chosen for hyperparemter tuning. A simple but effective solution to the demonstrated problem is to view the dataset profiling process as part of hyperparameter searching. However, profiling each dataset-model combination makes the naive assumption that cost can grow towards infinity, particularly with high cost DA datasets. Individual dataset profiling is

Data Set	Classes	Images
CIFAR-10	10	60,000
CIFAR-100	100	60,000
MNIST	10	60,000
Imagenette	10	13,000
Leaf Defoliation	2	97,395
Fruits 360	131	90,483
Plant Village	38	87,000
Weed Seedlings	8	34,666

Table 1. Image data sets. Vision benchmarks are blue. Use-inspired agriculture data sets are green.

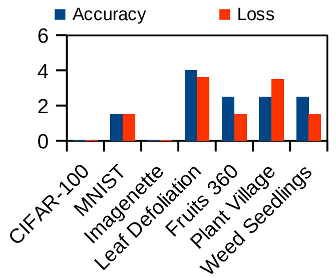

Figure 1. Euclidean distance between rank order of models on **CIFAR-10** versus other data sets.

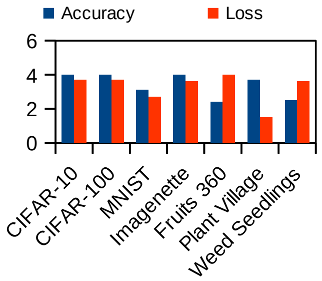

Figure 2. Distance between AI model rankings of **Leaf Defoliation Data Set** versus other data sets.

not feasible on a large scale for the majority of developers. A new method of model benchmarking and dataset profiling is needed. This section details our experimentation using early training weights to predict final training accuracy.

4.1 Methodology

We designed an experiment to test the use of weights to predict final testing accuracy on a complex DA dataset: the Leaf Defoliation dataset ((Zhang et al. 2022). To create a solution space, we tested a variety of hyperparameter configurations using VGG16 with pre-trained ImageNet weights combined with a small, fully connected classification neural network.

We chose 35 different hyperparameter configurations to explore, varying optimizer, learning rate, and final layer activation function. We ran each of the 35 configurations for 75 epochs and record the final accuracy per configuration.

Each epoch we saved the weights to later calculate summary statistics. We calculated the mean, variance, and q-th percentiles where $q \in \{0, 25, 50, 75, 100\}$ (Unterthiner et al. 2020) for biases and kernel weights separately. We calculated these statistics for each neural network layer, creating a 2x7 vector for each layer. Combining all 17 layers into a single matrix, we generated a 17x2x7 representation of means, variances, and percentiles. This matrix is then mapped to the final testing accuracy of its respective model's configuration. Because we created mappings at an epoch level, we significantly increase the sample space we explore. In total, we created 2625 accuracy mappings from the Leaf Defoliation dataset.

We performed an 80/20 train/test split of our vector accuracy mappings. In contrast to typical train/test splits, we did not randomize the placement of the mappings. Instead, we ensured the 20% in the test set is composed entirely of hyperparameter combinations that do not exist in the training set. The nonrandom nature of the test set is to prevent overfitting and leakage from the training set to the test set. Our approach resulted in a training set with 31 hyperparameter configurations and a testing set with four unseen hyperparameter configurations.

For prediction, we selected gradient boosting machines implemented in XGBoost's gradient boosting forest package ((Chen and Guestrin 2016)). We split our training set into a train and validation set at an 80/20 ratio. Using that validation set, we performed hyperparameter tuning, resulting in a model of 128 estimators with a max depth of 7 per tree.

The testing data consists of the same number of hyperparameter configurations each time. However, we varied the percent of each model configuration's training cycle we include in the prediction process. By limiting the data inputted to the GBM to an artificial n-th epoch of training time, we simulated lightweight benchmarking runs. For example, by reducing input data to the first five epochs of weight data, we tested the accuracy of predictions given only a fraction of the total training time. We select epochs 4, 8, 19, 38, 56, 76 which represent 5%, 10%, 25%, 50%, 75%, and 100% of the training time respectively.

4.2 Results

After hyperparameter tuning on the validation set, we tested our trained model on each of the test splits. Using the entirety of the test data, we achieved an absolute accuracy of 81.33% and a relative root mean squared error (RRMSE) of 0.196. Crucially, there is a minimal decrease in accuracy if we reduce the number of epochs we use as an input for our test set. Using only 4 epochs of input data from the test set (representing roughly 5% of its training time) we achieved an absolute accuracy of 80.61% and an RRMSE of 0.201. This trend continues across all designated input data splits. Across all splits, there is less than a 1% change in absolute accuracy and RRMSE. This indicates that the number of epochs of input data has little impact on the accuracy of a fully trained model.

5 Discussion

We have shown that widely used benchmark datasets are not faithful reference points for DA datasets. Further, DA vision datasets are not always faithful reference points for other DA vision datasets. We argue for new approaches to create benchmark datasets for DA. One potential approach explored in this paper exploits weight distributions observed in previous DA neural networks to predict accuracy and efficacy. In section 4, we accurately predicted unseen neural network's final test set accuracy using weights obtained after a fraction of total training time. It is especially promising that weights obtained early in the training cycle produced similar levels of prediction accuracy to predictions based on the weights obtained in the last epoch before convergence. In future work, we will explore this technique which could improve the DA neural network design process by reducing the training time necessary to determine if a model and its hyper-parameter set are suitable for a complex dataset.

References

Beck, M. A.; Liu, C.-Y.; Bidinosti, C. P.; Henry, C. J.; Godee, C. M.; and Ajmani, M. 2020. An embedded system for the automated generation of labeled plant images to enable machine learning applications in agriculture. *PLOS ONE*, 15: 1–23.

Boubin, J.; Chumley, J.; Stewart, C.; and Khanal, S. 2019. Autonomic Computing Challenges in Fully Autonomous Precision Agriculture. In *2019 IEEE International Conference on Autonomic Computing (ICAC)*. IEEE.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. New York, NY, USA: ACM. ISBN 978-1-4503-4232-2.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE CVPR*.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Dhar, S.; and Shamir, L. 2021. Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks. *Visual Informatics*, 5(3): 92–101.

Domhan, T.; Springenberg, J. T.; and Hutter, F. 2015. Speeding up Automatic Hyperparameter Optimization of Deep Neural Networks by Extrapolation of Learning Curves. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 3460–3468. AAAI Press. ISBN 9781577357384.

Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 88: 303–338.

Fang, H.; and Liang, S. 2003. Retrieving leaf area index with a neural network method: Simulation and validation. *IEEE Transactions on Geoscience and Remote Sensing*, 41(9): 2052–2062.

Foreman, L. 2014. Characteristics and Production Costs of U.S. Corn Farms, Including Organic, 2010.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. *IEEE CVPR*.

Hughes, D. P.; and Salathe, M. 2015. An open access repository of images on plant health to enable the development of mobile disease diagnostics. In *arXiv*.

Koch, B.; Denton, E.; Hanna, A.; and Foster, J. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Lu, Y.; and Young, S. 2020. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture*, 178: 105760.

Mulla, D. J. 2013. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering*, 114(4): 358–371.

Mureșan, H.; and Oltean, M. 2018. Fruit recognition from images using deep learning. *Acta Universitatis Sapientiae, Informatica*, 10: 26–42.

Ponce, J.; Berg, T.; Everingham, M.; Forsyth, D.; Hebert, M.; Lazebnik, S.; Marszalek, M.; Schmid, C.; Russell, B.; Torralba, A.; Williams, C.; Zhang, J.; and Zisserman, A. 2007. *Dataset Issues in Object Recognition*, volume 4170, 29–48. ISBN 978-3-540-68794-8.

Simonyan, K.; and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *IEEE CVPR*. Los Alamitos, CA, USA: IEEE.

Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 97: 6105–6114.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528.

Unterthiner, T.; Keysers, D.; Gelly, S.; Bousquet, O.; and Tolstikhin, I. O. 2020. Predicting Neural Network Accuracy from Weights. *ArXiv*, abs/2002.11448.

Yamada, Y.; and Morimura, T. 2016. Weight features for predicting future model performance of deep neural networks. *IJCIA-16*.

Zhang, Z.; Khanal, S.; Raudenbush, A.; Tilmon, K.; and Stewart, C. 2022. Assessing the efficacy of machine learning techniques to characterize soybean defoliation from unmanned aerial vehicles. *Computers and Electronics in Agriculture*, 193: 106682.