# Addressing domain shift in deep learning: Challenges and insights from plant disease diagnosis and flower recognition

Jianqiang Sun[1,*]

[1] Research Center for Agricultural Information Technology, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, Japan
[*] Corresponding author

**Keywords:** deep learning, generalization performance, domain shift, plant disease diagnosis, flower classification, data diversity

## Abstract

Deep learning models have demonstrated potential in plant disease diagnosis and flower recognition; however, their real-world application is hindered by domain shift—where models trained on one dataset underperform when tested on datasets collected in different conditions. This study assessed the cross-dataset generalization of deep learning models, including ResNet, EfficientNet, Inception, and MobileNet, using multiple tomato disease and flower datasets. While models achieved high accuracy when trained and validated on the same dataset, their performance declined significantly when evaluated on different datasets, underscoring the impact of domain variability. This performance degradation was consistent across all model architectures, suggesting that domain shift cannot be addressed through model architecture alone. The results emphasize the need for more diverse and representative datasets to mitigate variability in agricultural data and enhance the practical utility of deep learning models in agricultural applications.

## Introduction

Machine learning, particularly deep learning, has shown great potential in addressing complex real-world challenges, including agricultural applications. However, a key barrier to its broader adoption is domain shift—the issue where models trained under specific conditions fail to perform well when applied to new, unseen conditions (Quinonero-Candela et al., 2022). While domain shift has been extensively studied in other fields (Kouw and Loog, 2021; Subbaswamy and Saria, 2020), its impact on agricultural research, especially in plant disease and pest (PDP) diagnosis using image recognition, remains underexplored. This gap raises critical concerns regarding the generalizability of machine learning models in real-world agricultural contexts, which often differ significantly from controlled research settings.

PDPs pose a severe threat to global agriculture, accounting for an estimated 20-40% annual crop loss in major staples such as wheat, rice, maize, potatoes, and soybeans (Savary et al., 2019). Early and accurate detection of PDPs is crucial for preventing yield losses and curbing their spread, thereby ensuring a sustainable food supply. This urgent need has spurred the development of advanced diagnostic tools that assist farmers without requiring specialized expertise.

Recently, deep learning-based image recognition has gained prominence as a diagnostic tool for PDPs (Liu and Wang, 2021; Shoaib et al., 2023; Sun et al., 2024). However, most studies rely on controlled datasets (Abbas et al., 2021; Yu et al., 2023; Zhang et al., 2022) or self-collected data from limited field conditions (Huang et al., 2022; Shoaib et al., 2022; Xu et al., 2022). For instance, the widely used PlantVillage dataset (Hughes and Salathé, 2015) contains images taken against uniform backgrounds, which fail to capture the variability found in real-world agricultural environments. In practice, field conditions are far more complex, involving variations in lighting, plant orientation, and background elements. Similarly, self-collected datasets often focus on specific pests or diseases in controlled environments, further restricting the models' ability to generalize across broader conditions.

These controlled datasets do not adequately reflect the complexities of real-world environments, leading to inflated performance metrics that do not translate well outside laboratory settings. This gap highlights the need to assess the generalization capabilities of models trained on controlled datasets when applied to practical agricultural scenarios. Despite the crucial role of PDP diagnosis in global food security, no comprehensive studies have systematically evaluated the generalization performance of deep learning models across diverse, real-world agricultural data.

To address this gap, the current study evaluates the cross-dataset generalization of deep learning models using multiple datasets, including those collected under both controlled and field conditions, for image recognition-based PDP diagnosis. Additionally, a comparative analysis is conducted using flower classification tasks, which

typically present fewer complexities than PDP diagnosis. The variability in symptoms based on disease stage and other factors makes PDP diagnosis particularly challenging, often leading to misidentifications. By contrasting these tasks, this study aims to provide insights into the practical applicability of these models, focusing on their performance in real-world conditions beyond controlled environments.

## Materials and Methods

### *Data preprocessing of tomato disease datasets*

Image-based plant disease diagnosis has primarily focused on tomato plants due to the availability of numerous datasets. For this study, tomato disease image datasets were selected based on their accessibility and relevance. The datasets include PlantVillage (PV) (Hughes and Salathé, 2015), Tomato-Village (TV) (Gehlot et al., 2023), PlantDoc (Singh et al., 2020), and Tomato Leaf Disease (TLD) (Bryan, 2023), all of which are publicly available in online repositories.

The PV dataset is a seminal resource in plant disease image diagnostics and is widely used in the development of deep learning algorithms. It contains images of individual leaves, detached from the plant, and photographed against a uniform background. This dataset features 38 categories of plant species and diseases, 10 of which are relevant to tomato diseases and used in this study: healthy (HL), bacterial spot (BS), early blight (EB), late blight (LB), leaf mold (LM), tomato mosaic virus (MV), spider mites (SM), Septoria leaf spot (SS), target spot (TS), and yellow leaf curl virus (YV).

The TV dataset consists of a majority of images similar to those in the PV dataset, with isolated tomato leaves photographed against a uniform background while a smaller subset of the images is captured in field conditions. This dataset includes five categories of tomato diseases (HL, EB, LB, leaf miner (MN), and spotted wild virus (SW)) and three categories of physiological disorders (potassium deficiency, magnesium deficiency, and nitrogen deficiency). Physiological disorders were excluded from this study as they represent nutritional imbalances requiring different diagnostic approaches and do not classify as diseases.

The PlantDoc and TLD datasets contain images from various sources, including PV, web materials, digitized presentation slides, and field-captured images. Due to the significant overlap between these two datasets, they were merged and treated as a single dataset. Then, images sourced originally from the PV dataset were excluded to avoid redundancy. The final categories in this merged dataset included HL, BS, EB, LB, LM, MN, MV, SM, SS, and YV.

To maintain consistency, categories with fewer than 50 images were removed from each dataset to ensure adequate representation for model training and to improve the reliability of the comparative analysis. Additionally, to visualize the image distribution across datasets, uniform manifold approximation and projection

(UMAP) (McInnes et al., 2018) was employed. Fifty images from each category in each dataset were randomly sampled and resized to 224 × 224 pixels. These images were transformed using principal component analysis (PCA), with the top 128 principal components serving as input to UMAP for dimensionality reduction to two dimensions. UMAP parameters were set to `n_neighbors = 10` and `min_dist = 0.1`.

***Data preprocessing of flower datasets***

Three flower datasets were utilized in this study: the TensorFlow Flowers Dataset (TF) (Team, 2019), the 102 Category Flower Dataset (F102) (Nilsback and Zisserman, 2008), and the YNU Flower Dataset (YNU) (Yu et al., 2024). The TF dataset consists of five categories—daisy, dandelion, rose, sunflower, and tulip—with most images sourced from websites. The YNU dataset includes 11 categories of commercially important flowers, all photographed under controlled conditions. Since the images in the YNU dataset have a resolution of several thousand pixels, they were resized to a short side of 500 pixels to match the input requirements of the deep learning models, which typically use images with a resolution of a few hundred pixels. The F102 dataset contains 102 flower categories commonly found in the United Kingdom, with most images taken in natural environments.

For this study, only flower categories that appeared in at least two datasets and contained more than 50 images—including buttercup, carnation, daisy, dandelion, rose, and sunflower—were selected for model training and validation. Additionally, UMAP visualizations were applied to the flower datasets following the same procedure used for the tomato disease datasets. UMAP parameters were set to `n_neighbors = 20` and `min_dist = 0.01` for the flower datasets.

***Model training and validation***

Four deep learning architectures—ResNet (He et al., 2016), Inception v3 (Szegedy et al., 2016), MobileNet v3 (Howard et al., 2019), and EfficientNet v2 (Tan and Le, 2021)—were used to evaluate cross-dataset generalization for plant disease diagnosis and flower recognition. The evaluations were conducted across all combinations of these architectures and datasets.

For instance, to evaluate ResNet using the PV dataset, the model was trained on 80% of the images randomly sampled from PV and validated on the remaining 20%. To assess cross-dataset generalization, the trained model was also validated on 20% of images randomly sampled from other datasets (e.g., TV and TLD). Each evaluation process was repeated five times to ensure robust performance estimates, and the average performance metrics were calculated for each combination of architecture and dataset. Confusion matrices were generated for ResNet models trained on the PV dataset and validated on PV, TV, and TLD, based on the first run of validation. Additionally, Grad-CAM (Selvaraju et al., 2017) was applied to several HL images in the TV and TLD datasets that were incorrectly classified as LB by the ResNet model trained on the PV dataset.

All model architectures were implemented using predefined classes from the PyTorch library (Paszke et al., 2019). Specifically, ResNet, Inception v3, MobileNet v3, and EfficientNet v2 were instantiated using the `resnet18`, `inception_v3`, `mobilenet_v3_small`, and `efficientnet_v2_s` classes from the `torchvision.models` module, respectively. The initial weights for these architectures were set to the default pretrained weights from the ImageNet dataset (Jia et al., 2009).

### *Performance metrics*

Model performance was evaluated using the macro averages of precision, recall, and F1 score. These metrics were defined as follows: precision = TP / (TP + FP), recall = TP / (TP + FN), and F1 score = (2 × precision × recall) / (precision + recall), where TP, FP, and FN represent true positive, false positive, and false negative counts, respectively. Using macro averages ensures that each class is weighted equally, providing a balanced assessment of model performance across all categories.

## Results and Discussion

### *Dataset characteristics*

The PV dataset contains significantly more images than the other tomato disease datasets, while TV and TLD have relatively fewer images (**Table 1**). In terms of disease category coverage, all datasets, except for TV, encompass a broad range of disease categories. Only three categories—healthy (HL), early blight (EB), and late blight (LB)—are common across all datasets.

UMAP visualizations revealed that tomato disease images were clearly separated by dataset but not well differentiated by disease type (**Figure 1A**). This illustrates the issue of domain shift, where models trained on images captured under consistent conditions may generalize poorly when tested on images from varying conditions. The lack of distinction by disease type likely stems from the similar compositions within each dataset. For example, PV images often depict a single leaf on a monochromatic background, which likely becomes the dominant feature captured by PCA and UMAP. In contrast, the subtle differences in disease symptoms are overshadowed by the overall image composition, making it more challenging to distinguish disease types.

For the flower datasets, the number of images per category is summarized in **Table 2**. Three categories—daisy, rose, and sunflower—are present in all three datasets. The YNU dataset contains over 10,000 images, significantly more than the other two datasets, while the F102 dataset, despite covering all six categories, has the fewest images. UMAP visualizations for the flower datasets indicated that the TF and F102 datasets had similar distributions, while YNU, though adjacent, displayed a distinct pattern (**Figure 1B**). This reflects the

different image environments: TF and F102 images were captured in natural settings with complex backgrounds, whereas YNU images were taken under controlled conditions.

### *Cross-dataset generalization performance of plant disease diagnosis models*

Performance metrics such as precision, recall, and F1-score for the plant disease diagnosis models are shown in **Figures 2A-C**. As expected, the models performed best when the training and validation datasets were the same. However, performance dropped significantly when the validation dataset differed from the training dataset. For instance, ResNet, when trained on the PV dataset—which contains simple images with uniform backgrounds—achieved a high F1-score of 0.987 when validated on PV. In contrast, when validated on the TV and TLD datasets, the F1-scores plummeted to 0.012 and 0.130, respectively, even though the TV dataset includes many images similar to PV, featuring isolated tomato leaves under controlled conditions. Conversely, models trained on TV and TLD also struggled when validated on PV, highlighting the bidirectional nature of domain shift.

Training models on a combined dataset of PV, TV, and TLD (referred to as ALL) mitigated performance drops when validated on individual datasets, although declines were still evident compared to models trained on a single dataset. Notably, performance deterioration was minimal when models trained on ALL were validated on PV, likely because the ALL dataset included a high proportion of PV images.

A deeper analysis of the confusion matrices provided further insights into these performance gaps. ResNet, when trained on PV, accurately identified diseases in PV images (**Figure 2D**). However, when validated on TV, a large percentage of images (85.5%) in the MN and SW categories—absent from the training data—were misclassified into other categories, particularly BS, EB, and LB (**Figure 2E**). Even after excluding the MN and SW categories, performance remained poor (F1-score = 0.092), primarily due to the misclassification of healthy (HL) images into BS, EB, LB, LM, and YV categories. Similarly, when ResNet trained on PV was validated on TLD, MN images were frequently misclassified as EB and LB (**Figure 2F**), paralleling the misclassification trends observed with TV. Common categories across TV and TLD also displayed significant misclassification, such as most HL images being predicted as LB, and LM images frequently misclassified as EB, LB, or SS. Additionally, frequent misclassifications were observed between EB, LB, and SS categories.

A particularly noteworthy trend was the consistent misclassification of HL images as LB across different validation datasets (**Figures 2E-F**). Grad-CAM visualizations suggested that ResNet trained on PV focused on the salient areas of the leaf (**Figure 3**), but often misclassified them. This may be due to physiological wilting in healthy leaves mimicking the dry and wilted symptoms of LB. Misclassification into categories other than LB (e.g., EB, LM, MV, SS, TS) occurred less frequently, likely because those diseases display symptoms across the entire leaf, unlike the localized wilting typical of LB.

Models trained on PV struggled with the complexity of field images in TLD, which is understandable given the controlled conditions of PV images. However, the fact that models trained on TLD also failed to generalize to PV images suggests that complex training data alone does not resolve domain shift issues. These trends were consistent across all deep learning architectures tested, including ResNet, EfficientNet, Inception, and MobileNet, indicating that domain shift affects performance regardless of the architecture used.

***Cross-dataset generalization performance of flower recognition***

The flower classification models, which dealt with objects characterized by distinct colors and shapes, exhibited similar trends to plant disease models, albeit with less pronounced performance declines. For instance, ResNet, when trained on the F102 flower dataset, achieved an F1-score of 0.966 during validation on F102. However, it experienced a moderate decrease in performance, dropping to 0.395 and 0.457 when validated on the TF and YNU datasets, respectively (**Figure 4**). These patterns were consistent across other architectures, including EfficientNet, Inception, and MobileNet, highlighting the widespread impact of domain shift. Nonetheless, the domain shift was more significant in plant disease diagnosis than in flower recognition, probably recognizing a flower is easier than recognizing a lesion spot on leaves.

***Challenges in dataset availability and accessibility***

The difficulties encountered in cross-dataset generalization underscore the broader challenges of deploying deep learning models in real-world agricultural environments, where data variability is inevitable. Mitigating domain shift requires access to larger and more diverse datasets; however, the efforts to obtain additional tomato disease datasets were hindered by limited availability. A survey of existing tomato disease datasets surprisingly observed that many researchers continue to rely on simple and controlled images (e.g., PV) for algorithm development. Only a few studies utilized more complex, field-collected data that could better represent the realities of agricultural environments, and many of which remain unpublished or inaccessible. Despite frequent claims in Data Availability statements that "datasets are available from the corresponding author upon reasonable request," the requests for access were unanswered, even for recently published studies. This limited access to diverse, field-derived datasets hinders progress in agricultural artificial intelligence research.

The development and sharing of practical, large-scale datasets that are easily accessible to researchers would significantly accelerate advancements in agricultural artificial intelligence, enhance food security, and support sustainable farming practices. However, issues related to data privacy and the high costs of data collection pose substantial barriers to widespread data sharing. Even when data is collected at considerable expense, making it freely available presents an additional challenge. Overcoming these obstacles and facilitating the release of high-quality agricultural datasets will be crucial to driving innovation and research in agricultural artificial intelligence.

## Conclusion

This study underscores the significant impact of domain shift on the cross-dataset generalization of deep learning models used for plant disease diagnosis and flower recognition. Although models performed well when trained and validated on the same dataset, their effectiveness declined markedly when tested on different datasets, revealing limited generalizability. These findings were consistent across various deep learning architectures, confirming that domain shift is a widespread issue that cannot be resolved by merely altering the model architecture. This highlights the challenges posed by domain variability in real-world agricultural applications. To improve cross-dataset generalization and enhance the practical use of deep learning in agriculture, future efforts should prioritize the collection, sharing, and public availability of diverse datasets. Addressing these challenges will not only advance agricultural artificial intelligence research but also promote the development of sustainable agricultural solutions.

## References


Abbas, A., Jain, S., Gour, M., Vankudothu, S., 2021. Tomato plant disease detection using transfer learning with C-GAN synthetic images. Comput. Electron. Agric. 187, 106279.

Bryan, 2023. Tomato Leaf Disease Dataset. Roboflow Universe.

Gehlot, M., Saxena, R.K., Gandhi, G.C., 2023. "Tomato-Village": a dataset for end-to-end tomato disease detection in a real-world environment. Multimed. Syst. 29, 3305–3328.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 770–778.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., Le, Q., 2019. Searching for MobileNetV3, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, pp. 1314–1324.

Huang, M.-L., Chuang, T.-C., Liao, Y.-C., 2022. Application of transfer learning and image augmentation technology for tomato pest identification. Sustainable Computing: Informatics and Systems 33, 100646.

Hughes, D.P., Salathé, M., 2015. An open access repository of images on plant health to enable the development of mobile disease diagnostics. ArXiv.

Jia, D., Wei, D., Richard, S., Li, L.-J., Li, K., Li, F.-F., 2009. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvprw.2009.5206848

Kouw, W.M., Loog, M., 2021. A Review of Domain Adaptation without Target Labels. IEEE Trans. Pattern Anal. Mach. Intell. 43, 766–785.

Liu, J., Wang, X., 2021. Plant diseases and pests detection based on deep learning: a review. Plant Methods 17, 22.

McInnes, L., Healy, J., Saul, N., Großberger, L., 2018. UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. 3, 861.

Nilsback, M.-E., Zisserman, A., 2008. Automated flower classification over a large number of classes, in: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. Presented at the Image Processing (ICVGIP), IEEE, pp. 722–729.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E.B., Garnett, R. (Eds.), NeurIPS. pp. 8024–8035.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D., 2022. Dataset shift in machine learning, Neural Information Processing series. MIT Press, London, England.

Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N., Nelson, A., 2019. The global burden of pathogens and pests on major food crops. Nat. Ecol. Evol. 3, 430–439.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV). Presented at the 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, pp. 618–626.

Shoaib, M., Hussain, T., Shah, B., Ullah, I., Shah, S.M., Ali, F., Park, S.H., 2022. Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease. Front. Plant Sci. 13, 1031748.

Shoaib, M., Shah, B., Ei-Sappagh, S., Ali, A., Ullah, A., Alenezi, F., Gechev, T., Hussain, T., Ali, F., 2023. An advanced deep learning models-based plant disease detection: A review of recent research. Front. Plant Sci. 14, 1158933.

Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N., 2020. PlantDoc: A Dataset for Visual Plant Disease Detection, in: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, CoDS COMAD 2020. Association for Computing Machinery, New York, NY, USA, pp. 249–253.

Subbaswamy, A., Saria, S., 2020. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics 21, 345–352.
Sun, J., Cao, W., Fu, X., Ochi, S., Yamanaka, T., 2024. Few-shot learning for plant disease recognition: A review. Agronomy Journal 116, 1204–1216.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. cv-foundation.org, pp. 2818–2826.
Tan, M., Le, Q.V., 2021. EfficientNetV2: Smaller Models and Faster Training, in: International Conference on Machine Learning.
Team, T.T., 2019. Flowers. URL http://download.tensorflow.org/example_images/flower_photos.tgz
Xu, M., Yoon, S., Fuentes, A., Yang, J., Park, D.S., 2022. Style-Consistent Image Translation: A Novel Data Augmentation Paradigm to Improve Plant Disease Recognition. Front. Plant Sci. 12. https://doi.org/10.3389/fpls.2021.773142
Yu, M., Sun, Y., Ma, R., Mu, H., Liu, D., Li, Z., 2024. YNU Flower: A market-oriented benchmark dataset for flower classification. Sci. Hortic. (Amsterdam) 338, 113510.
Yu, S., Xie, L., Huang, Q., 2023. Inception convolutional vision transformers for plant disease identification. Internet of Things 21, 100650.
Zhang, L., Zhou, G., Lu, C., Chen, A., Wang, Y., Li, L., Cai, W., 2022. MMDGAN: A fusion data augmentation method for tomato-leaf disease identification. Appl. Soft Comput. 123, 108969.

## Availability of data and materials

The PlantVillage dataset is accessible at https://github.com/spMohanty/PlantVillage-Dataset. The Tomato-Village dataset can be found at https://github.com/mamta-joshi-gehlot/Tomato-Village. The Tomato Leaf Disease dataset is available at https://universe.roboflow.com/bryan-b56jm/tomato-leaf-disease-ssoha/dataset/63. The TensorFlow Flowers dataset is downloadable from https://download.tensorflow.org/example_images/flower_photos.tgz. The 102-category flower dataset is accessible at https://www.robots.ox.ac.uk/~vgg/data/flowers/102/. The YNU Flower dataset is available at https://doi.org/10.5281/zenodo.12525071.

## Author contributions

JS: Conceptualization, supervision, funding acquisition, investigation, data curation, methodology, formal analysis, validation, writing – original draft, writing – review & editing.

## Acknowledgments

Part of the computations were conducted on the SHIHO supercomputer at the National Agriculture and Food Research Organization.

## Competing interests

The author declares no competing interests.

## Funding

This research was supported by JSPS KAKENHI Grant Number 22H05179.

# Tables

**Table 1**

**Number of images in plant disease and pest datasets.** Values represent the number of images per category for each of the PlantVillage (PV), Tomato-Village (TV), and Tomato Leaf Disease (TLD) datasets used for model evaluation in this study.

| category | PV | TV | TLD |
|---|---|---|---|
| healthy (HL) | 1,591 | 226 | 212 |
| bacterial spot (BS) | 2,127 | - | 110 |
| early blight (EB) | 1,000 | 62 | 423 |
| late blight (LB) | 1,909 | 125 | 350 |
| leaf mold (LM) | 952 | - | 290 |
| leaf miner (MN) | - | 1,824 | 380 |
| mosaic virus (MV) | 373 | - | 116 |
| two spotted spider mites (SM) | 1,676 | - | 54 |
| Septoria leaf spot (SS) | 1,771 | - | 288 |
| spotted wild virus (SW) | - | 611 | - |
| target spot (TS) | 1,404 | - | - |
| yellow leaf curl virus (YV) | 5,357 | - | 145 |
| Total | 18,160 | 2,848 | 2,368 |
| Categories | 10 | 5 | 10 |

**Table 2**

**Number of images in flower datasets.** Values represent the number of images per category for each of the TensorFlow Flower (TF), 102 Flower (F102), and YNU Flower (YNU) datasets used for model evaluation in this study.

| category | TF | F102 | YNU |
|---|---|---|---|
| buttercup | - | 71 | 1,225 |
| carnation | - | 52 | 2,389 |
| daisy | 633 | 176 | 1,765 |
| dandelion | 898 | 92 | - |
| rose | 641 | 171 | 4,364 |
| sunflower | 699 | 61 | 654 |
| Total | 2,871 | 623 | 10,397 |
| Categories | 4 | 8 | 6 |

# Figures

## Figure 1

**Distribution of images.** Panels A and B display the UMAP embedding results for the tomato disease and flower datasets, respectively. DIM1 and DIM2 correspond to the first and second dimensions calculated by UMAP. The shapes indicate different datasets, while the colors represent the disease types.

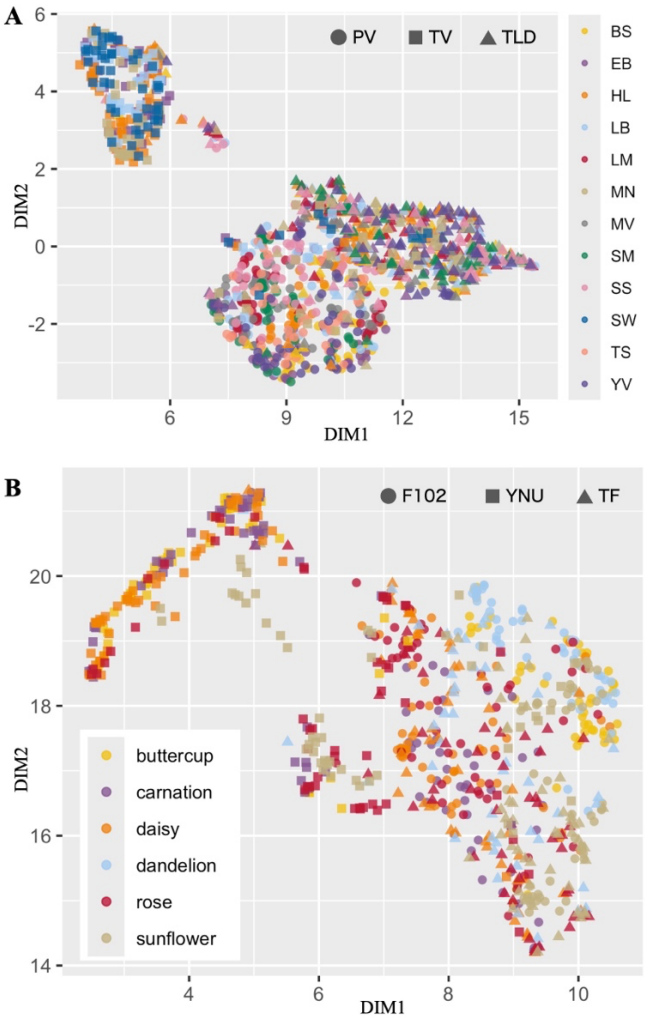


**Figure 2**

**Performance metrics of plant disease diagnosis models.** Panels A, B, and C show the precision, recall, and F1-scores, respectively, for models trained on the datasets listed in the rows and validated on the datasets listed in the columns. Panels D, E, and F are heatmaps illustrating the confusion matrices of ResNet trained on the PV dataset and validated on the PV, TV, and TLD datasets, respectively.

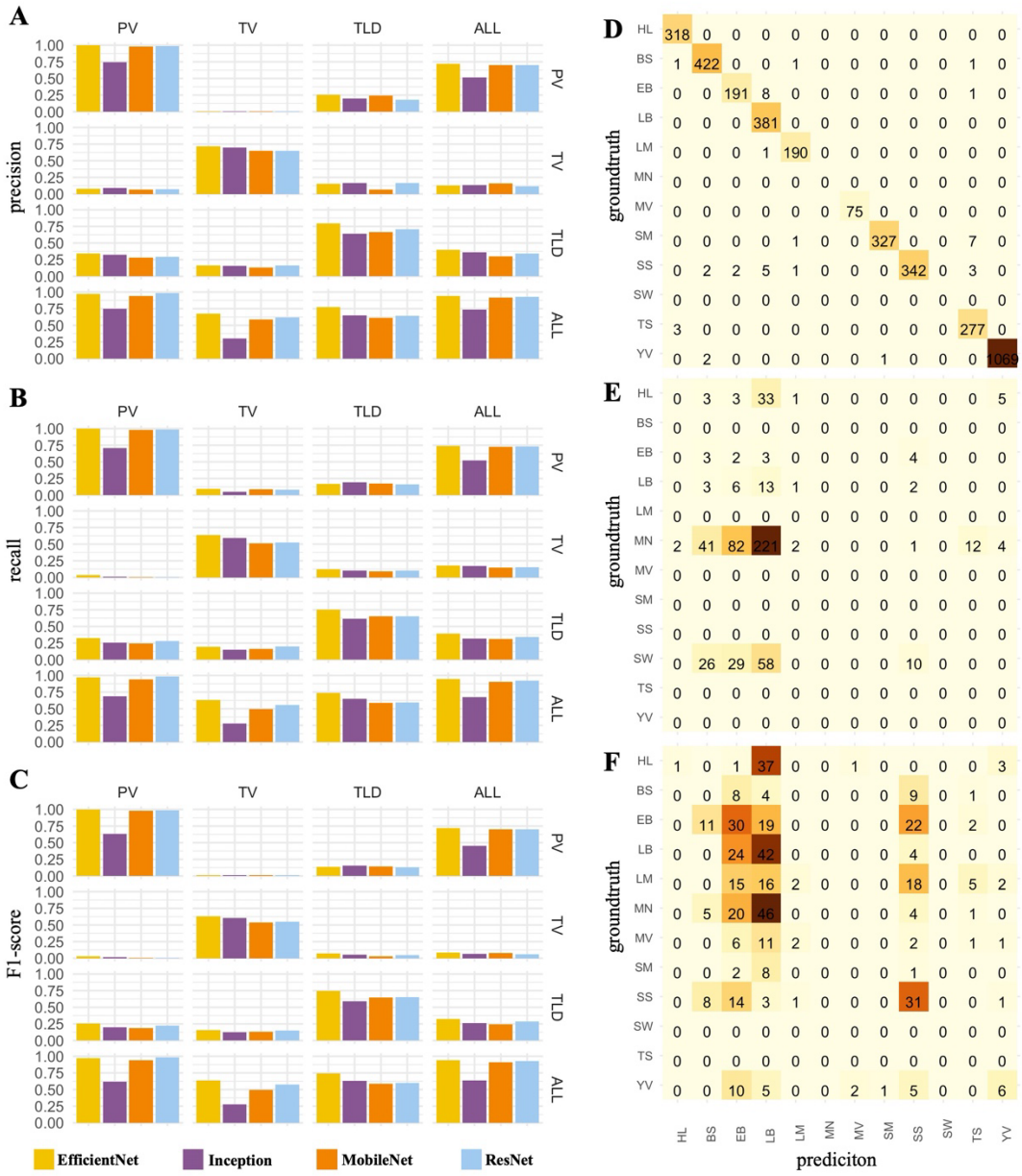
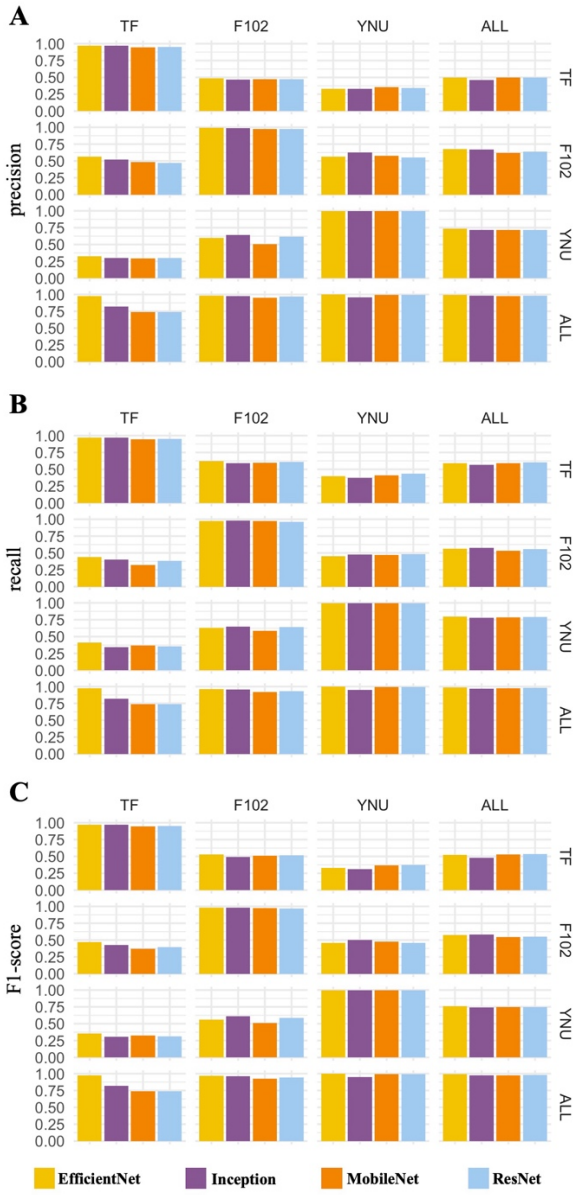


**Figure 3**

**Visualization of image regions used for classification.** Panels A and B display three original HL images from the TV and TLD datasets, respectively, along with the corresponding heatmaps generated by Grad-CAM. The heatmaps use a color gradient from blue to red to indicate the importance of different regions for classification. Red highlights areas of high importance, while blue denotes regions ignored during classification.

**Figure 4**

**Performance metrics of flower recognition models.** Panels A, B, and C show the precision, recall, and F1-scores, respectively, for models trained on the datasets listed in the rows and validated on those listed in the columns.