REAL-WORLD CHALLENGES IN PEST DETECTION USING DEEP LEARNING: AN INVESTIGATION INTO FAILURES AND SOLUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning models have shown significant promise in pest detection tasks within controlled environments, but their performance often degrades when deployed in real-world agricultural settings. This study investigates the challenges hindering the generalization of these models, focusing on data quality issues, environmental variability, and inherent model limitations. Through extensive experiments, including learning rate optimization and multi-dataset training, we demonstrate that while lower learning rates can enhance generalization, models still struggle with robustness to environmental changes. Our findings highlight critical pitfalls in deploying deep learning models for pest detection and offer insights into potential solutions for improving their real-world applicability.

Comment: The rest of the paper primarily focuses on environmental variability and noise in images.

Comment: This might be setting an unnecessarily high bar.

1 INTRODUCTION

Accurate pest detection is crucial for protecting crops and ensuring global food security. Deep learning models have emerged as powerful tools for automating pest detection tasks, achieving high accuracy in controlled environments. However, their performance often degrades significantly when deployed in real-world agricultural settings . This gap between controlled experiments and real-world applications poses a serious challenge for precision agriculture and highlights the need for robust, generalizable models. Understanding and addressing the reasons behind these performance drops is essential for advancing AI in agriculture.

Comment: Both sentences require citations to be substantiated.

In this work, we investigate the factors contributing to the failures of deep learning models in real-world pest detection scenarios. We hypothesize that issues such as data quality, environmental variability, and inherent model limitations play significant roles in hindering model generalization. Through a series of experiments, we explore these challenges in depth. Our findings reveal that while optimizing hyperparameters, such as the learning rate, can lead to improved validation accuracy, deep learning models still struggle to maintain robustness under environmental changes. Moreover, multi-dataset training and domain adaptation techniques, aimed at enhancing generalization across different datasets, present their own set of challenges, including increased computational demands and inconsistent performance gains.

Comment: Is it really in depth?

Comment: The domain adaptation experiment in Section 5.2 appears to highlight the challenges of adapting ImageNet-trained models to other vision datasets when environmental noise is present, but this experiment isn't directly related to the main focus of the paper.

By presenting these negative and inconclusive results, we aim to highlight the real-world pitfalls and challenges in deploying deep learning models for pest detection. Our research provides valuable insights for the agricultural and machine learning communities, contributing to the development of more reliable AI tools for precision farming.

2 RELATED WORK

Deep learning has been widely applied in agricultural contexts for tasks such as pest and disease detection, showing high accuracy in controlled settings (Mustakim et al., 2024; Kumar et al., 2022). Li et al. (2023b) highlighted the limitations of traditional deep learning methods in practical applications, noting issues such as overfitting and sensitivity to environmental variations. Several studies have explored methods to improve model robustness and generalization. Data augmentation techniques have been employed to enhance dataset diversity and reduce overfitting (Abdulkareem

et al., 2024). Domain adaptation strategies have been proposed to address domain shifts and improve performance in new environments (Prasad & Agniraj, 2024; Li et al., 2023a). However, these approaches often do not fully address the challenges faced in real-world deployment.

Reviews like Teixeira et al. (2023) and Hu (2023) have identified gaps in current research, emphasizing the need for models that generalize well to diverse, real-world conditions. Additionally, Amir et al. (2024) discussed the limitations of deep learning models when encountering out-of-distribution inputs, underscoring the importance of verifying model generalization. Our work distinguishes itself by focusing on the failures and limitations of deep learning models in real-world pest detection scenarios, providing an in-depth investigation into the underlying causes and proposing insights for improvement.

3 Methodology

We employed deep learning models for pest detection, focusing on evaluating their performance and robustness in real-world agricultural settings. We utilized the ResNet-18 architecture (**?**), pretrained on ImageNet , and fine-tuned it on the Crop Pest and Disease dataset, which includes 22 classes of pests and diseases collected from local farms. To investigate the challenges, we designed experiments to assess the impact of learning rates on model performance. We hypothesized that optimizing the learning rate could improve generalization. We also implemented data augmentation techniques to simulate environmental variability, such as brightness and contrast changes, Gaussian blur, and random affine transformations, to evaluate the models' robustness. Additionally, we explored multi-dataset training using datasets such as EuroSAT (**?**), MedMNIST (**?**), and CIFAR-10 (**?**) to assess the potential of domain adaptation and transfer learning in improving model generalization across different agricultural domains.

Comment: Cite. Also would be nice to show some example images in the appendix.

Comment: EuroSAT might make sense but using MedMNIST and CIFAR-10 to "assess the potential of domain adaptation...across different agricultural domains" is a stretch b/c these two are clearly not from the agricultural domain.

Comment: The citation key 'he2016deep' exists in the latex file but not in references.bib.

Comment: All three datasets citation key exist in the latex file but not in the bib file.

4 Experimental Setup

The Crop Pest and Disease dataset comprises 25,126 images across 22 classes of pests and diseases affecting crops such as cashew, cassava, maize, and tomato. We split the dataset into training (70%), validation (15%), and testing (15%) sets. For the baseline experiments, we conducted a grid search to optimize the learning rate, evaluating values of $\{1e^{-4}, 5e^{-4}, 1e^{-3}, 5e^{-3}, 1e^{-2}\}$. We trained the ResNet-18 model for 10 epochs for each learning rate, using a batch size of 32 and the Adam optimizer. To simulate challenging environmental conditions, we applied data augmentations during testing, including brightness and contrast adjustments, Gaussian blur, and random affine transformations. We introduced the Environmental Robustness Score (ERS), calculated as the ratio of model accuracy under challenging conditions to that under normal conditions, to quantify robustness.

In the research experiments, we trained models on additional datasets—EuroSAT, MedMNIST, and CIFAR-10—to investigate the effects of multi-dataset training on model generalization. We used similar training settings and evaluated models using accuracy, loss, and ERS.

Comment: used a subset (around 2500 images)

5 Experiments and Results

5.1 Impact of Learning Rate on Model Performance

To evaluate the impact of learning rates on model performance, we trained the ResNet-18 model on the Crop Pest and Disease dataset using different learning rates. Figure 1 illustrates the aggregated accuracy, loss, and ERS across different learning rates.

As shown in Figure 1, lower learning rates ($1e^{-4}$ and $5e^{-4}$) result in smoother convergence of training and validation accuracy, and a steady decrease in training loss. The ERS remains more stable for these learning rates, suggesting enhanced robustness to environmental variability. In contrast, higher learning rates lead to overfitting and unstable loss patterns, with significant fluctuations in ERS scores. These results indicate that optimizing hyperparameters like learning rate is crucial for improving model generalization and robustness in real-world settings.

Comment: The figure seems to suggest the lower lr group overfits more

Comment: only in this specific condition and definition of environmental robustness

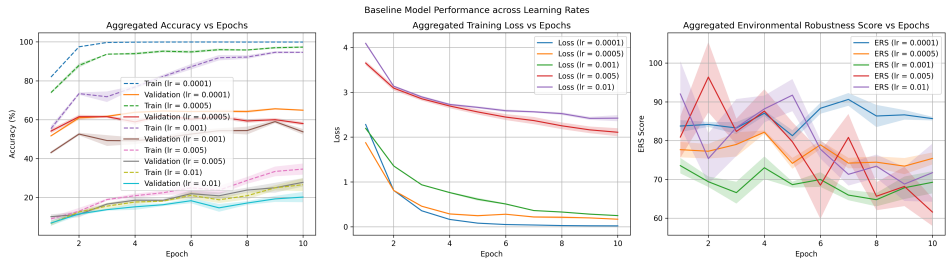

Figure 1: Baseline model performance across learning rates. Aggregated training and validation accuracy, training loss, and Environmental Robustness Score (ERS) over epochs for different learning rates. Lower learning rates yield higher validation accuracy and more stable ERS scores, indicating better generalization and robustness.

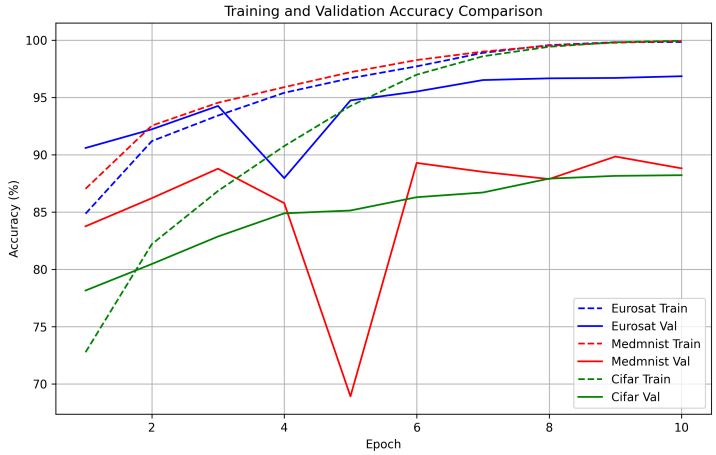

Figure 2: Comparison of training and validation accuracy across different datasets. Models trained on EuroSAT and CIFAR-10 exhibit stable and high accuracy, whereas the model trained on MedMNIST shows erratic accuracy patterns, indicating challenges in generalization due to domain discrepancies.

Comment: Should explain that this is more about generalization from ImageNet to EuroSAT, MedMNIST, or CIFAR.

5.2 Challenges in Multi-Dataset Training

To investigate model generalization across different domains, we trained the ResNet-18 model on additional datasets: EuroSAT, MedMNIST, and CIFAR-10. Figure 2 presents the comparison of training and validation accuracy across these datasets.

From Figure 2, the models trained on EuroSAT and CIFAR-10 achieve high and stable training and validation accuracy over epochs, suggesting effective learning and better generalization. In contrast, the MedMNIST model displays fluctuating accuracy, highlighting difficulties in adapting to the pest detection task. This suggests that significant domain shifts can negatively impact learning, leading to overfitting and decreased robustness.

Comment: a stretched argument. Difficulty of adopting ResNet-18 to MedMNIST doesn't immediately suggest it's also difficult to adapting to the pest detection task. At least it needs more explanation.

To further examine the robustness of these models, we analyzed the ERS across epochs, as shown in Figure 3.

Figure 3 illustrates that the models trained on EuroSAT and CIFAR-10 maintain higher ERS scores across epochs, suggesting robustness to environmental augmentations applied during testing. The MedMNIST model's low ERS scores indicate vulnerability to such changes, underscoring the challenges posed by domain differences. These findings highlight the varying impact of dataset characteristics on model generalization and robustness. While multi-dataset training and domain adapta-

Comment: Should refer to Appendix B, as it provides more detailed information on the noise used

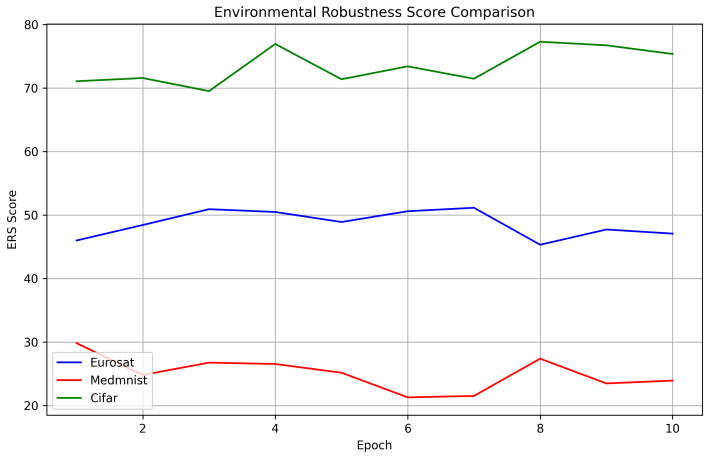

Figure 3: Environmental Robustness Score (ERS) comparison across different datasets. The EuroSAT and CIFAR-10 models maintain higher and more stable ERS scores, indicating better robustness to environmental variability. The MedMNIST model shows low and unstable ERS scores, reflecting sensitivity to environmental changes.

Comment: The legend can be improved

tion can offer potential improvements, they also present challenges, such as increased computational demands and inconsistent performance gains, which must be carefully managed.

6 Discussion

Our experiments reveal significant challenges in deploying deep learning models for pest detection in real-world agricultural settings. Optimizing hyperparameters like learning rates enhances model generalization and robustness to some extent, as evidenced by the improved performance and stable ERS scores at lower learning rates. However, models still struggle with environmental variability, indicating that hyperparameter optimization alone is insufficient to achieve robust real-world performance.

The exploration of multi-dataset training provides valuable insights into domain adaptation challenges. The varying performance across datasets underscores the importance of dataset selection and the potential pitfalls of naively combining datasets with different characteristics. The model's poor performance on MedMNIST suggests that significant domain shifts can negatively impact learning, leading to overfitting and decreased robustness.

Comment: This can be misleading, since multi-dataset training could mean a model trained on multiple datasets, but here multiple models are trained, where each model is trained on a single dataset

Comment: This only makes sense if "combining" refers to ImageNet pretraining followed by fine-tuning on a different dataset, but it usually gives the impression that the model is trained on multiple datasets.

These findings emphasize the need for specialized strategies to address data quality issues and environmental variability. Data augmentation techniques that more accurately reflect real-world conditions, robust training methods, and domain-specific model adaptations may be necessary to improve model performance in practical applications.

7 Conclusion

Our study highlights critical pitfalls in deploying deep learning models for pest detection in real-world agricultural settings. While optimizing hyperparameters like learning rates can enhance model generalization and robustness, challenges remain due to environmental variability and domain discrepancies. Multi-dataset training introduces additional complexities, and its benefits depend on the compatibility of the datasets involved. Future work should focus on developing advanced techniques tailored to real-world conditions, such as improved data augmentation strategies that mimic environmental changes, robust training methods that enhance model resilience, and architectures designed for adaptability. By addressing these challenges, we can move closer to deploying reliable AI tools in precision agriculture that are resilient to real-world variability.

REFERENCES

Ismael M. Abdulkareem, Faris K. Al-Shammri, Noor Aldeen A. Khalid, and Natiq A. Omran. Proposed approach for object detection and recognition by deep learning models using data augmentation. *Int. J. Online Biomed. Eng.*, 20:31–43, 2024.

Guy Amir, Osher Maayan, Tom Zelazny, Guy Katz, and Michael Schapira. Verifying the generalization of deep learning to out-of-distribution domains. *ArXiv*, abs/2406.02024, 2024.

Jiangfeng Hu. Application of deep learning in smart agriculture research. *Applied and Computational Engineering*, 2023.

Raj Kumar, Dinesh Singh, A. Chug, and A. Singh. Evaluation of deep learning based resnet-50 for plant disease classification with stability analysis. In *International Conference Intelligent Computing and Control Systems*, pp. 1280–1287, 2022.

A. Li, Elisa Bertino, Rih-Teng Wu, and Ting Wu. Building manufacturing deep learning models with minimal and imbalanced training data using domain adaptation and data augmentation. *2023 IEEE International Conference on Industrial Technology (ICIT)*, pp. 1–8, 2023a.

Manzhou Li, Siyu Cheng, Jingyi Cui, Changxiang Li, Zeyu Li, Chang Zhou, and Chunli Lv. High-performance plant pest and disease detection based on model ensemble with inception module and cluster algorithm. *Plants*, 12, 2023b.

M. Mustakim, Aditya Rezky Pratama, Imam Ahmad, Teguh Arifianto, Kelik Sussolaikah, and Sepriano Sepriano. Image classification of corn leaf diseases using cnn architecture resnet-50 and data augmentation. In *2024 International Conference on Decision Aid Sciences and Applications (DASA)*, pp. 1–6, 2024.

Pulicherla Siva Prasad and Senthilrajan Agniraj. Cross-domain adaptation techniques for robust plant disease detection: A dann-coral hybrid approach. *International Journal of Experimental Research and Review*, 2024.

A. Teixeira, José Ribeiro, R. Morais, J. Sousa, and António Cunha. A systematic review on automatic insect detection using deep learning. *Agriculture*, 2023.

SUPPLEMENTARY MATERIAL

A ADDITIONAL FIGURES AND DETAILED RESULTS

We provide additional figures and detailed results to supplement the main text. These figures offer deeper insights into the models' behaviors under different experimental conditions.

Figure 4 shows that the EuroSAT and CIFAR-10 models show consistent decreases in training loss, reflecting effective learning. The MedMNIST model's erratic loss suggests that the model struggles to minimize the loss function, possibly due to significant differences between medical images and agricultural pest images.

Comment: This should be ImageNet images

B IMPLEMENTATION DETAILS

All models were implemented using PyTorch 1.9.0. The ResNet-18 architecture was initialized with ImageNet pretrained weights. For optimization, we used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of $1e^{-4}$. No learning rate schedules or gradient clipping were applied.

Comment: weight decay is 0.01 instead of $1e^{-4}$.

Data augmentations for simulating challenging conditions included `ColorJitter` with brightness and contrast factors of 0.5, `GaussianBlur` with a kernel size of 3, and `RandomAffine` transformations with degrees up to 15 and translation up to 10%. These augmentations were applied during testing to evaluate the Environmental Robustness Score (ERS).

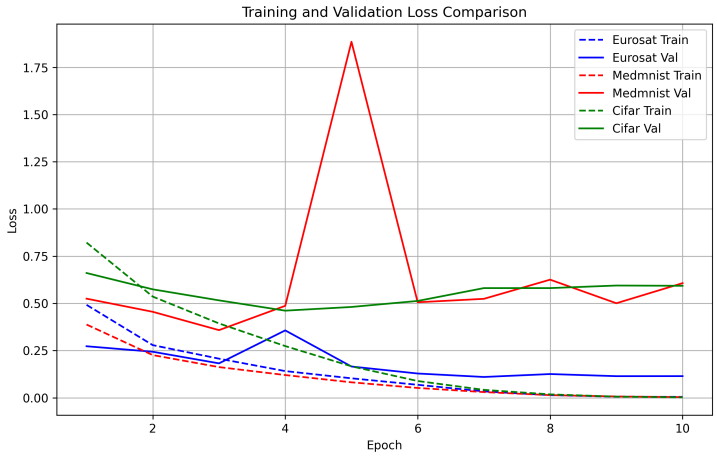

Figure 4: Comparison of training loss across different datasets. Models trained on EuroSAT and CIFAR-10 datasets demonstrate a steady decrease in loss, while the model trained on MedMNIST exhibits erratic loss curves, indicating instability during training due to domain mismatch.

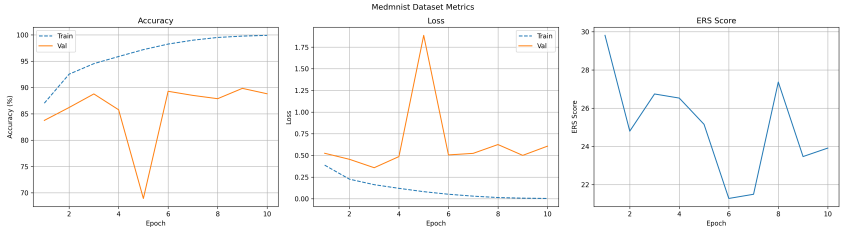

Figure 5: Performance metrics for the model trained on MedMNIST dataset. The erratic behavior in accuracy and loss indicates challenges in model convergence and generalization when applying the MedMNIST dataset to pest detection tasks.

Comment: This statement is incorrect again. It should something like "...applying/using the MedMNIST dataset with the ImageNet-pretrained model." Also, the figure isn't mentioned in the main text.

The Environmental Robustness Score (ERS) is defined as:

$$\text{ERS} = \frac{\text{Accuracy under challenging conditions}}{\text{Accuracy under normal conditions}} \tag{1}$$

This metric quantifies the model's robustness to environmental changes by comparing its performance under augmented test sets to that under standard conditions.

The additional datasets used for multi-dataset training were:

- **EuroSAT**: A dataset consisting of 27,000 labeled Sentinel-2 satellite images covering 10 classes (**?**).
- **MedMNIST**: A collection of lightweight medical image datasets covering various tasks (**?**).
- **CIFAR-10**: A well-known dataset consisting of 60,000 32x32 color images in 10 classes (**?**).

For the multi-dataset training, we used a batch size of 64 to accommodate the increased data volume. Training was conducted for 30 epochs, and early stopping was applied if validation loss did not decrease for 5 consecutive epochs.