

Enhancing Plant Disease Detection in Smart Agriculture Using Deep Learning Models and Advanced Data Augmentation Techniques

by

Nurbol Mukhametzhan

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

at the

NAZARBAYEV UNIVERSITY

June 2026

© Nazarbayev University 2026. All rights reserved.

Author

Department of Computer Science

11 April 2026

Certified by.....

Askar Boranbayev

Professor

Thesis Supervisor

Certified by.....

Talgar Bayan

Postdoctoral Scholar

Thesis Supervisor

Accepted by

Yelyzaveta Arkhangelsky

Dean, School of Engineering and Digital Sciences

Enhancing Plant Disease Detection in Smart Agriculture

Using Deep Learning Models and Advanced Data Augmentation Techniques

by

Nurbol Mukhametzhan

Submitted to the Department of Computer Science
on 11 April 2026, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science

Abstract

Deep learning for automatic detection of plant diseases has emerged as an innovative method to address agricultural productivity and food security issues. While deep CNNs have achieved near-ceiling performance on benchmark datasets under controlled conditions, the results do not necessarily translate to real-world agricultural environments. The reason for this is domain shift: models are typically trained on clean laboratory images. Field images include background clutter, illumination variations, occlusions, and inconsistent capture conditions [1, 2, 10, 11].

This thesis investigates cross-domain plant disease classification under a controlled-to-field setting using transfer-learned CNN architectures, namely EfficientNet-B0, MobileNetV3-Large, and ResNet-50 [21, 22, 20]. PlantVillage is used as the source domain for training, while PlantDoc is used as the target domain for evaluation and adaptation. The study evaluates a 12-class controlled-to-field transfer task using a harmonized subset of overlapping classes from both datasets. The implemented framework compares four training pipelines: P1, which uses preprocessing only and no augmentation; P2, which applies basic geometric and photometric augmentation; P3, which applies stronger environmental augmentation together with mixed-image regularization using MixUp and CutMix; and P4, which serves as an ablation study by applying P3’s environmental transforms without image mixing. [24, 26, 27].

The results show that source-domain validation performance does not translate into strong field performance. In the zero-shot cross-domain setting, where models are trained on PlantVillage and evaluated directly on PlantDoc without target-domain training, the best configuration reaches only 24.84% accuracy on the filtered PlantDoc FULL evaluation set of 1405 images, confirming a substantial lab-to-field generalization gap. When target-domain supervision is introduced through fine-tuning on the filtered PlantDoc training split and evaluation on the untouched filtered PlantDoc test split, performance improves substantially, reaching up to 59.32% accuracy and 0.5734 macro-F1. The results further show that basic augmentation is more reliable

than the most aggressive augmentation setting for zero-shot transfer, while supervised fine-tuning on target-domain data provides the largest reduction in the deployment gap. Furthermore, a target-domain data efficiency ablation reveals that supervised adaptation scales logarithmically. Models recover the majority of their fine-tuned performance using only a fraction of the target dataset, demonstrating that bridging the lab-to-field gap does not strictly require massive field-collected datasets.

Overall, this research studies plant disease recognition using field data. It shows that while domain adaptation improves robustness and makes models more deployable, augmenting the source training data alone is not sufficient to solve the controlled-to-field generalization problem.

Thesis Supervisor: Askar Boranbayev
Title: Professor

Thesis Supervisor: Talgar Bayan
Title: Postdoctoral Scholar

Acknowledgments

I would like to express my sincere gratitude to Nazarbayev University for providing high-quality machine learning courses that profoundly shaped my understanding of the field. The knowledge and skills I gained throughout these courses played a crucial role in guiding the development of this thesis.

My deepest appreciation goes to my supervisor, Professor Askar Boranbayev, for his continuous support, insightful guidance, and patience throughout the entire research process. His expertise and feedback have been invaluable. I would also like to express my sincere gratitude to my co-advisor, Talgar Bayan, for his valuable insights, constructive feedback, and guidance, which significantly contributed to the development of this work. I also owe immense thanks to my family, whose unwavering encouragement, understanding, and belief in me sustained my motivation from start to finish. Their support made this work possible. Finally, I am grateful to my friends and colleagues who inspired me, offered help when needed, and made this journey both productive and memorable.

Contents

1	Introduction	13
2	Literature Review and Related Works	19
2.1	Domain Shift, Domain Generalization, and Domain Adaptation . . .	24
2.2	Model Selection Rationale	25
3	Methodology	31
3.1	Datasets and Domain Shift	32
3.2	Label Harmonization	33
3.3	Data Preprocessing	37
3.4	Model Architectures and Transfer Learning	38
3.5	Training Pipelines and Augmentation Strategy	39
3.6	Training Procedure and Hyperparameter Selection	44
3.6.1	Source-Domain Training	44
3.6.2	Target-Domain Fine-Tuning	46
3.7	Evaluation Protocol and Metrics	47
3.7.1	Implementation Environment	48
4	Results and Discussion	51
4.1	Evaluation Sets and Reporting Conventions	51
4.2	Zero-Shot Cross-Domain Results	52
4.3	Ablation Study: Impact of Augmentation Strength and Image-Mixing Regularization.	54

4.4	Architecture Comparison Under Domain Shift	56
4.5	Domain Adaptation Results	57
4.6	Error Analysis	60
4.6.1	Zero-Shot Error Patterns	61
4.6.2	Residual Errors After Fine-Tuning	61
4.6.3	Implications for Deployment	62
4.7	Discussion	64
4.7.1	Answers to Research Questions	67
4.8	Threats to Validity and Limitations	69
5	Conclusion	73
A	Figures	77

List of Figures

1-1	Overall thesis architecture. End-to-end pipeline from dataset preparation and label harmonization through four training pipelines, three CNN architectures, zero-shot evaluation, and supervised domain adaptation.	17
3-1	An example of some of the images contained within the PlantVillage dataset captured under controlled conditions.	32
3-2	An example of some of the variability present in the PlantDoc dataset images (taken in the field).	33
3-3	Representative PlantVillage samples processed through each training pipeline. Each row shows the same leaf image after P1 (no augmentation), P2 (basic geometric/photometric), P4 (environmental noise only), and P3 (environmental + MixUp/CutMix). The progressive increase in visual distortion is evident across columns. Additional samples across all four pipelines are shown in Figure A-2.	41
4-1	Training and validation curves for ResNet-50 on PlantVillage, showing strong in-domain optimization despite weak zero-shot transfer.	55
4-2	Training and validation curves for EfficientNet-B0 on PlantVillage, showing strong in-domain optimization despite weak zero-shot transfer.	55
4-3	Training and validation curves for MobileNetV3-Large on PlantVillage, showing strong in-domain optimization despite weak zero-shot transfer.	56

4-4	Target-domain data efficiency learning curves. Mean Accuracy (left) and Macro-F1 (right) evaluated on the PlantDoc test split, shown as a function of the percentage of PlantDoc training data used for fine-tuning. Models were initialized from Pipeline P2 (BasicAug) source-domain checkpoints, which were identified in the P4 ablation study as providing the most stable foundation for adaptation. Error bars indicate the standard deviation across 3 random seeds.	58
4-5	Representative misclassified PlantDoc test images for the best-performing fine-tuned model.	63
A-1	Validation loss and macro-F1 convergence curves for all three architectures trained under Pipeline P2, illustrating the two-phase training strategy and early stopping behavior.	77
A-2	Additional PlantVillage samples transformed by each training pipeline (10 random samples).	78
A-3	Early stopping epoch for each model-pipeline combination, trained on PlantVillage source domain.	79
A-4	Training time and early stopping epoch for each model-pipeline combination, trained on PlantVillage source domain.	79

List of Tables

3.1	Label mapping from PlantDoc folder names to PlantVillage canonical class labels. Only the 12 overlapping classes retained after shared-class filtering are shown. The index column corresponds to the class index used throughout all experiments. Classes without a counterpart in either dataset were excluded prior to this mapping.	35
3.2	Per-class image counts after label harmonization and shared-class filtering. The PlantVillage column shows the source-domain training data, while the PlantDoc Train and Test columns show the target-domain partitions used for fine-tuning and evaluation, respectively. Class imbalance is evident in both datasets, with PlantVillage counts ranging from 373 to 3,208 and PlantDoc test support ranging from 6 to 15 images per class.	36
3.3	Summary of the three evaluated CNN architectures. Parameter counts correspond to the full model including the replaced 12-class classification head. FLOPs are reported for a single forward pass at 224×224 input resolution.	39

3.4	Correspondence between real-world field conditions and augmentation transforms across the four training pipelines. A check mark (✓) indicates that the condition is explicitly simulated by at least one transform in the pipeline; a dash (-) indicates no coverage. Pipeline P1 applies no augmentation; P2 covers basic acquisition variability; P3 provides the most advanced augmentation including environmental degradations and mixed-image regularization; and P4 isolates the environmental degradations from P3 to serve as an ablation study. Several common field conditions remain unsimulated across all pipelines. . . .	43
4.1	Zero-shot cross-domain performance (train on PlantVillage; evaluate on PlantDoc FULL, 1405 images, 12 classes) comparing four training pipelines. P4 serves as an ablation study to isolate the effects of environmental augmentation from the mixed-image regularization used in P3. Best result per metric is shown in bold	53
4.2	Domain adaptation results on the filtered PlantDoc test split (118 images, 12 classes) using source models initialized from P3-trained (advanced augmentation) checkpoints.	57
4.3	Maximum supervised domain adaptation performance on the PlantDoc test split (118 images, 12 classes). Results represent the performance ceiling achieved by initializing models from the optimal P2 source-domain checkpoints and fine-tuning on 100% of the available target-domain training data.	57
4.4	Comparison of zero-shot (Pipeline P2, PlantDoc FULL) and best fine-tuned performance (PlantDoc test split, 118 images). Both use P2-trained source checkpoints.	59
4.5	Per-class F1 scores comparing the best zero-shot model (ResNet-50, Pipeline P2) and the best fine-tuned model (MobileNetV3-Large, Pipeline P2), both evaluated on the filtered PlantDoc test split (118 images, 12 classes). Classes are sorted by improvement magnitude.	60

Chapter 1

Introduction

Traditional methods of plant disease diagnosis rely on the expertise of farmers and agronomists, who inspect plants visually and perform laboratory tests. Visual inspections and laboratory tests provide a high degree of accuracy; however, they are extremely labor intensive and time-consuming. Furthermore, large-scale application to commercial agriculture will be difficult, if not impossible. Additionally, due to the subjective nature of both the visual inspections and laboratory tests, there exists an opportunity for significant amounts of human error. These limitations hinder early detection and timely intervention at scale, underscoring the need for intelligent, automated systems within the framework of smart agriculture.

Deep Learning (DL), specifically Convolutional Neural Networks (CNNs), is becoming increasingly popular for classifying plant diseases. DL/CNN based approaches differ fundamentally from traditional image processing methods which rely on manual definition of feature types such as color, texture and/or shape. The ability of DL/CNN to extract higher level feature representations without manual intervention provides significant advantages compared to traditional processing methods. As such, CNNs are able to attain high accuracies on standard benchmarks; i.e., 95-99% or better in controlled testing settings [1, 7]. In addition, advances in GPU computing and large annotated databases have increased the feasibility of utilizing DL/CNN based systems for real world applications of detecting plant diseases.

While DL/CNN has provided great success, there is a fundamental limitation asso-

ciated with them. A model that has been trained in a controlled environment may fail to perform well outside the same controlled environment. Typically, most commonly used datasets (e.g., PlantVillage) were designed and collected using images obtained under highly controlled conditions (i.e., consistent backgrounds and low amounts of noise) [1]. Consequently, models that are trained using datasets such as Plant Village can obtain near perfect results in environments similar to laboratories where the image collection occurred. Conversely, when the model is applied to images that include varying levels of clutter in the background, differences in lighting, occlusion, etc. the accuracy of the model will likely drop substantially. This phenomenon of decreased accuracy in the application environment due to differences between the training and test environments is called domain shift. Domain shift remains a significant barrier to widespread adoption of DL/CNN technology into agricultural practices [10, 11].

Research has previously demonstrated a large difference between how well models perform when trained solely on lab data versus when applying those models to images collected in a field environment. Singh et al. (2020) established the PlantDoc dataset consisting of images of plants that were taken in a field environment rather than a laboratory environment. They illustrated through experiments that even though models can be trained using only lab data, there is considerable distance between the accuracy of the model when tested using only images taken in a laboratory and the accuracy of the model when tested on images taken in a natural field environment [3]. Additionally, they showed that the inclusion of data taken from the target domain via fine-tuning improved model performance by as much as 31% when evaluating the classification accuracy. Therefore, including data from the target domain is critical to increasing model robustness against domain shift.

This work assesses a 12-class transfer scenario by establishing a commonality in labeling domains between PlantVillage and PlantDoc. By means of an explicit mapping and filtering procedure, this work considers a shared set of common pathologies among three types of crops: pepper, potato and tomato. Establishing this commonality is necessary prior to performing a valid cross-domain evaluation since it ensures that both the source-trained classifier and the target evaluation set operate in an

identical index space.

The thesis then compares four training pipelines of progressively increasing regularization strength. Pipeline P1 serves as a no-augmentation baseline. Pipeline P2 applies moderate augmentation to introduce source-domain diversity without distorting disease morphology. Pipeline P3 represents the most aggressive configuration, combining environmental degradations with mixed-image regularization. Pipeline P4 is introduced as an ablation of P3, isolating the contribution of environmental noise from that of structural image mixing. These pipelines are evaluated under two separate protocols. First, zero-shot cross-domain generalization is measured by training on PlantVillage and evaluating directly on the filtered PlantDoc FULL set, which merges the PlantDoc train and test partitions after shared-class filtering, resulting in 1405 evaluation images. Second, supervised domain adaptation is measured by fine-tuning the source-trained models on the filtered PlantDoc training split of 1287 images and evaluating on the untouched filtered PlantDoc test split of 118 images. Additionally, a data-fraction ablation study is conducted during fine-tuning to quantify the data efficiency of target-domain adaptation, scaling the available target data from 10% to 100%. This separation is critical because it distinguishes pure out-of-domain transfer from target-domain adaptation and prevents confusion between zero-shot evaluation and fine-tuning-based improvement. In this way, the contribution of the thesis is not simply to show that augmentation does help, but to quantify, under a reproducible experimental setup, how much augmentation alone can improve cross-domain robustness and how much additional benefit is obtained once target-domain supervision is introduced.

To systematically address these objectives, this thesis is guided by the following primary research questions:

RQ1: How severe is the generalization gap when convolutional neural networks (specifically ResNet-50, EfficientNet-B0, and MobileNetV3-Large) trained on controlled laboratory imagery are deployed directly on real-world field imagery?

RQ2: To what extent can varying intensities and types of data augmentation (basic vs.

aggressive environmental, and mixed-image regularization) mitigate the domain shift between controlled and field environments?

RQ3: How does supervised domain adaptation, achieved through target-domain fine-tuning, compare to data augmentation in recovering classification performance and bridging the lab-to-field gap?

RQ4: How data-efficient is supervised domain adaptation? Specifically, what fraction of labeled target-domain field data is required to effectively bridge the lab-to-field generalization gap?

The remainder of the thesis will be structured as follows to address each question. Chapter 2 provides a review of prior work on deep learning based systems for detecting plant diseases; domain shift and the selection criteria for the chosen models. Chapter 3 outlines the methodology used for this project including how the datasets were made consistent across all projects, the configuration of the four training pipelines and how the performance of the different approaches would be evaluated. Chapter 4 describes the results obtained from the experiments and compares supervised domain adaption with zero-shot generalization, supplemented by an error analysis and an ablation study that assesses the impact of structural augmentation. Finally, Chapter 5 concludes the study, identifies potential shortcomings of the current work and offers possible paths forward for further investigation.

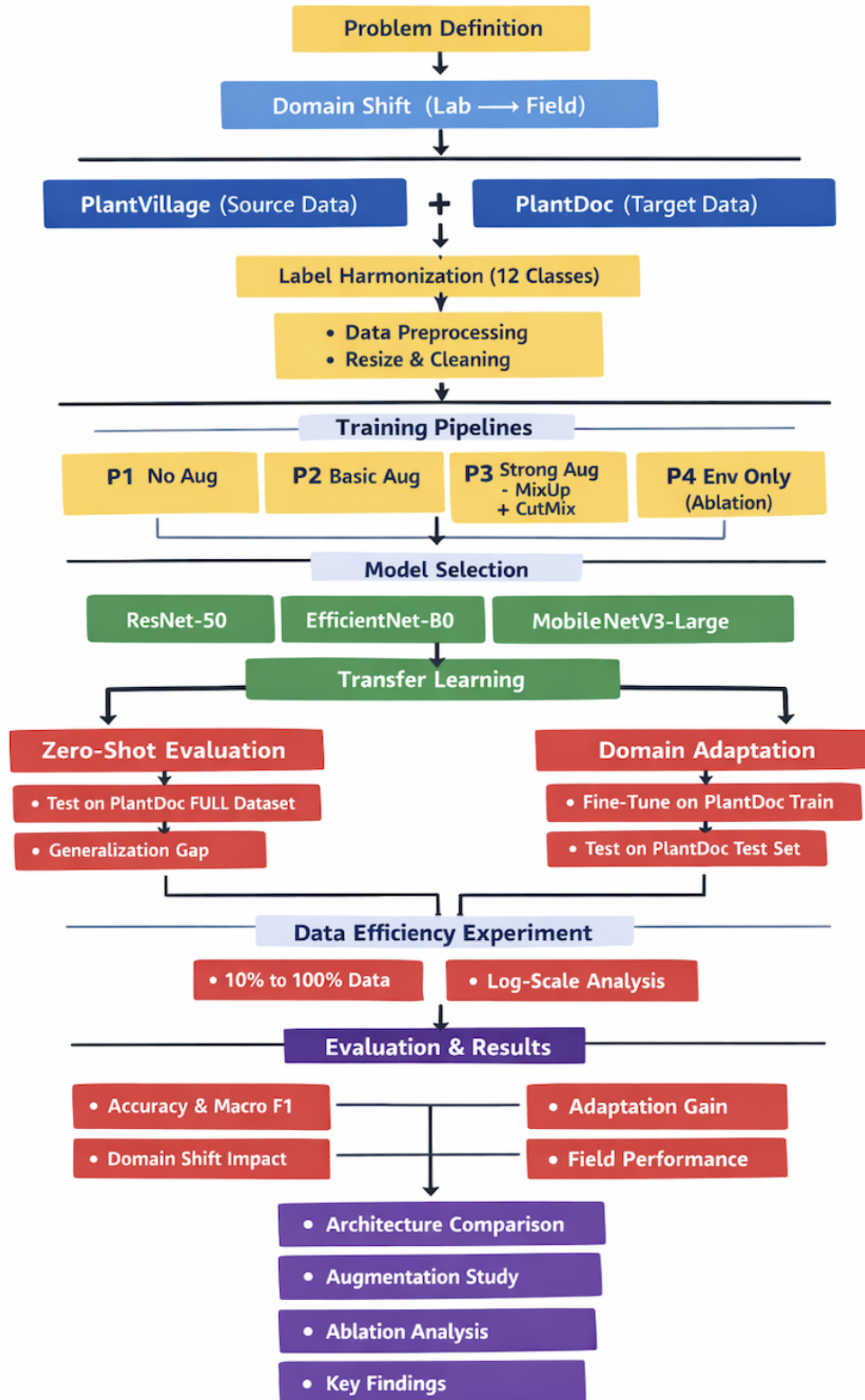


Figure 1-1: Overall thesis architecture. End-to-end pipeline from dataset preparation and label harmonization through four training pipelines, three CNN architectures, zero-shot evaluation, and supervised domain adaptation.

Chapter 2

Literature Review and Related Works

Early success in the application of deep learning technology toward the detection of plant disease led to rapid advancement within the field. The ability to detect disease in plants at an accuracy rate of over 99% across 38 crop-disease classifications was demonstrated by Mohanty et al. (2016), utilizing data from the Plant Village dataset [13]; similarly, Ferentinos (2018) reported accuracy rates of up to 99.53% across 25 different plant diseases using multiple CNN models [14]. These successful applications were followed by numerous other studies published between 2020 and 2025 which focused on improving architectures and training methodologies for detecting plant diseases. Recent studies have also investigated the use of various CNN's including VGG16, ResNet, Inception, DenseNet, and EfficientNet for identifying plant diseases. A systematic review of studies investigating plant disease concluded that while VGG16 has traditionally been the most common CNN model used in leaf disease classification; it is no longer being utilized as frequently due to its lower accuracy and performance when compared to newer network models such as ResNet-50 and EfficientNet [1, 8, 3]. An example of the potential of using advanced CNN architectures is the paper by Atila et al. (2021), which introduced an EfficientNet based model to classify plant disease using PlantVillage data. They reported 99.91% - 99.97% accuracy, after applying both transfer learning and data augmentation, outperforming previously established state of the art results [5]. The above indicates that when given sufficient training data, advanced CNN architectures can

develop highly discriminative features for plant disease. Another important area in the literature is the practice of using transfer learning with pre-trained models. As researchers have transitioned away from training from scratch (which requires massive amounts of data), they have turned toward fine-tuning models that have already been trained on large scale image databases (for example, ImageNet). With the increased effectiveness of transfer learning (fine-tuning models trained on large image corpora such as ImageNet) being applied to smaller agricultural datasets [1, 9], it has quickly become the standard method for plant disease classification. The authors of Chen et al. (2021) illustrated the potential of this method by adding attention mechanisms to a pre-trained MobileNetV2 model, employing a two-stage fine-tuning process, and achieving 99.67 percent accuracy on a publicly available leaf disease database [18]. While still providing mobile deployment capabilities, this method was successful. Fine-tuning can be combined with full-network fine-tuning or hybrid methods; for instance, Mamun et al. (2024) employed an SVM classifier with deep features from ResNet-50 and achieved an F1 score of 0.9838 on a rice disease dataset that outperformed an end-to-end CNN approach [30]. Results collectively indicate that pre-trained representations represent a solid foundation for detecting plant diseases across architectures and dataset sizes [1, 2]. Results presented within the literature support the advantages of applying pre-trained deep representations to detect plant diseases. Due to the small sample sizes and highly controlled nature of many publically available plant disease databases, there exists a high risk of overfitting and poor generalization, making data augmentation nearly universally applicable as part of all current training pipelines [2]. A variety of basic augmentation methods including random flip, crop, rotate, brightness and/or contrast jittering have been demonstrated to significantly reduce overfitting and increase CNN accuracy on leaf disease-related tasks through an increase in the effective diversity of the training set without requiring the collection of additional data [24].

In addition to geometric and photometric transformations, generative-based approaches have also been explored. Abbas et al. (2021) utilized a CycleGAN to generate realistic images of diseased and healthy tomato leaves and showed improvement

in detection performance when the generated images were included in the training set [4]. However, there continues to be debate regarding the practicality of generating synthetic disease images that appear indistinguishably from real ones as this task is currently technologically difficult and computationally expensive [2].

As a distinct body of research explores mixed-image augmentation, which generates novel training examples by integrating parts of two input images together with their associated labels. MixUp [26] interpolates the pixel value and label information between two images, whereas CutMix [27] replaces a rectangular section of one image with a portion of another. Both techniques encourage the model to learn from incomplete or blended inputs and have proven to diminish overly confident predictions. Li and Tanone (2023) documented that the inclusion of MixUp and CutMix augmented classification robustness for corn leaf diseases [16], while Jiang et al. (2025) integrated both methods into a few shot learning framework for cross-domain plant disease recognition [17]. Random erasure has also been utilized to simulate the partial occlusion that can occur when leaves or fruit overlap in field images [25].

Overall, the literature establishes that augmentation is essential for improving model robustness, but its effectiveness depends on the realism, magnitude, and task-relevance of the transforms applied. Models trained with greater and more varied augmentation are generally less prone to overfitting; however, whether augmentation alone can bridge the gap between controlled and field imagery remains an open question that the present thesis directly investigates.

Field deployment and cross-domain generalization have become increasingly important in plant disease recognition research because performance on laboratory benchmarks does not necessarily predict performance under real agricultural conditions. Controlled datasets such as PlantVillage contain relatively clean samples with centered leaves, limited occlusion, and simple backgrounds, which makes them well suited for benchmarking representation learning but less representative of deployment conditions. In contrast, field images often contain cluttered backgrounds, multiple leaves, variable lighting, scale changes, motion blur, and partial occlusion [15]. As a direct consequence of this, the actual distribution of the underlying data

at deployment will likely be different from what was seen during the model’s training phase; this represents a well known type of ‘dataset shift’, often referred to as “domain shift”. Reviews have recently highlighted how this problem has remained one of the key open issues for agricultural computer vision. Moreover, many of the plant disease models which have been developed and published in recent years, remain largely limited to being tested in an artificially curated environment (e.g. using labeled datasets), and not in real-world field environments [1, 2, 29].

The issue addressed in this part of the thesis is particularly relevant to the experimental studies described herein. The classification performance on a single dataset is not solely evaluated by the studies. Instead, the studies are designed to specifically assess whether or not there exists a controlled-to-field transfer, i.e., whether a model trained on PlantVillage images performs well when it is used to classify images in the PlantDoc dataset. The aligned report shows that this setting produces a severe performance drop in the zero-shot case, even when source-domain validation performance is near ceiling. This behavior is fully consistent with the broader literature: strong in-domain accuracy does not imply robustness to changes in background, acquisition conditions, or dataset construction. Therefore, the most relevant prior work for this thesis is not only work that reports very high classification accuracy, but also work that studies why such performance degrades under deployment mismatch and what interventions can partially reduce that degradation [2].

Among available datasets, PlantDoc is especially important in this context because it was introduced specifically to expose the limitations of laboratory-trained plant disease classifiers under more realistic visual conditions. Prior work using PlantDoc has shown that models trained only on laboratory imagery perform poorly on field images, while incorporating target-domain data through fine-tuning can substantially improve performance. This distinction maps directly onto two evaluation settings used in the present thesis: zero-shot cross-domain generalization, where no target-domain training is allowed, and supervised domain adaptation, where source-trained models are further fine-tuned on PlantDoc training data before being evaluated on a held-out PlantDoc test set. Framing the literature in this way is important be-

cause it clarifies that augmentation and transfer learning are not being studied here in isolation; they are being studied as interventions under a specific controlled-to-field distribution shift [3, 28, 29].

In summary, the literature establishes three points that are directly relevant to this thesis. First, deep CNNs and transfer learning can achieve very high accuracy on curated plant disease datasets, particularly when training and test images are drawn from similar controlled conditions. Secondly, data augmentation is utilized throughout the source domain to diminish overfitting and increase robustness, yet its efficacy is heavily reliant upon the level of realism, magnitude, and task-relatedness of each transformation applied. Thirdly and most importantly, however, it has been found that model performance can decrease drastically when laboratory images are replaced with field images, demonstrating that the primary challenge is not merely the ability to classify images but the ability to maintain robustness across domains. The aforementioned observations have driven the experimental design of this dissertation. As opposed to assessing architecture(s) solely based upon their accuracy in their respective domain(s), this study will assess how three pre-trained CNN backbones operate under a controlled-to-field (laboratory images to field images) transfer condition subsequent to explicit labeling harmonization between PlantVillage and PlantDoc. The study will also compare three progressively stronger training pipelines, providing a means to examine the impact of augmentation strength as an ablation, versus assuming that augmentation is always beneficial. Additionally, by distinguishing between zero-shot evaluation and supervised fine-tuning using labeled target-domain data, the dissertation will better adhere to the domain-adaptation literature, thereby providing a more realistic representation of the requirements needed for practical application of image-based systems in agricultural settings [1, 2, 28, 29].

2.1 Domain Shift, Domain Generalization, and Domain Adaptation

The simplest form of the cross-domain learning problem is that where the model is first trained on one domain and then tested on another for which no test labels exist. This setting can also be referred to as zero-shot cross-domain testing (or domain generalization) [29] or simply as domain generalization. When some amount of labeled data exists in the target domain and this data is used when training the model, the problem is now a matter of supervised domain adaptation [28]. There is an important distinction here in terms of the type of question being asked. Zero-shot testing asks how much feature extraction from the source domain will be applicable in the target domain. Fine-tuning with labeled data from the target domain addresses how much of the domain difference will be removed if you expose the model to the target distribution.

The main reason for distinguishing between the two cases is to interpret the results of both evaluations correctly. The evaluation of a model with poor performance in a zero shot case does not mean the model is a bad classifier but rather the model may be very well-suited to the source domain. On the other hand, an improved performance after fine tuning does not mean the model had a good (robust) cross-domain representation; it means the model needed additional supervised learning using data from the target domain to learn how the conditions in the target domain are different from those of the source domain [28]. The aligned studies experiment in this thesis follow exactly this line of reasoning: the performance of a model in a zero shot case will be measured on a filtered version of the full PlantDoc dataset while the performance of a model when fine tuned will be measured by fine tuning on a filtered version of the PlantDoc training split and measuring its performance on a filtered version of the PlantDoc test split. Thus, this design puts the empirical work of this dissertation into the literature of domain shift as opposed to the literature of pure in-domain benchmarking [7].

2.2 Model Selection Rationale

Recent studies have tested and compared three of the most common architectures in deep learning for plant disease detection: ResNet-50, MobileNetV3, and EfficientNet-B0, based on the three primary criteria: the level of accuracy achieved, the speed at which images are processed, and the feasibility of implementation. ResNet-50: ResNet-50 is a 50 layer deep CNN, a residual neural network. It has a high ranking in many image-based computer vision tasks and has been heavily used in classifying plant diseases, as it can capture powerful features. Studies have reported that ResNet-50, after being fine-tuned for specific plant disease classification tasks, achieves very high accuracy scores, typically serving as a reference point. For example, Mamun et al., in their study of rice leaf disease classification, used the deep features obtained by ResNet-50 along with an SVM classifier and reached a high F1 score of 0.9838 [30]. Similarly, studies based on ResNet-50 have reported high accuracy rates ($> 95\%$) on datasets that include multiple types of leaf diseases (e.g., PlantVillage) [1, 14]. One of the major advantages of ResNet-50 is the fact that it is capable of extracting very discriminant features; however, there are some disadvantages associated with ResNet-50 including the very large number of parameters (more than 23 million) and the very high computational cost (more than 4 billion floating-point operations per second/inference) that can make it difficult to implement in real time and/or in constrained environments (e.g., mobile devices, embedded systems). Therefore, while ResNet-50 is a good reference point for evaluating the accuracy of deep learning models for plant disease classification, researchers have generally looked to other models that can provide better trade-offs with respect to processing speed and/or resource usage. However, in certain scenarios where the maximum possible accuracy is required and computational resources are sufficient, ResNet (and even deeper versions of ResNet, e.g., ResNet-101/152) can be used as part of an ensemble approach to achieve higher accuracy performance. Examples of such scenarios include Ahad et al. (2023), who developed an ensemble method consisting of DenseNet, MobileNet, and ResNet152 for rice disease detection to en-

hance robustness [33]. MobileNetV3: MobileNetV3 is a family of CNN architectures that have been specifically designed for application in mobile and embedded vision tasks. Both the Small and Large variants of MobileNetV3 utilize a combination of techniques, including depthwise separable convolution and squeeze-and-excitation module, to greatly decrease the number of parameters, while maintaining comparable accuracy to much deeper models. As such, MobileNet models have demonstrated excellent performance in plant disease classification, given their efficiency. For example, Hassan et al. (2021) demonstrated real-time inference using MobileNetV2 and reported accuracy levels up to 99.56% on 38 different types of plant diseases (PlantVillage dataset), essentially matching the performance of significantly heavier networks [10]. MobileNetV3 improves upon the architecture of V2 and has been incorporated into more recent studies. Karim et al. (2024) developed a modified version of the Large MobileNetV3 for detecting grape leaf diseases and deployed the model onto a Jetson Nano. The model achieved an accuracy level of 99.42% in testing and was able to run in real time on the edge device due to its compact size (less than 5 million parameters, and approximately <1 million parameters after quantization) [11]. The major advantage of MobileNetV3 is that it is computationally efficient and can process images at a much faster rate and consume less memory/power than alternative architectures. The main disadvantage of MobileNetV3 is that they tend to slightly underperform when compared to very deep models. However, studies have indicated that the loss in accuracy is minimal relative to the efficiency gained through the reduced computational requirements. For example, Chen et al. (2021) were able to develop an attention mechanism-enhanced MobileNetV2, which achieved an accuracy level of 99.67%, demonstrating that lightweight models can achieve equivalent performance to larger networks on benchmark datasets [18]. Overall, MobileNetV3 (and other lightweight CNNs) are optimal for speed and deployability. As such, they are well-suited for on-site disease monitoring applications (e.g., via drones, smartphone apps, etc.). The primary drawback of MobileNetV3 is potentially limited to highly complex classification tasks, or high resolution images, where larger models would have captured additional detail. However, for the majority of leaf disease

classification tasks examined to date, MobileNet-based models provided near state-of-the-art accuracy with significantly less computational overhead. MobileNetV3 is based upon the compound scaling approach utilized in EfficientNet-B0 to find the optimal balance between network depth, input resolution, and number of layers for optimal accuracy/computation trade-off. MobileNetV3 has approximately 5.3 million parameters, comparable to MobileNet; however, this model includes a more sophisticated architecture that was co-designed through neural architecture search. After 2019, EfficientNet-B0 and its larger variants (B3, B5, etc.) became very popular in plant disease detection. EfficientNet-B0 and larger variants have demonstrated good performance in multi-class problems. For instance, Atila et al. (2021) utilized the EfficientNet-B5 and B4 models (which are both larger than B0) on the PlantVillage database and obtained over 99.9% accuracy with data augmentation [5]. While the smaller EfficientNet-B0 is still a very capable model, it has shown strong results. Faisal et al. (2023) compared EfficientNet-B3 (a scaled-up version of B0) against ResNet-50 and MobileNet, and concluded that EfficientNet models performed better for the classification of citrus leaf diseases [1]. In addition to the aforementioned example, an attention-augmented EfficientNet-B0 achieved 99.39% accuracy on a 39-class dataset, demonstrating a substantial improvement relative to a plain CNN (which had an accuracy of 46.7%) on the same task [8]. Similarly, Adnan et al. (2023) applied adaptive augmentation to EfficientNetB3 for multi-class plant disease classification and demonstrated strong accuracy gains on PlantVillage data [12]. This demonstrates the ability of EfficientNet to achieve high accuracy with relatively low computational complexity. Although EfficientNet-B0 will generally require more computations than MobileNetV3 of similar size due to the compounded depth of EfficientNet-B0 and the higher input resolution, EfficientNet-B0 typically achieves greater accuracy per parameter than MobileNetV3. Additionally, EfficientNet-B0 has been reported to possess robustness; it performs well when fine-tuned to new data. Like MobileNet, however, EfficientNet is optimized for efficiency and thus would be desirable for edge deployment. Specifically, Hassan et al. (2021) pointed out that EfficientNet-B0, combined with depthwise convolutions and proper training, achieved 99.5% accuracy on

the PlantVillage database while being deployable on mobile devices [8]. Therefore, ultimately, the choice between utilizing either MobileNet V3 or EfficientNet B0 will be based upon the specific constraints (resource) which are present, as well as whether or not it is necessary to obtain slightly higher levels of precision from the model than what is provided by MobileNet V3. MobileNetV3 can be selected if extreme limitations on memory and/or inference time exist. On the other hand, ResNet-50 may be reserved for server-side analysis or as part of an ensemble to validate results for plant disease identification rather than stand-alone use on the farm. Comparison Summary: In total, the employment of pre-trained CNNs has significantly aided the rapid development of plant disease detection systems. The fine-tuning of a robust architecture (ResNet, EfficientNet, MobileNet, etc.) on plant images has now become a standard workflow for detecting plant diseases, frequently resulting in classification accuracies of 95–99% [8]. ResNet-50 represents a more accurate option for plant disease detection; however, ResNet-50 is computationally more expensive than MobileNetV3 and EfficientNet-B0. Conversely, MobileNetV3 and EfficientNet-B0 offer a favorable trade-off between accuracy and computational efficiency [8, 21, 22]. Consequently, the choice of model can be determined by the intended deployment scenario: if a cloud-based system using drone images is desired, a more complex model (e.g., ResNet-50, EfficientNet-B3/B5) can be employed to optimize accuracy. If a farmer’s smartphone application or an edge device located in the field is desired, a smaller model (e.g., MobileNetV3, EfficientNet-B0) should be selected, as illustrated by several studies that successfully achieved real-time detection with MobileNet/EfficientNet without suffering a meaningful loss in accuracy [8]. Model interpretability and user trust are important considerations in agriculture; as mentioned above, several studies have incorporated attention modules or Grad-CAM visualizations into these models (primarily EfficientNet- and MobileNet-based models) to visualize diseased areas and provide explanations to users regarding predictions [8]. The experiment utilize all of these insights in this research: it evaluates EfficientNet-B0 and MobileNetV3 (both pre-trained on ImageNet) as base models because they are highly accurate and deployable, and also include ResNet-50 in comparison as a baseline. The intention is

to determine what trade-offs exist within the above-mentioned models, and then the one with the highest level of both accuracy and speed for use in smart agriculture will be chosen.

Field images have several environmental conditions that contribute to visual degradation (e.g., illumination variation). Illumination variance results from changes in sunlight due to sun angle, clouds, and canopy shade causing shadows, over exposure, and non-uniformity of light throughout the leaf's surface. Atmospheric variances (fog, haze, etc.) result in reduced contrast and color saturation particularly at times of day when farmers typically examine their crops (i.e. early am/late pm) and may also be subject to rain. Leaf motion resulting from handheld camera operation or by wind induced leaf movement occurs commonly outside. Camera sensor noise is increased through use of less than ideal low-light conditions or use of lower quality camera systems common to many smartphones. Beyond these image degradations, physical contaminants (soil, dust, water drops, insects, etc.) on the leaf surface may hide symptomology of diseases or produce classifications similar to lesions thereby adding additional classification problems that do not occur in laboratory captured datasets. Leaves covered with dew and/or rain will alter how the reflective characteristics of the leaf surface appear and thus the color and texture of both healthy and diseased tissues. The previously listed environmental factors rarely occur independently of each other; therefore it is rare that a single field image does not contain multiple forms of degradation (i.e. shadow, blur, and dust), further complicating the ability to accurately classify.

Chapter 3

Methodology

Overview

The objective of this study is to determine how well an automatic plant disease classifier performs when there has been a shift in domains; that is, the image data used for training were controlled laboratory images and the test images are real world images taken in a field environment. Six major steps comprise the method used in this study. These include (1) selecting appropriate datasets with which to compare the performance across different environments as well as characterize the extent of the domain shift, (2) perform label harmonization so that both sets of images have common classes or categories for evaluation purposes, (3) preprocess each set of images by removing noise and adjusting size and color to be consistent and compatible with the classifier, (4) define four separate but increasingly strong pipelines that can augment the original images using random transformations such as rotation and flipping, (5) choose model(s), use a form of transfer learning to allow models trained on one type of dataset to perform reasonably well on other types of datasets, and train and evaluate these models under two scenarios: zero shot cross domain generalization and supervised domain adaptation using fine tuning.

3.1 Datasets and Domain Shift

Two publicly available datasets were used: PlantVillage and PlantDoc.

The PlantVillage database is comprised of about 54,306 images, which are organized into 38 crop-disease categories and include 14 crop types. Each type has its own "healthy" category. Because all of the images were taken using the same lab setup (with consistent background and reduced image noise) the models that have been created to be used with PlantVillage, will generally obtain good results in lab settings as well [1]. Representative samples of the controlled setting images of PlantVillage can be found in Figure 3-1.



Figure 3-1: An example of some of the images contained within the PlantVillage dataset captured under controlled conditions.

PlantDoc is a new dataset that represents our target domain. It contains over 2,598 images of 13 plant species and their 17 diseases. Images from PlantDoc were taken in actual field environments and thus have great variation in lighting, background distractions, and degree of occlusion by leaves; cameras also had varying degrees of quality. Researchers have shown previously that models that have been trained only on lab-created data tend to fail badly on test datasets like PlantDoc. This demonstrates the problem of the domain shift [3]. Samples of what the field

images look like are illustrated in Figures 3-2.



Figure 3-2: An example of some of the variability present in the PlantDoc dataset images (taken in the field).

The two datasets differ not only in their visual characteristics but also in their labeling conventions, which prevents direct cross-dataset comparison without preprocessing. PlantVillage uses a structured naming format of the form `Crop__Disease`, while PlantDoc uses more informal folder names that may omit the crop prefix or use different disease terminology. Resolving this discrepancy is the subject of the following section.

3.2 Label Harmonization

The PlantVillage and PlantDoc datasets differ in their folder and class naming conventions, making direct comparison between the two datasets impossible unless the label space is preprocessed. PlantVillage uses a structured naming format of the form `Crop__Disease`, while PlantDoc uses more informal and variable folder names that may omit the crop prefix or use different disease terminology. Therefore, to ensure validity of cross-dataset evaluation, this study identifies an explicit mapping from PlantDoc class names to PlantVillage canonical class labels, as shown in Table 3.1.

After mapping, non-overlapping classes are discarded. Only those classes that appear in both datasets post-mapping are retained. Several PlantDoc classes could not be matched to any PlantVillage class due to the absence of an equivalent category in the source dataset. For example, `Apple_Scab_Leaf`, `Corn_Gray_leaf_spot`, and `grape_leaf_black_rot` were present in PlantDoc but had no corresponding PlantVillage class because the filtered PlantVillage directory only contained pepper, potato, and tomato crop categories. In total, 16 PlantDoc classes were excluded on this basis, including all apple, blueberry, cherry, corn, grape, peach, raspberry, soybean, squash, and strawberry categories. Similarly, certain PlantVillage classes that have no counterpart in PlantDoc were excluded from training, such as `Potato___healthy` and `Tomato__Target_Spot`, which had no matching folder in the PlantDoc dataset. In order to accomplish the above mentioned process (the "filtering" part) of this research, twelve classes are identified, and thus a common label space, which is illustrated in Table 3.2, and Table 3.1, are identified. The common label space is utilized for all the training, validation, zero shot testing and fine tuning activities conducted in this study. The task of matching labels with one another is not simply an option nor a cosmetic procedure; it has to be accomplished. In fact, matching labels is a mandatory methodology. For example, if cross-domain evaluations are going to have any meaning at all, then there must exist a common label space that is utilized by the classifier trained on the source domain, and the test data from the target domain. Without a common label space, a model trained on all 38 classes of PlantVillage and tested on the 17 classes of PlantDoc would produce output probabilities that would not correspond to class index values present in either dataset. Therefore, any reported results based upon accuracy would be meaningless.

Index	PlantDoc Folder Name	PlantVillage Canonical Label
0	Bell_pepper_leaf_spot	Pepper__bell___Bacterial_spot
1	Bell_pepper_leaf	Pepper__bell___healthy
2	Potato_leaf_early_blight	Potato___Early_blight
3	Potato_leaf_late_blight	Potato___Late_blight
4	Tomato_leaf_bacterial_spot	Tomato_Bacterial_spot
5	Tomato_Early_blight_leaf	Tomato_Early_blight
6	Tomato_leaf_late_blight	Tomato_Late_blight
7	Tomato_mold_leaf	Tomato_Leaf_Mold
8	Tomato_Septoria_leaf_spot	Tomato_Septoria_leaf_spot
9	Tomato_leaf_yellow_virus	Tomato__Tomato_YellowLeaf__Curl_Virus
10	Tomato_leaf_mosaic_virus	Tomato__Tomato_mosaic_virus
11	Tomato_leaf	Tomato_healthy

Table 3.1: Label mapping from PlantDoc folder names to PlantVillage canonical class labels. Only the 12 overlapping classes retained after shared-class filtering are shown. The index column corresponds to the class index used throughout all experiments. Classes without a counterpart in either dataset were excluded prior to this mapping.

Table 3.2 summarizes the per-class image counts for each dataset partition after the harmonization and filtering described above. The resulting shared label space contains 12 classes covering pepper, potato, and tomato diseases.

Class	PlantVillage	PlantDoc Train	PlantDoc Test
Pepper__bell___Bacterial_spot	997	74	9
Pepper__bell___healthy	1478	34	8
Potato___Early_blight	1000	157	14
Potato___Late_blight	1000	200	8
Tomato_Bacterial_spot	2127	101	9
Tomato_Early_blight	1000	79	9
Tomato_Late_blight	1909	101	10
Tomato_Leaf_Mold	952	85	6
Tomato_Septoria_leaf_spot	1771	145	12
Tomato_Tomato_YellowLeaf_Curl_Virus	3208	223	15
Tomato__Tomato_mosaic_virus	373	44	10
Tomato_healthy	1591	44	8
Total	17406	1287	118

Table 3.2: Per-class image counts after label harmonization and shared-class filtering. The PlantVillage column shows the source-domain training data, while the PlantDoc Train and Test columns show the target-domain partitions used for fine-tuning and evaluation, respectively. Class imbalance is evident in both datasets, with PlantVillage counts ranging from 373 to 3,208 and PlantDoc test support ranging from 6 to 15 images per class.

Class imbalance is present in both datasets. In PlantVillage, image counts per class range from 373 (Tomato mosaic virus) to 3,208 (Tomato Yellow Leaf Curl Virus). In PlantDoc, the imbalance is more pronounced due to the smaller dataset size and the natural variability of field-collected imagery. This imbalance motivated the use of class-weighted cross-entropy loss during training, as described in Section 3.6.

For evaluation purposes, two distinct target-domain sets are derived from the filtered PlantDoc data. The *PlantDoc FULL* set merges the filtered PlantDoc training and test partitions to produce 1,405 evaluation images; this set is used exclusively for zero-shot cross-domain evaluation where the model has never seen any target-

domain data. For supervised domain adaptation, the original PlantDoc partitions are preserved: the filtered training split (1,287 images) is used for fine-tuning, and the filtered test split (118 images) is used for evaluation. By ensuring that these two testing regimes were mutually exclusive, we ensured that they would provide answers to two different Research Questions through Non-Overlapping Experimental Designs. Once the Class Space is determined to be Shared by Both Datasets, All Images in both datasets will go through the Common Preprocessing Pipeline before entering the training phase.

3.3 Data Preprocessing

All images were loaded and converted to three-channel RGB format to ensure a consistent input representation, since some images in the raw datasets were stored as grayscale or RGBA. Each image was then resized to 224×224 pixels, which is the default input resolution for all three evaluated architectures when initialized from their standard ImageNet-pretrained checkpoints. This resolution was selected because it matches the spatial dimensions used during the original ImageNet pretraining of ResNet-50, EfficientNet-B0, and MobileNetV3-Large, and deviating from it would require the pretrained positional statistics in batch normalization layers to adapt to a different feature map geometry, potentially slowing convergence or degrading transfer performance. Following resizing, all pixel values were normalized using the per-channel ImageNet mean $\mu = (0.485, 0.456, 0.406)$ and standard deviation $\sigma = (0.229, 0.224, 0.225)$. These specific normalization statistics were chosen because they correspond to the channel-wise distribution of the ImageNet training set on which all three backbone models were originally pretrained. Using the same normalization ensures that the input distribution seen by the pretrained feature extractors at the start of training is consistent with the distribution they were optimized for, which is a standard requirement when applying transfer learning from ImageNet-initialized models. No additional preprocessing steps, such as histogram equalization or background removal, were applied, in order to preserve the natural appearance

of both laboratory and field images and to avoid introducing preprocessing-induced artifacts that could confound the cross-domain evaluation. The preprocessed images serve as input to three pretrained CNN architectures, each representing a different trade-off between classification capacity and computational efficiency.

3.4 Model Architectures and Transfer Learning

Three pretrained convolutional neural network architectures were evaluated: ResNet-50 [20], EfficientNet-B0 [21], and MobileNetV3-Large [22]. The designs illustrated provide distinct balances among these competing performance characteristics. A quantitative summary of each of the architectures is provided in Tab. 3.3, which lists (1) the number of parameters used by each model; (2) the relative computational costs associated with each architecture for inference at fixed-size inputs; and (3) the maximum size of each input image that can be processed without additional processing resources. All three of these design metrics have implications for the practical deployment of deep learning-based systems – an implication supported by their inclusion in the comparative evaluation described in this paper. ResNet-50 is designed to serve as a large-capacity baseline, representing the largest feasible number of trainable parameters and computational requirements for deploying such systems on servers. EfficientNet-B0 and MobileNetV3-Large both target efficiency-constrained environments, but differ in their architectural design: EfficientNet-B0 uses compound scaling to balance depth, width, and resolution, while MobileNetV3-Large uses depthwise separable convolutions and hardware-aware neural architecture search to minimize latency on mobile processors. Including all three architectures allows the study to assess whether the domain shift problem is influenced by model capacity or architectural design, or whether it is primarily a property of the dataset gap itself.

Architecture	Parameters (M)	FLOPs (G)	Input Size
ResNet-50	23.53	4.09	224×224
EfficientNet-B0	4.02	0.38	224×224
MobileNetV3-Large	4.22	0.22	224×224

Table 3.3: Summary of the three evaluated CNN architectures. Parameter counts correspond to the full model including the replaced 12-class classification head. FLOPs are reported for a single forward pass at 224×224 input resolution.

For each architecture, the pretrained classification head was replaced with a new fully connected layer whose output dimensionality matches the 12 shared classes defined after label harmonization. All remaining layers were initialized with ImageNet-pretrained weights [9]. This ensures that all three backbones are evaluated under an identical class space and that architectural comparisons are not confounded by differences in output definition. Each of these architectures is trained under four augmentation pipelines of increasing strength, designed to test whether source-domain data diversity can reduce the controlled-to-field generalization gap.

3.5 Training Pipelines and Augmentation Strategy

The goal for the four distinct pipelines for testing augmentation and regularization in the context of domain shift is to test two variables separately with the same set of training data for the PlantVillage dataset. All four pipelines use the same validation splits and evaluation transforms; therefore, the only other variable is the augmentation strategy used in the training phase for the source domain.

The baseline pipeline, P1, does not augment the images beyond resizing, tensor conversion, and ImageNet normalization [19] in order to measure how much each model overfits to the "clean" visual characteristics of PlantVillage when no additional source-domain variability is introduced.

The second pipeline, P2, adds a moderate level of torchvision-based augmentation to the images to create benign acquisition variability that is not too de-

structive to the overall shape of the lesions. The transform sequence consists of `RandomResizedCrop(224, scale=(0.8, 1.0))`, `RandomHorizontalFlip(p=0.5)`, `RandomVerticalFlip(p=0.2)`, `RandomRotation(degrees=20)`, `ColorJitter(brightness=0.2, contrast=0.2, saturation=0.1)` followed by tensor conversion and ImageNet normalization. Compared with P1, P2 increases appearance diversity while remaining relatively conservative. The parameters for the specific pipeline P2 were selected as a way to create reasonable variations in how images are acquired, but still conservatively enough so that the morphologic features of symptoms of diseases can remain intact. As such, the upper and lower limits of (0.8, 1.0), of the crop scale, will ensure that at least 80% of the original image will continue to be displayed, thus precluding excessive loss of details regarding lesions due to over-advanced cropping. Since most leaf structures display some form of bilaterally symmetric structure, the value for the horizontal flip probability of 0.5, represents this characteristic well. However, it is less common for leaves to be captured with their vertical axis flipped. Therefore, the value for the vertical flip probability was set at 0.2. The choice of a rotation angle range of ± 20 degrees represents an attempt to model the natural variations in leaf orientation during image capture. However, it also does not represent extremely large rotations which may occur in actual practice. Lastly, the choices of jittering values for the brightness and contrast of 0.2 and for the saturation jitter of 0.1 represent moderate levels of illumination variations while representing large enough color variations to make visually distinguishable changes to symptoms associated with color differences along the edges of lesions.

P3 uses a more advanced version of the field-based transformation from the previous pipeline P2. The transformation order for P3 includes `RandomResizedCrop(scale=(0.8,1.0))`, random horizontal and vertical flips, rotation by up to 30° , `ColorJitter` with larger changes, and additional environmental augmentations including `RandomShadow`, `RandomFog`, `MotionBlur`, `GaussNoise`, and `CoarseDropout`, prior to normalization and conversion to tensors.

The mixed-image regularization parameters for Pipeline P3 were configured following the original MixUp and CutMix recommendations. The MixUp interpolation co-

efficient $\alpha = 0.4$ was chosen, producing mixing weights sampled from a $\text{Beta}(0.4, 0.4)$ distribution. This distribution is bimodal and tends to generate mixing ratios close to 0 or 1, meaning that most mixed images remain visually dominated by one of the two source images, preserving lesion recognizability. A CutMix coefficient $\alpha = 1.0$ results in uniform distributions of the size of patches encountered by the model; therefore, both smaller and larger replacement of patches can be experienced while training.

An average activation probability of 0.5 (i.e., on average) allows for a balanced exposure to both clean images and mixed images over all of the training batches. Additionally, when the mixing is active an equal selection probability of 0.5 will be used to select which method of mixing (Mixup or Cutmix) will be applied as a means to prevent one type of regularizing signal from being dominant.



Figure 3-3: Representative PlantVillage samples processed through each training pipeline. Each row shows the same leaf image after P1 (no augmentation), P2 (basic geometric/photometric), P4 (environmental noise only), and P3 (environmental + MixUp/CutMix). The progressive increase in visual distortion is evident across columns. Additional samples across all four pipelines are shown in Figure A-2.

Building on the environmental portion of P3, the Pipeline P4 (EnvAugOnly) configuration was developed. The purpose of this configuration is a controlled ablation study. The main difference between configurations P3 and P4 is that both apply the same environmental transformations (RandomShadow, RandomFog, MotionBlur, GaussNoise, and CoarseDropout) but P4 removes MixUp and CutMix completely. This separation is important for methodological purposes; by including both environmental distortion and the structural image mixing at the same time, it is not possible to identify which factor(s) are causing the changes in performance. The ability to compare P4 directly with P3 allows researchers to assess whether performance changes occur due to model difficulties associated with field-like visual noise or due to the image-mixing regularization used to obscure local lesion features needed to accurately diagnose diseases.

The environmental augmentations in P3 namely RandomShadow, RandomFog, MotionBlur, and GaussNoise were included because each of these transformations corresponds to a specific type of visual degradation observed in the PlantDoc field images. RandomShadow simulates the effect of uneven canopy cover and partial shading by other plants or structures. RandomFog approximates the haze and reduced contrast that occur during humid or early-morning field conditions. MotionBlur simulates the effect of camera movement during handheld image capture in the field. GaussNoise models the sensor noise that is more prominent in images taken under lower light conditions or with lower-quality camera hardware. CoarseDropout was included to simulate partial occlusion from overlapping leaves, which is common in dense canopy environments and is a feature that distinguishes field images from the cleanly isolated leaves in PlantVillage.

Table 3.4 summarizes the correspondence between the environmental conditions commonly encountered in field deployment and the augmentation transforms included in the four pipelines. While Pipeline P3 covers several major sources of visual degradation, certain real-world conditions are not explicitly simulated. These include water droplets and dew on leaf surfaces, soil or dust contamination, insect damage co-occurring with disease symptoms, and the effect of leaves at varying growth stages.

The absence of these conditions from the augmentation pipeline may partially explain the persistent domain gap even under the most advanced augmentation setting, since the synthetic training distribution still differs from the full range of field variability.

Field Condition	P1	P2	P3	P4
Varying leaf orientation	–	✓	✓	✓
Illumination variation	–	✓	✓	✓
Scale / distance changes	–	✓	✓	✓
Shadow / uneven lighting	–	–	✓	✓
Fog / haze	–	–	✓	✓
Motion blur	–	–	✓	✓
Sensor noise	–	–	✓	✓
Partial occlusion	–	–	✓	✓
Soil / dust on leaves	–	–	–	–
Water droplets / dew	–	–	–	–
Insect damage co-occurrence	–	–	–	–
Multiple leaves in frame	–	–	–	–
Mixed-Image Regularization (MixUp/CutMix)	–	–	✓	–

Table 3.4: Correspondence between real-world field conditions and augmentation transforms across the four training pipelines. A check mark (✓) indicates that the condition is explicitly simulated by at least one transform in the pipeline; a dash (–) indicates no coverage. Pipeline P1 applies no augmentation; P2 covers basic acquisition variability; P3 provides the most advanced augmentation including environmental degradations and mixed-image regularization; and P4 isolates the environmental degradations from P3 to serve as an ablation study. Several common field conditions remain unsimulated across all pipelines.

In addition, the goal of the design is to subject the model to a greater variety of environmental conditions in comparison to the original image, similar to what would be found in an actual field environment. The types of transformations used are based on what has been seen in PlantDoc; specifically illumination variation, par-

tial occlusions, motion blur, and environmental effects including shadows and haze. Additionally, Pipeline P3 applies mixed-image regularization during training through MixUp and CutMix. The proposed pipeline uses MixUp with $\alpha = 0.4$, CutMix with $\alpha = 1.0$, activates mixed-image augmentation at the batch level with probability 0.5, and chooses MixUp versus CutMix with probability 0.5 when mixing is triggered. Thus, P3 is not simply a stronger version of P2; it combines environmental augmentation with label-mixing regularization in the same training pipeline. However, because P3 introduces two significant variables simultaneously—environmental distortions and structural image mixing a fourth configuration, Pipeline P4 (EnvAugOnly), was implemented as a controlled ablation study. P4 utilizes the exact suite of environmental augmentations found in P3 (such as fog, shadows, and motion blur) but excludes MixUp and CutMix. The inclusion of P4 is methodologically vital; it allows the study to decouple the impact of visual noise from the impact of mixed-image regularization. This evaluation will provide sufficient information as to whether the degradation of performance experienced with the more sophisticated pipelines results from the model’s limitations with handling very noisy data in a manner similar to that observed in real-world fields (i.e., field-like noise), or due to the masking effect of the mixing logic on local fine-scale detail which are required to accurately identify various forms of plant pathology.

3.6 Training Procedure and Hyperparameter Selection

3.6.1 Source-Domain Training

The PlantVillage data set (filtered) was separated into training and validation sets through an 80 / 20 stratified method so as to maintain the number of instances per class within each subset. Thus the 80 / 20 split provided a balance between sufficient training samples to allow for discrimination among all twelve classes and sufficient validation samples to allow for reliable early stopping and scheduling of the

learning rate. By employing stratification, it was ensured that each class had the same representation percentage-wise in both subsets; this was important since there was some degree of class imbalance in the filtered PlantVillage data set. Consistent ratios for splitting and consistent seeds were used in all architectures and pipelines to avoid confusion regarding whether observed differences in performance were due to differences in the architecture or augmentation strategies versus differences in how the data was being split.

As such, all experiments were run at a fixed random seed (of 42), with seeded workers and determinate settings enabled where possible by the frameworks used, to improve reproducibility. All experiments were run at a batch size of 32 for all architectures evaluated. Larger batch sizes were constrained by the available GPU memory, particularly for the ResNet-50 architecture, while smaller batch sizes were avoided because they increase gradient noise and, under class-weighted loss, can produce batches in which minority classes are absent, leading to unstable weight updates. A batch size of 32 is also commonly used in comparable plant disease classification studies, which supports comparability with prior work.

To reduce the effect of class imbalance and overconfident predictions, the models were optimized using class-weighted cross-entropy with label smoothing of 0.1. The class weights were computed from inverse class frequencies on the training subset and normalized before being passed to the loss function. Optimization was performed using the AdamW optimizer together with a ReduceLROnPlateau scheduler, and early stopping was applied to limit overfitting. Mixed-precision training was enabled when CUDA was available to improve computational efficiency. Per-configuration early stopping epochs are reported in Figure A-3.

A two-phase transfer learning strategy was adopted for source-domain training. In Phase 1, the pretrained backbone was frozen and only the newly replaced classification head was trained for 5 epochs. This warm-up stage allowed the classifier layer to adapt to the 12-class plant disease task without immediately perturbing the pretrained feature extractor. In Phase 2, the full network was unfrozen and fine-tuned end-to-end for up to 25 additional epochs using a reduced learning rate, giving a maximum total of

30 training epochs across both phases. As is common with two phase transitions that follow the general structure of typical transfer learning methods this training process resulted in greater stability during optimization and faster convergence. Convergence plots for each architecture under pipeline p2 can be found in appendix figure A-1.

Hyperparameters that were ultimately chosen were from a previous random search procedure where many different configurations were tested based on source domain validation performance . The best performing configuration for each architecture was then selected and those configuration settings were carried forward into the current study [23]. Those selected values included: Resnet50 had an initial learning rate set at 1×10^{-3} and a weight decay set at 1×10^{-4} ; EfficientNetB0 had an initial learning rate of 5.034×10^{-4} and a weight decay set at 4.994×10^{-5} ; MobileNetV3Large had an initial learning rate of 9.802×10^{-4} and a weight decay set at 2.18×10^{-4} . All of these were constant throughout the training, zero shot testing, and domain adaption studies to provide consistency across models and pipeline comparisons.

3.6.2 Target-Domain Fine-Tuning

For the supervised domain adaptation method, source-trained checkpoints were trained in a multi-data regime with a two-stage approach on the filtered PlantDoc Training Set. This multi-regime strategy allowed us to assess the adaptability of the fine-tune training and allow for sampling variation. The plantdoc training dataset was subsampled as equally distributed as possible into six fractions that included; 10 % (about 100 images) , 20 % (roughly 200 images), 30 % (approximately 300 images), 50 % (about 500 images), 70 % (nearly 700 images), and 100 % (all images) of the total number of available images. Given the limited size of the smallest datasets used (i.e., less than or equal to 10% of the original dataset or roughly <100 images), we also needed to have statistically reliable results from our data. Therefore, all of the above mentioned sub-sampling and training processes were repeated in triplicate using three different random seed numbers (42, 123, and 7). In each of the above-mentioned runs, the fraction that was selected for training was then divided in half again to create an “adaptation” training set and an “internal-validation” test set based

on a stratified 80/20 split to provide some measure of fine-tune progress as well as enable early stopping decisions. All models were ultimately evaluated on the fixed, untouched PlantDoc test split (118 images), with performance reported as the mean and standard deviation across the three seeds.

In Phase A, the pretrained backbone was frozen and only the classification head was retrained for 3 epochs using the AdamW optimizer with a learning rate of 3×10^{-4} and a ReduceLROnPlateau scheduler with a reduction factor of 0.5 and patience of 1. This brief head-only phase allows the classifier to recalibrate to the target-domain feature distribution before the full network is modified. In Phase B, the full network was unfrozen and trained end-to-end for up to 12 epochs using a reduced learning rate of 8×10^{-5} , with the same AdamW optimizer and a weight decay of 1×10^{-4} . The scheduler patience was increased to 2 in this phase, and early stopping with patience 4 was applied based on validation macro-F1, with the best checkpoint restored for final evaluation.

The batch size of 32 was retained from source training, and class-weighted cross-entropy was used as the loss function. However, label smoothing was reduced from 0.1 to 0.05 during fine-tuning. This reduction reflects the smaller target-domain dataset and the model’s already-established task knowledge: because the feature extractor has already learned disease-relevant representations during source training, a sharper supervision signal is more appropriate than the stronger regularization needed during initial training from pretrained weights.

3.7 Evaluation Protocol and Metrics

Model performance is reported using four metrics: accuracy, macro-precision, macro-recall, and macro-F1. Among these, macro-F1 is treated as the primary summary metric because it weights all classes equally regardless of support, making it more informative than accuracy alone when class frequencies are uneven across disease categories. AUC (Area Under the ROC Curve) is also reported for zero-shot results to assess whether the models retain discriminative ranking ability even when their

hard class predictions are incorrect.

Two distinct evaluation protocols are used, corresponding to the two target-domain set definitions established in Section 3.2. Protocol A (zero-shot cross-domain evaluation) trains models exclusively on PlantVillage and evaluates them directly on the PlantDoc FULL set without any target-domain supervision. Protocol B (supervised domain adaptation) initializes from the PlantVillage-trained checkpoints, fine-tunes on the PlantDoc training split, and evaluates on the held-out PlantDoc test split. The two protocols answer different research questions and are therefore reported separately throughout the thesis: Protocol A measures how well source-domain representations transfer without adaptation, while Protocol B measures how much of the lab-to-field gap can be reduced once labeled target-domain data are introduced.

3.7.1 Implementation Environment

All experiments were implemented in Python 3.10 using PyTorch 2.1 as the primary deep learning framework. Model architectures and pretrained ImageNet weights were loaded from the `torchvision.models` module. Data augmentation for Pipelines P1 and P2 was implemented using `torchvision.transforms`, while the environmental augmentations in Pipeline P3 were implemented using the Albumentations library (version 1.3). Evaluation metrics were computed using `scikit-learn` (version 1.3).

Training was performed on a single NVIDIA GeForce RTX 5080 GPU with 16 GB of video memory (VRAM). Mixed-precision training was enabled via PyTorch’s `torch.cuda.amp` module, which uses automatic casting between 32-bit and 16-bit floating-point representations to reduce memory consumption and accelerate matrix operations without measurable loss in model accuracy. The use of mixed precision was particularly important for ResNet-50, whose 23.5 million parameters and 4.09 GFLOPs per forward pass would otherwise approach the memory ceiling of the available GPU at batch size 32. Training times per model-pipeline configuration are reported in Figure A-4

Deterministic behavior was enforced by setting a fixed random seed of 42 across all random number generators (Python, NumPy, PyTorch, and CUDA). Additionally,

`torch.backends.cudnn.deterministic` was set to `True` and `torch.backends.cudnn.benchmark` was set to `False` to ensure reproducibility of convolutional operations across runs. `DataLoader` workers were seeded using a worker initialization function to prevent non-deterministic behavior during parallel data loading.

Chapter 4

Results and Discussion

This chapter presents the empirical results of the study under two distinct evaluation protocols. The first protocol measures zero-shot cross-domain generalization by training the models on the shared-class PlantVillage source domain and evaluating them directly on the filtered PlantDoc FULL target set. The second protocol measures supervised domain adaptation by initializing from the PlantVillage-trained checkpoints, fine-tuning on the filtered PlantDoc training split, and evaluating on the untouched filtered PlantDoc test split. These two protocols must be interpreted separately because they answer different questions: the first evaluates out-of-domain transfer without target supervision, while the second evaluates how much the domain gap can be reduced once labeled target-domain data are introduced.

4.1 Evaluation Sets and Reporting Conventions

All reported results are based on the harmonized 12-class subset obtained after label mapping and shared-class filtering. For zero-shot evaluation, the target set is the filtered PlantDoc FULL dataset containing 1405 images, created by merging the filtered PlantDoc training and test partitions. For domain adaptation, the model is fine-tuned on the filtered PlantDoc training split of 1287 images and evaluated on the untouched filtered PlantDoc test split of 118 images. Because these protocols use different target-domain sample definitions, their results should not be compared as if

they were measured on the same evaluation set.

Performance is reported using accuracy and macro-F1 as the primary summary metrics. Macro-F1 is especially important in this study because it weights all disease classes equally and therefore gives a more balanced view of cross-domain behavior than accuracy alone. In the discussion below, source-domain validation performance is not treated as evidence of real-world robustness; instead, the focus is on target-domain behavior under domain shift.

4.2 Zero-Shot Cross-Domain Results

Table 4.1 summarizes zero-shot cross-domain performance for all three architectures and all three source-domain training pipelines. The results show a severe lab-to-field generalization gap. Although the models achieve very strong performance on source-domain validation during training, their accuracy on PlantDoc FULL remains low, ranging from 17.86% to 24.84%, with macro-F1 values ranging from 0.1318 to 0.1966. This shows that good performance using controlled laboratory imagery does not translate well to field imagery due to differences of background, lighting, occlusion and capture artifact. In comparison, the zero-shot results are at least 16 percentage points above random chance (approximately 8.3%), and as such indicate that while there is certainly discriminative information transferred from the source domain, it is still insufficient for field use.

An additional observation from Table 4.1 is that the AUC scores are substantially higher than the accuracy values across all configurations, ranging from 0.6713 to 0.7380 compared to accuracies of 17.86% to 24.84%. This discrepancy indicates that the models retain a degree of discriminative ranking ability even when their hard class predictions are incorrect. In other words, it is very common for the correct class to have one of the highest or most likely predictions from the model but not necessarily the top prediction. The fact that source-domain representation has captured some of this potentially useful information that is degraded, but not completely lost when moving to the target domain also provides insight into how well the source domain will

support the use of these models in real world applications where you are satisfied with getting either a top- k prediction or a prediction above a certain confidence threshold (i.e. as opposed to an exact single classification)

Architecture	Pipeline	Acc (%)	Precision	Recall	Macro-F1	AUC
ResNet-50	P1 (NoAug)	23.91	0.3260	0.2545	0.1966	0.6846
ResNet-50	P2 (BasicAug)	24.84	0.2834	0.2729	0.1927	0.7177
ResNet-50	P3 (AdvAug)	19.86	0.3326	0.2193	0.1681	0.6713
ResNet-50	P4 (EnvAugOnly)	22.28	0.3715	0.2292	0.1831	0.6903
EfficientNet-B0	P1 (NoAug)	20.78	0.3210	0.2239	0.1605	0.6947
EfficientNet-B0	P2 (BasicAug)	22.99	0.3193	0.2675	0.1936	0.7122
EfficientNet-B0	P3 (AdvAug)	19.50	0.3675	0.2215	0.1494	0.7380
EfficientNet-B0	P4 (EnvAugOnly)	17.22	0.3729	0.1997	0.1395	0.7101
MobileNetV3-Large	P1 (NoAug)	21.49	0.3926	0.2393	0.1841	0.7213
MobileNetV3-Large	P2 (BasicAug)	21.49	0.2770	0.2330	0.1784	0.6994
MobileNetV3-Large	P3 (AdvAug)	17.86	0.3036	0.2091	0.1318	0.6879
MobileNetV3-Large	P4 (EnvAugOnly)	18.36	0.3818	0.2223	0.1441	0.6749

Table 4.1: Zero-shot cross-domain performance (train on PlantVillage; evaluate on PlantDoc FULL, 1405 images, 12 classes) comparing four training pipelines. P4 serves as an ablation study to isolate the effects of environmental augmentation from the mixed-image regularization used in P3. Best result per metric is shown in **bold**.

Several key findings can be inferred from Table 4.1 . In particular, no model has been able to reach the reported levels of performance found in many PlantVillage within-dataset experiments documented in the literature. Therefore, it appears that a classification capacity issue does not represent the major problem facing this dissertation, but rather the difference between the training (source) and testing (target) data distributions. Additionally, as evidenced by Table 4.1, ResNet-50 achieves the best overall zero-shot accuracy with an accuracy of 24.84% when using pipeline P2; however, as also seen in Table 4.1, EfficientNet-B0 uses pipeline P2 to achieve the largest macro-f1 score of 0.1936. The differences noted above are very minimal and therefore do not provide any evidence that would suggest that either or both backbone architectures have successfully resolved the issues associated with the source-to-target data distribution.

4.3 Ablation Study: Impact of Augmentation Strength and Image-Mixing Regularization.

A central goal of the thesis is to test whether stronger augmentation improves cross-domain robustness. The results show that this is only partially true. Pipeline P2, which applies moderate geometric and photometric augmentation, provides the most reliable zero-shot improvement over the baseline P1. For ResNet-50, accuracy increases from 23.91% to 24.84%. The highest accuracy for EfficientNet-B0 increased by 2.21 percentage points (from 20.78% to 22.99%), while the highest macro-F1 for this model increased by 3.31 percentage points (from 0.1605 to 0.1936). The highest accuracy for MobileNetV3-Large did not change using the two models of P2; it remained at 21.49%. Macro-F1 decreased slightly compared to P1. Overall, P2 has the least variation in results between models.

In contrast, the results for Pipeline P3 (AdvAug) show a consistent loss of performance across all architectures. This deterioration is specifically illuminated by the results of the P4 (EnvAugOnly) ablation study. Across all models, P4 outperformed P3 in zero-shot accuracy; for instance, ResNet-50 achieved 22.28% accuracy under P4 compared to only 19.86% under P3. This gap provides empirical evidence that the mixed-image regularization (MixUp/CutMix) in P3 is the primary driver of performance loss. While the environmental transforms in P4 are difficult for the model to handle, the addition of image mixing in P3 further destroys the localized lesion textures and morphological signatures required for the model to generalize. This confirms that the models are more sensitive to the loss of structural information (due to CutMix) than they are to visual noise (due to environmental transforms). However, since P4 still underperforms relative to the basic augmentations of Pipeline P2 (24.84%), it is evident that advanced environmental distortions alone create a training distribution that is too far removed from the 'clean' source domain to provide a reliable benefit for zero-shot transfer.

The overall findings do not indicate that augmentation can improve cross-dataset generalization in all cases. Instead, the overall evidence suggests that augmentation

will only be useful when both the degree and form of augmentation remain consistent with the level of variability of the target domain.

Visual support for this interpretation is provided by the source-domain training dynamics in Figures 4-1, 4-2, 4-3. The training and validation curves for the source domain (i.e., the PlantVillage domain) provide additional evidence that high validation performance was obtained during source-domain training, yet poor target-domain zero-shot performance still occurred. Therefore, validation performance on the source domain alone cannot provide adequate evidence that cross-domain generalization has been achieved.

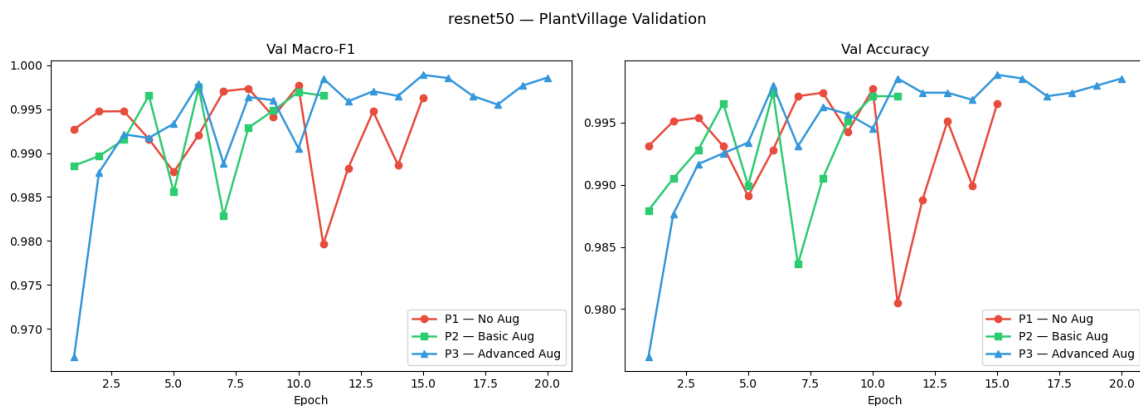


Figure 4-1: Training and validation curves for ResNet-50 on PlantVillage, showing strong in-domain optimization despite weak zero-shot transfer.

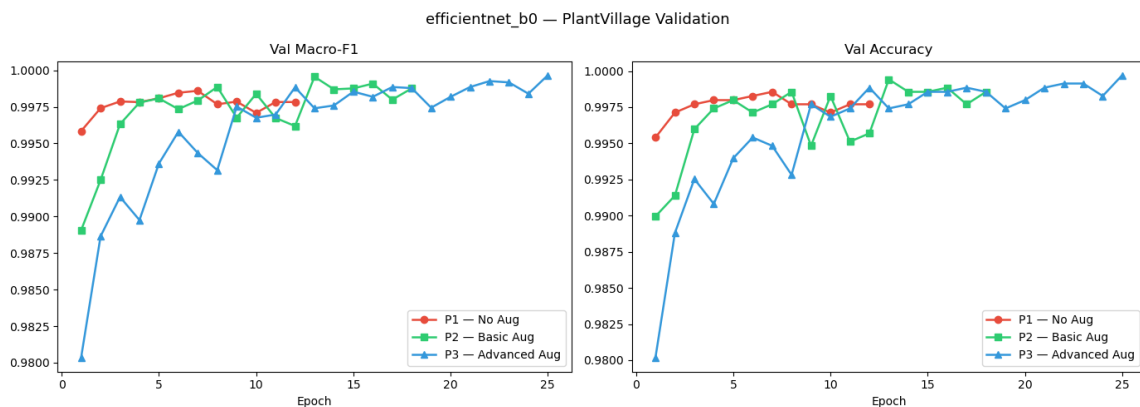


Figure 4-2: Training and validation curves for EfficientNet-B0 on PlantVillage, showing strong in-domain optimization despite weak zero-shot transfer.

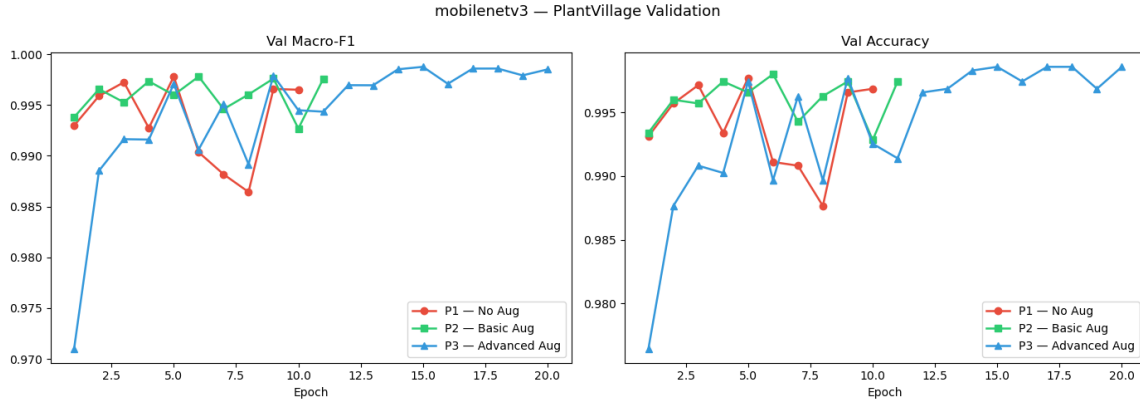


Figure 4-3: Training and validation curves for MobileNetV3-Large on PlantVillage, showing strong in-domain optimization despite weak zero-shot transfer.

4.4 Architecture Comparison Under Domain Shift

When comparing architectures based on the zero-shot evaluation, the distinction between them is relatively minor when compared to the scope of the domain disparity. Under Pipeline 2, ResNet-50 achieved the best zero-shot accuracy; however, under the same pipeline, EfficientNet-B0 yielded virtually equivalent macro-F1 levels. The MobileNetV3-Large performed generally lower than the two other architectures in each setting. However, there were some instances where it remained competitive as part of the lighter pipelines. This suggests that while architecture choice may influence what level of zero-shot performance will be realized, the generalization problem of source-trained models toward PlantDoc is not changed by architecture choice.

This result is significant to the understanding of prior work. Many plant disease studies report high accuracy using modern CNN architectures, but those results are often obtained under within-domain or laboratory-style evaluation. The present experiments show that even well-established pretrained backbones such as ResNet-50, EfficientNet-B0, and MobileNetV3-Large remain vulnerable when the deployment domain differs substantially from the source training data. Consequently, architecture selection alone should not be treated as a sufficient solution to real-world robustness.

4.5 Domain Adaptation Results

The second protocol evaluates whether target-domain supervision can reduce the observed lab-to-field gap. In this setting, the source-trained models are fine-tuned on the filtered PlantDoc training split and then evaluated on the untouched filtered PlantDoc test split. The results show a substantial improvement for all three models after fine-tuning. Table 4.2 reports domain adaptation results when fine-tuning from P3-initialized source checkpoints. This configuration tests whether advanced source-domain augmentation provides a better starting point for target-domain adaptation.

Architecture	After Fine-tune Acc (%)	After Fine-tune Macro-F1
ResNet-50	44.92	0.4212
EfficientNet-B0	38.98	0.3702
MobileNetV3-Large	33.05	0.3045

Table 4.2: Domain adaptation results on the filtered PlantDoc test split (118 images, 12 classes) using source models initialized from P3-trained (advanced augmentation) checkpoints.

The gains over the zero-shot baseline are substantial across all three architectures: ResNet-50 reaches 44.92% accuracy and 0.4212 macro-F1, EfficientNet-B0 reaches 38.98% accuracy and 0.3702 macro-F1, and MobileNetV3-Large reaches 33.05% accuracy and 0.3045 macro-F1. These improvements far exceed any gains obtained by changing the augmentation pipeline alone, confirming that exposure to labeled target-domain data is a more effective intervention than source-only augmentation for reducing the domain gap. Table 4.3 then reports the best-performing adaptation configuration obtained in this study, which uses P2-initialized source checkpoints rather than P3.

Architecture	Accuracy (%)	Macro-F1
ResNet-50	55.93	0.5491
EfficientNet-B0	57.63	0.5393
MobileNetV3-Large	59.32	0.5734

Table 4.3: Maximum supervised domain adaptation performance on the PlantDoc test split (118 images, 12 classes). Results represent the performance ceiling achieved by initializing models from the optimal P2 source-domain checkpoints and fine-tuning on 100% of the available target-domain training data.

These results represent the best-performing configuration obtained in the study and show that simpler source-domain initialization followed by supervised target-domain fine-tuning is more effective than initializing adaptation from the most advanced augmentation pipeline. Among the three architectures, MobileNetV3-Large achieves the highest accuracy at 59.32% and the highest macro-F1 at 0.5734, followed closely by EfficientNet-B0 and ResNet-50. This ordering differs from the zero-shot setting and further supports the view that cross-domain robustness and adaptation behavior cannot be inferred solely from source-domain optimization performance.

While these results establish the performance ceiling using the full target-domain dataset, practical agricultural deployments often face severe data collection constraints. To understand how much target-domain data is practically required to achieve these gains, the models were further evaluated across stratified data fractions ranging from 10% to 100%, averaged over three random seeds to account for variance. Following the comparative analysis in Section 4.3 which utilized the P4 ablation study to identify Pipeline P2 as the most effective source-domain configuration this pipeline was selected as the fixed initialization for the data-efficiency study.

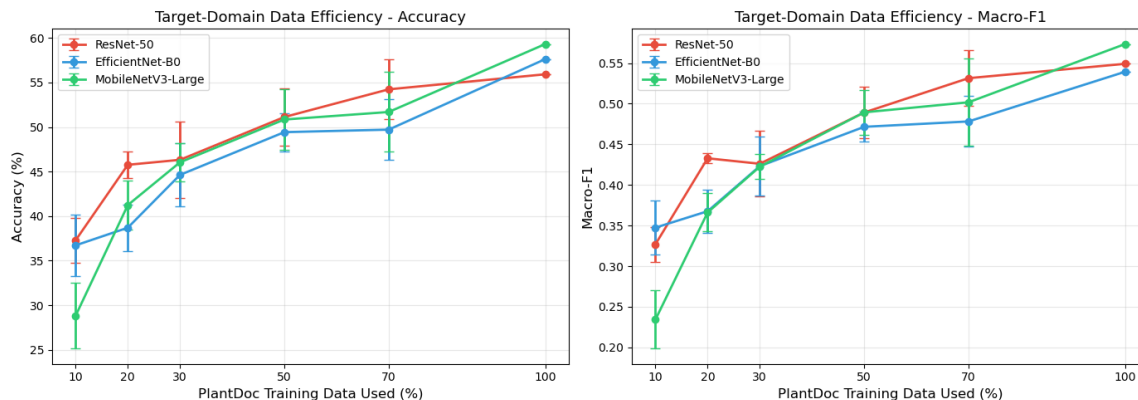


Figure 4-4: Target-domain data efficiency learning curves. Mean Accuracy (left) and Macro-F1 (right) evaluated on the PlantDoc test split, shown as a function of the percentage of PlantDoc training data used for fine-tuning. Models were initialized from Pipeline P2 (BasicAug) source-domain checkpoints, which were identified in the P4 ablation study as providing the most stable foundation for adaptation. Error bars indicate the standard deviation across 3 random seeds.

The learning curves in Figure 4-4 reveal that performance improves sublinearly

as more target-domain data is introduced, indicating a high degree of data efficiency. Even at extreme low-data regimes, the pretrained representations adapt rapidly. When fine-tuned on just 10% of the PlantDoc training data (102 images), ResNet-50 achieves 37.29% mean accuracy, which easily surpasses its zero-shot baseline of 24.84%. Similarly, EfficientNet-B0 reaches 36.72% mean accuracy with 10% data, up from its 22.99% zero-shot performance. As the data fraction increases, diminishing returns become apparent. For instance, using 50% of the target data (514 images), ResNet-50 recovers 51.13% mean accuracy remarkably close to its maximum performance of 55.93% achieved using the full 100% dataset. MobileNetV3-Large exhibits a slightly steeper learning curve, climbing from 28.82% at the 10% fraction to 51.70% at the 70% fraction, before peaking at 59.32% with the full dataset. These findings demonstrate that while the initial domain shift is severe, overcoming it does not strictly require massive field-collected datasets. The pretrained source representations are highly data-efficient and can be successfully recalibrated to field conditions with a relatively small annotation budget of roughly 300 to 500 target-domain images.

To make this transition from zero-shot to fine-tuned performance easier to interpret, the comparison is visualized in Table 4.4. Such a figure provides a compact summary of the central empirical result of the thesis, namely that supervised target-domain fine-tuning leads to a much larger improvement than changing augmentation policy alone.

Architecture	Zero-Shot			Fine-Tuned		
	Acc (%)	Macro-F1	Prec.	Acc (%)	Macro-F1	Prec.
ResNet-50	24.84	0.1927	0.2834	55.93	0.5491	0.5628
EfficientNet-B0	22.99	0.1936	0.3193	57.63	0.5393	0.5392
MobileNetV3-Large	21.49	0.1784	0.2770	59.32	0.5734	0.5918

Table 4.4: Comparison of zero-shot (Pipeline P2, PlantDoc FULL) and best fine-tuned performance (PlantDoc test split, 118 images). Both use P2-trained source checkpoints.

Class	Zero-Shot F1	Fine-Tuned F1	Improvement	Support
Tomato_Leaf_Mold	0.2000	0.2857	+0.0857	6
Tomato_Early_blight	0.4242	0.6000	+0.1758	9
Potato___Early_blight	0.2105	0.4348	+0.2243	14
Potato___Late_blight	0.1818	0.4211	+0.2393	8
Tomato_Bacterial_spot	0.0000	0.3333	+0.3333	9
Tomato_healthy	0.0000	0.3636	+0.3636	8
Pepper__bell___Bacterial_spot	0.4000	0.7778	+0.3778	9
Tomato_YellowLeaf__Curl_Virus	0.3636	0.7692	+0.4056	15
Pepper__bell___healthy	0.4667	0.8889	+0.4222	8
Tomato_Septoria_leaf_spot	0.2353	0.6667	+0.4314	12
Tomato_Late_blight	0.2500	0.7000	+0.4500	10
Tomato__Tomato_mosaic_virus	0.0000	0.6400	+0.6400	10

Table 4.5: Per-class F1 scores comparing the best zero-shot model (ResNet-50, Pipeline P2) and the best fine-tuned model (MobileNetV3-Large, Pipeline P2), both evaluated on the filtered PlantDoc test split (118 images, 12 classes). Classes are sorted by improvement magnitude.

4.6 Error Analysis

Although fine-tuning substantially improves performance, errors remain on the PlantDoc test set. These misclassifications are important because they reveal which aspects of the deployment domain remain difficult even after supervised adaptation. To provide a quantitative basis for this analysis, Table 4.5 reports per-class F1 scores for both the best zero-shot and best fine-tuned configurations, evaluated on the same 118-image PlantDoc test split to ensure a fair comparison. Due to the small per-class support (ranging from 6 to 15 images), individual F1 scores should be interpreted as indicative of broad trends rather than precise estimates.

4.6.1 Zero-Shot Error Patterns

The zero-shot per-class results in Table 4.5 reveal that the domain shift does not affect all classes equally. Three classes are completely unrecognizable in the zero-shot setting: `Tomato_healthy` ($F1 = 0.0000$), `Tomato__Tomato_mosaic_virus` ($F1 = 0.0000$), and `Tomato_Bacterial_spot` ($F1 = 0.0000$). In contrast, `Pepper__bell___healthy` achieves an $F1$ of 0.4667 and `Tomato_Early_blight` reaches 0.4242, suggesting that their visual symptoms are distinctive enough to partially survive the domain shift.

The confusion pair analysis provides further insight into why certain classes fail. The most frequent error is the misclassification of `Potato___Early_blight` as `Pepper__bell___Bacterial_spot` (7 errors), followed by `Tomato_Septoria_leaf_spot` misclassified as `Tomato_Early_blight` (5 errors) and `Tomato__Tomato_mosaic_virus` misclassified as `Pepper__bell___healthy` (5 errors). These patterns indicate that the model’s learned decision boundaries collapse under domain shift, mapping visually ambiguous field images onto a small number of high-confidence source-domain categories. The `Pepper__bell___Bacterial_spot` and `Pepper__bell___healthy` classes act as attractors, collectively absorbing misclassified samples from multiple other classes.

The complete failure of the `Tomato_healthy` class ($F1 = 0.0000$) is particularly notable. Healthy leaves in PlantVillage have uniform green coloring against clean backgrounds, whereas healthy leaves in PlantDoc appear within cluttered field scenes with variable lighting and surrounding vegetation. Without the clean background as a distinguishing cue, the model cannot separate healthy from diseased tissue, and all 8 healthy samples are redistributed across disease classes, with 4 being misclassified as `Tomato_Late_blight`.

4.6.2 Residual Errors After Fine-Tuning

After fine-tuning, all classes show improvement, though the magnitude varies considerably. The strongest gains are observed for the three classes that completely failed

in the zero-shot setting: `Tomato__Tomato_mosaic_virus` improves from 0.0000 to 0.6400, `Tomato_healthy` from 0.0000 to 0.3636, and `Tomato_Bacterial_spot` from 0.0000 to 0.3333. At the other end, `Tomato_Leaf_Mold` shows the smallest improvement (+0.0857), remaining the weakest class after fine-tuning at $F1 = 0.2857$. This suggests that the limited number of Leaf Mold samples in the PlantDoc training split (support = 6 in the test set) may be insufficient for the model to learn reliable field-domain features for this class.

The dominant error patterns also change qualitatively after fine-tuning. The most frequent confusion pair shifts to `Potato___Early_blight` misclassified as `Potato___Late_blight` (7 errors), which is a biologically meaningful error because early and late blight on potato leaves can be visually similar in intermediate disease stages, where lesion color and size have not yet fully differentiated. Additional within-crop confusion is observed between `Tomato_Bacterial_spot` and `Tomato_Septoria_leaf_spot` (3 errors) and between `Tomato__Tomato_YellowLeaf__Curl_Virus` and `Tomato__Tomato_mosaic_virus` (3 errors). The `Tomato_healthy` class, while improved, still has 3 samples misclassified as `Tomato__Tomato_mosaic_virus`, suggesting difficulty in distinguishing healthy leaves from leaves with subtle viral symptoms in field conditions.

4.6.3 Implications for Deployment

The two implications for the application of the class-by-class analysis are: Firstly, the per-class analysis illustrates how domain adaptation can be achieved with fine tuning which will not only enhance average performance as well as alter the type of error. The zero shot errors are primarily caused by the differences in domains (cross-domain) related to background and lighting conditions. In these cases, the predictions fall into a couple of attractor classes. Following fine-tuning, the errors shift from confusing visually similar diseases in the same crop to the biologically meaningful confusions of disease relevant features instead of domain specific artifacts.

Second, the classes that remain difficult after fine-tuning particularly `Tomato_Leaf_Mold` ($F1 = 0.2857$), `Tomato_Bacterial_spot` ($F1 = 0.3333$), and

Tomato_healthy ($F1 = 0.3636$)—share two characteristics: low support in the PlantDoc test split and subtle visual symptoms that are difficult to distinguish from other conditions. Further improvement for these classes would likely require larger target-domain training sets, disease-stage-aware data collection, or the incorporation of additional discriminative cues beyond single-image appearance.

To support this discussion, a panel of representative misclassified examples is shown in Figure 4-5 to illustrate the visual characteristics of failure cases more concretely.

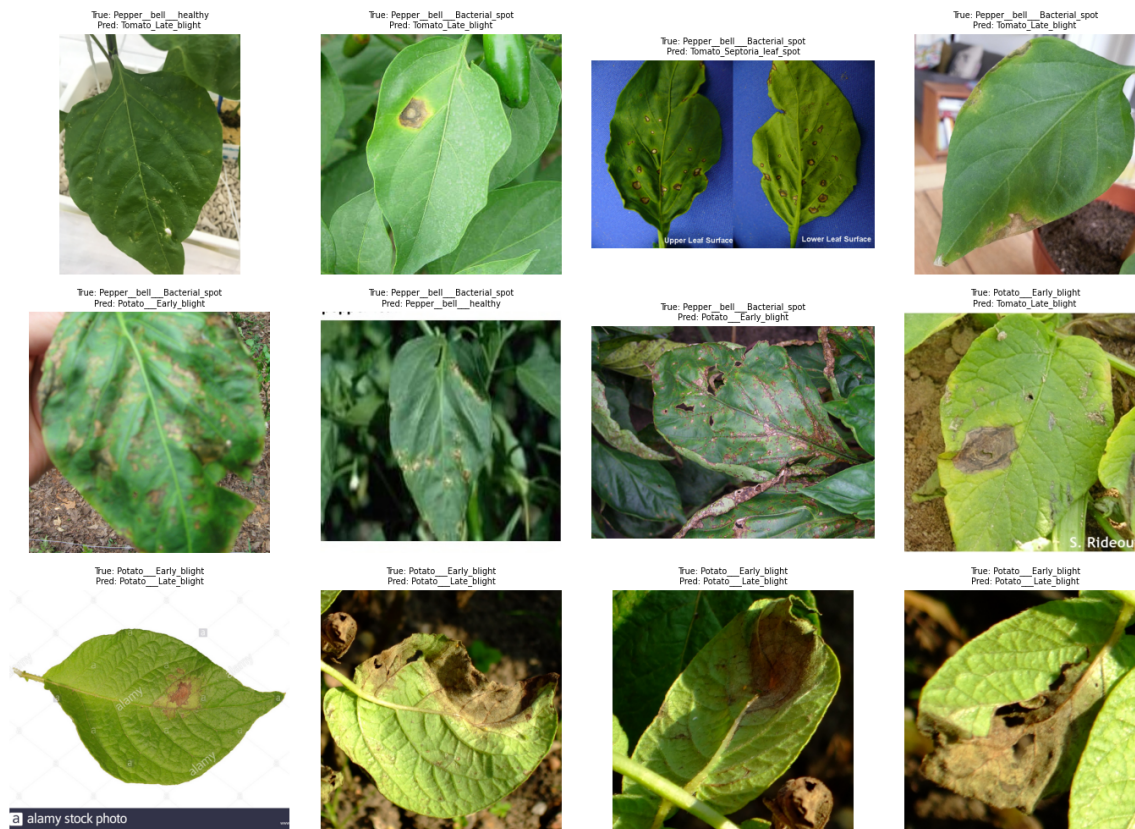


Figure 4-5: Representative misclassified PlantDoc test images for the best-performing fine-tuned model.

4.7 Discussion

Taken together, the results support three main conclusions. First, the domain shift between PlantVillage and PlantDoc is severe. The low zero-shot performance across all models and pipelines shows that high source-domain validation accuracy is not a reliable proxy for field robustness. Second, augmentation must be interpreted carefully. Moderate augmentation provides limited but sometimes useful gains, whereas stronger augmentation combined with mixed-image regularization does not improve zero-shot transfer in this study and often degrades it. Third, supervised domain adaptation through fine-tuning on PlantDoc is substantially more effective than augmentation alone for reducing the lab-to-field gap.

Adding a class-by-class perspective to our findings from section 4.6 provides additional insight into how the domain shift does not uniformly reduce performance over all classes. Classes with distinctive visual symptoms e.g., Pepper__bell___healthy ($F1 = 0.4667$) and Tomato_Early_blight ($F1 = 0.4242$) can still partially be recognized when no data from their respective domains was seen during training. However, the other classes where visual symptom recognition relies heavily upon subtle texture based visual cues i.e., Tomato_Bacterial_spot, Tomato__Tomato_mosaic_virus, and Tomato_healthy, all completely lost recognizable attributes ($F1 = 0.0000$). These differences suggest that while there may have been some sort of "generalization gap" observed, it would not necessarily have been a simple matter of scaling back the level of difficulty experienced during testing since this could be a function of how well visually distinguishable a disease's symptoms are under different imaging conditions. As such, reporting total system accuracy as a measure of cross domain readiness has significant limitations and therefore, in addition to total system accuracy, the authors recommend that future studies report the performance of individual classes as standard procedure in evaluating readiness for cross domain use in computer vision research applications for agriculture.

Pipelining P3 resulted in consistent degradations which, although can be explained by the fact that strong augmentation does not result in benefits every time, two other

mechanisms are also possible and are theoretically separable regardless of whether or not the study design allows for separation of these mechanisms. The first mechanism concerns the interaction between mixed-image regularization and fine-grained texture classification. MixUp and CutMix were originally validated on object recognition benchmarks such as ImageNet, where the discriminative signal is distributed broadly across an image. Plant disease classification is a fundamentally different problem: the diagnostic signal resides in localized lesion texture, color gradation at lesion margins, and symptom morphology that occupies a small and variable fraction of the leaf area. When CutMix replaces a rectangular region of one disease image with a patch from another, it systematically destroys exactly the localized features that the model must learn to recognize. Zhang et al. (2018) and Yun et al. (2019) acknowledge that CutMix assumes spatially distributed features an assumption that does not hold for fine-grained medical or biological image classification [26, 27]. Subsequent work on fine-grained recognition has shown that mixed-image regularization can degrade performance when class-discriminative regions are small and spatially concentrated [31, 32]. This directly predicts the pattern observed here: P3 hurts more for MobileNetV3-Large (which has fewer parameters and relies more heavily on discriminative texture) than for ResNet-50 (which has greater representational capacity to recover from corrupted samples). The second mechanism concerns the additive effect of environmental augmentations - RandomFog, RandomShadow, MotionBlur, and GaussNoise on top of the mixed-image regularization. Each of these individually simulates a plausible field condition. However, their simultaneous application generates a training distribution that is both more complex and qualitatively different from PlantDoc field images, which tend to exhibit one or two confounding factors at a time rather than all simultaneously. This is a form of over-regularization: the model must account for so many simultaneous perturbations that it cannot maintain stable feature representations for disease-relevant texture. The result is not improved robustness but rather a failure to converge on any coherent discriminative representation for the source domain, which then transfers even less well to the target. These two mechanisms together destruction of localized diagnostic features by CutMix, and

over-regularization from simultaneous environmental perturbations provide a principled explanation for why P3 underperforms P2 in zero-shot transfer despite its greater nominal diversity. The validity of this explanation is confirmed by the P4 ablation results. By removing the image-mixing component while keeping the environmental perturbations identical, the model regained a significant portion of its discriminative ability. This proves that while environmental noise is a hurdle, the systematic destruction of localized diagnostic features by CutMix is the critical 'breaking point' for cross-domain plant disease classification. Consequently, the performance gap between P4 and P3 (e.g., 2.42 percentage points for ResNet-50) serves as a direct measurement of the penalty imposed by mixed-image regularization in this specific task.

The environment coverage analysis in Table 3.4 provides additional context for interpreting the augmentation results. Although Pipeline P3 simulates six environmental conditions that are absent from P1 and P2, several common field conditions remain uncovered, including soil contamination, water droplets, and the presence of multiple overlapping leaves. These uncovered conditions are frequently present in PlantDoc images and may account for a portion of the residual domain gap that persists even after advanced augmentation. Moreover, as discussed in Section 4.3, the simultaneous application of all P3 transforms creates training images that combine multiple degradations at once, whereas real field images typically exhibit one or two conditions simultaneously. Future augmentation strategies could benefit from a more selective approach, applying individual environmental transforms with calibrated probabilities that match the empirical frequency of each condition in the target domain.

The additional comparison between P3-initialized adaptation and the best fine-tuning configuration further refines this conclusion. It shows that stronger source-domain augmentation does not necessarily produce a better starting point for downstream adaptation. In the present study, the best adaptation results were obtained from a simpler source-training setup, suggesting that excessive source-side regularization may harm the transferability of disease-relevant features when the target domain is already visually complex.

The above findings support the larger premise of this thesis as well. That is, the actual problem to be solved in plant disease image classification is not just building a large capacity classifier for curated images, it is creating training and testing methods that will function correctly when deployed. Therefore, the division into zero shot and fine tuning evaluation procedures has been justified by these experiments and they also clearly demonstrate the need for both. Zero shot evaluations measure the degree of domain transfer difficulty, whereas fine-tuned evaluation measures the amount of performance improvement that can occur from being able to use target domain examples. In general, the results demonstrate that in order to close the gap between laboratory and field based evaluations that explicit target domain adaptation is required. Architecture changes or augmentations on the source side are associated with less significant gains than those achieved by explicitly adapting to the target domain.

4.7.1 Answers to Research Questions

The experimental results presented in this chapter allow the four research questions posed in Chapter 1 to be addressed directly.

RQ1: How severe is the generalization gap when CNNs trained on controlled laboratory imagery are deployed directly on real-world field imagery?

There is a large difference in how well these models generalize to new data. For example, all of these models have validation accuracy over 98 percent for their original domain but the highest zero shot model had an accuracy of only 24.84 percent on the PlantDoc FULL model (ResNet-50, Pipeline P2) with macro-F1 values less than 0.20 for all different models shown in Table 4.1. Table 4.5, which provides per class results for each model also illustrates that the loss of recognition ability is not uniform and some classes are completely lost (Tomato_healthy, $F1 = 0.0000$), while other classes are still recognizable (Pepper__bell__healthy, $F1 = 0.4667$). This demonstrates that there is no guarantee that high benchmark performance will lead to being ready for deployment into a real world setting.

RQ2: To what extent can varying intensities of data augmentation

mitigate the domain shift?

Augmentation provides limited and non-monotonic improvement. Pipeline P2 (basic augmentation) yields small but consistent gains, peaking at a 2.21 percentage point increase for EfficientNet-B0. However, more advanced strategies proved counter-productive. Through the use of the P4 ablation study, it was discovered that while advanced environmental noise (haze, blur, shadows) limits performance, the inclusion of mixed-image regularization (MixUp/CutMix) causes a total collapse of certain class-level decision boundaries. These results demonstrate that augmentation alone cannot solve the controlled-to-field generalization problem, and that excessive structural augmentation is specifically detrimental when it obscures the fine-grained features required for disease identification.

RQ3: How does supervised domain adaptation compare to data augmentation in bridging the lab-to-field gap?

Supervised domain adaptation through fine-tuning is substantially more effective than any augmentation strategy including the isolated environmental transforms tested in Pipeline P4. The best fine-tuned configuration (MobileNetV3-Large, P2 initialization) achieves 59.32% accuracy and 0.5734 macro-F1 on the PlantDoc test split, representing an improvement of 37.83 percentage points in accuracy over its zero-shot baseline (Table 4.4). By contrast, the largest gain from changing augmentation pipeline alone is only 2.21 percentage points. Furthermore, the per-class analysis shows that fine-tuning does not merely improve aggregate metrics but restructures the error profile from domain-induced confusion to biologically meaningful confusion between visually similar diseases. This confirms that exposure to labeled target-domain data is the most effective intervention tested in this study for reducing the deployment gap.

RQ4: How data-efficient is supervised domain adaptation? Specifically, what fraction of labeled target-domain field data is required to effectively bridge the lab-to-field generalization gap?

Supervised domain adaptation is highly data-efficient. The learning curves in Figure 4-4 show that performance improves sublinearly as more target-domain data is

introduced. Even at the 10% fraction (approximately 102 images), all three architectures surpass their zero-shot baselines by a substantial margin: ResNet-50 reaches 37.29% accuracy (vs. 24.84% zero-shot) and EfficientNet-B0 reaches 36.72% (vs. 22.99%). Diminishing returns become evident beyond 50% of the training data (514 images), at which point ResNet-50 already recovers 51.13% accuracy close to its 55.93% ceiling at 100% training data. These findings indicate that bridging the lab-to-field gap does not require massive field-collected datasets; an annotation budget of roughly 300–500 target-domain images is sufficient to recover most of the fine-tuning benefit.

4.8 Threats to Validity and Limitations

Several limitations should be acknowledged. First, the experiments are restricted to 12 shared classes defined through manual label harmonization between PlantVillage and PlantDoc. Although this step is necessary for valid cross-dataset evaluation, it introduces a degree of human judgment and excludes non-overlapping disease categories. The excluded classes, which include all apple, corn, grape, and strawberry categories, may exhibit different domain shift characteristics than the retained pepper, potato, and tomato classes. Consequently, the findings of this study should be generalized to other crop-disease combinations with caution.

Second, the zero-shot and fine-tuning protocols use different target-domain set definitions. Zero-shot evaluation uses the PlantDoc FULL set of 1405 images, while fine-tuning evaluation uses the 118-image PlantDoc test split. While this is methodologically justified the zero-shot model has never seen any PlantDoc data, whereas the fine-tuned model has been trained on the PlantDoc training partition it means that their absolute scores should not be directly compared without noting the dataset difference. The per-class comparison in Table 4.5 addresses this by evaluating both models on the same 118-image test split.

Third, the small size of the target-domain data presents a limitation at both training and evaluation stages. During fine-tuning, the filtered PlantDoc training

split contains approximately 1,287 images across 12 classes, which is small relative to the capacity of the evaluated architectures. This contributes to the overfitting observed during adaptation, where training accuracy on PlantDoc reaches near-ceiling levels within 10 epochs while validation performance plateaus around 60%. Early stopping was critical in mitigating this effect, though it also limits how much the model can adapt before training is halted. During evaluation, the 118-image test split limits the statistical precision of the reported metrics, with per-class support ranging from 6 to 15 samples. At this scale, a single additional correct or incorrect prediction can shift a class-level F1 score by 0.10 or more. To quantify the uncertainty at the aggregate level: with 118 samples and an observed accuracy of 59.32%, a 95% binomial confidence interval spans approximately ± 9 percentage points. Future work should evaluate on larger field-condition datasets to obtain more precise performance estimates. To partially mitigate these small-sample effects, several measures were adopted throughout the study: class-weighted cross-entropy loss was used during both source training and fine-tuning, stratified splits preserved class proportions across partitions, and macro-averaged metrics were used to weight all classes equally regardless of support.

Fourth, while the strongest augmentation pipeline (P3) combines multiple interventions, this study successfully isolated the specific impact of mixed-image regularization through the Pipeline P4 ablation. However, the individual contributions of the remaining environmental perturbations such as the relative impact of shadow versus fog remain bundled. Future work could further decompose Pipeline P4 to determine which specific environmental factors impose the greatest accuracy penalty. Additionally, the study evaluates only three pretrained CNN backbones; exploring more recent architectures (e.g., Vision Transformers) or more advanced domain adaptation techniques like feature alignment or adversarial training may yield different results. Finally, the study does not include non-deep learning baselines, which could provide additional context for interpreting the observed performance gains.

Despite these limitations, the directional conclusions of the study that the domain gap is severe, that moderate augmentation outperforms advanced augmentation, and

that fine-tuning provides substantially larger gains than augmentation alone are supported consistently across all three architectures and are therefore unlikely to be artifacts of the small sample size.

Chapter 5

Conclusion

This research looked at how plants are classified with respect to their diseases when classifying them using images taken in a lab versus images of plants that were actually grown outdoors. The study was based on two domains (PlantVillage being the source domain; and PlantDoc the target) where it used 3 different types of pre-trained CNNs (Resnet50, Efficientnetb0, Mobilenetv3-large), all working under a consistent 12-class cross-domain setup. In contrast to many plant disease studies that report only strong within-dataset performance, this work explicitly examined whether models trained on clean source-domain data could generalize to visually different field images and whether this gap could be reduced through augmentation and target-domain fine-tuning.

The experimental evidence clearly illustrates that performing a "zero shot" cross domain adaptation experiment from PlantVillage to PlantDoc is quite challenging. Although the model was capable of achieving high source domain validation performance, it did poorly when applied to PlantDoc. This indicates that there is no relationship between achieving high performance on a dataset that has been carefully curated, and being able to perform well under the conditions of the real world. As such, the findings support one of the main hypotheses of this dissertation; namely, that while accurate classification in a controlled environment is an important factor, the major obstacle to using machine learning models to recognize plant diseases is the ability to effectively deal with domain shift.

The same types of additional experiments were performed to assess whether augmenting data would help improve the performance of the models across different domains. These experiments also revealed that higher levels of data augmentation do not always provide better results. Instead, augmentation at moderate levels (pipeline p2) produced the greatest improvement in zero-shot test data performance for all of the models evaluated. The largest amount of augmentation (Pipeline P3) resulted in decreased performance in nearly every case. These results indicate that augmentation must be carefully selected to match the desired variability. If augmentation becomes too extreme, or too artificial, then instead of increasing the invariance of disease-related features, it may actually obscure these features. This was empirically confirmed by the Pipeline P4 ablation study, which demonstrated that removing structural image-mixing while maintaining identical environmental noise consistently improved accuracy. This suggests that for plant pathology tasks, the systematic destruction of localized lesion textures by techniques like CutMix represents a critical failure point for cross-domain generalization.

The most important discovery of this dissertation was that supervised domain adaptation via fine-tuning on PlantDoc yields significantly improved performance on the target domain. Each of the three models improved with target-domain training, with MobileNetV3-Large yielding the best fine-tuned results. In addition, the improvement from domain adaptation was substantially greater than the improvement from source-domain augmentation. Furthermore, the data-fraction ablation study revealed that this adaptation is highly data-efficient. Performance scaled logarithmically with the amount of target-domain training data, meaning that the majority of the fine-tuning benefit can be recovered with a small annotation budget of roughly 300 to 500 field-collected images. This finding has direct practical implications: it suggests that collecting and labeling even a modest number of field images is sufficient to substantially close the domain gap, making deployment in new agricultural environments more feasible than a full dataset collection effort would suggest.

The overall results show that while transfer learning offers a strong foundation, relying solely on source-domain training is not sufficient for reliable deployment in

the field. For that reason, explicit label matching, careful evaluation with a common class space, and isolating zero-shot transfer from fine-tuning on the target domain were all required to obtain interpretable results. As such, this dissertation provides a more practical view of how well plant disease classification systems perform when deployed in real-world environments, and shows that cross-domain evaluation should be viewed as a necessary aspect of agricultural computer vision research rather than an optional add-on to within-dataset benchmarking.

There are many paths for additional research. Advanced methods of domain adaptation, including feature alignment, self-supervised representation learning, and semi-supervised learning with unlabeled target domain data, could potentially help further decrease the gap between the lab and the field. Future ablation studies will also have to isolate the effects of each individual augmentation component. While this thesis successfully isolated the impact of mixed-image regularization through Pipeline P4, the most effective augmentation pipeline examined in this thesis combined multiple environmental augmentations that still need individual decomposition. Additionally, future augmentation pipelines should consider a broader range of environmental conditions beyond those simulated in Pipeline P3, including soil and dust contamination on leaf surfaces, water droplets and dew, and the co-occurrence of insect damage with disease symptoms. Augmentation parameters could also be calibrated to match the empirical distribution of environmental conditions in specific agricultural regions, rather than applying all transforms uniformly, since the results of this study suggest that augmentation is most effective when it closely matches the variability of the target domain. The broader evaluation of additional field datasets, shared classes, and larger target-domain test sets will also increase the reliability and generality of the conclusions. Future research may study the use of lightweight models suitable for deployment along with adaptation techniques in a way that maximizes both robustness and accuracy while minimizing computational cost in agricultural settings.

In summary, this dissertation indicates that the gap between controlled and field environments for plant disease classification is large, augmentation strategies tested in this study specifically when using aggressive structural mixing—can be counter-

productive, and target-domain fine-tuning is the best of the three methods tested for improving performance in real-world settings. This study’s conclusions also support the necessity of evaluating plant disease models in a cross-domain setting that simulates real-world use cases. The study further supports the need for a stronger baseline for future work on robust agricultural image classification.

Appendix A

Figures

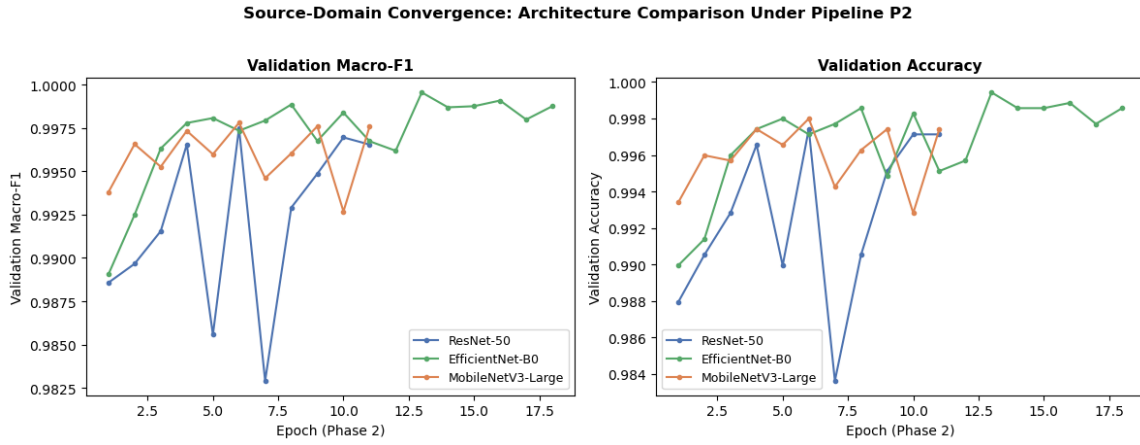


Figure A-1: Validation loss and macro-F1 convergence curves for all three architectures trained under Pipeline P2, illustrating the two-phase training strategy and early stopping behavior.

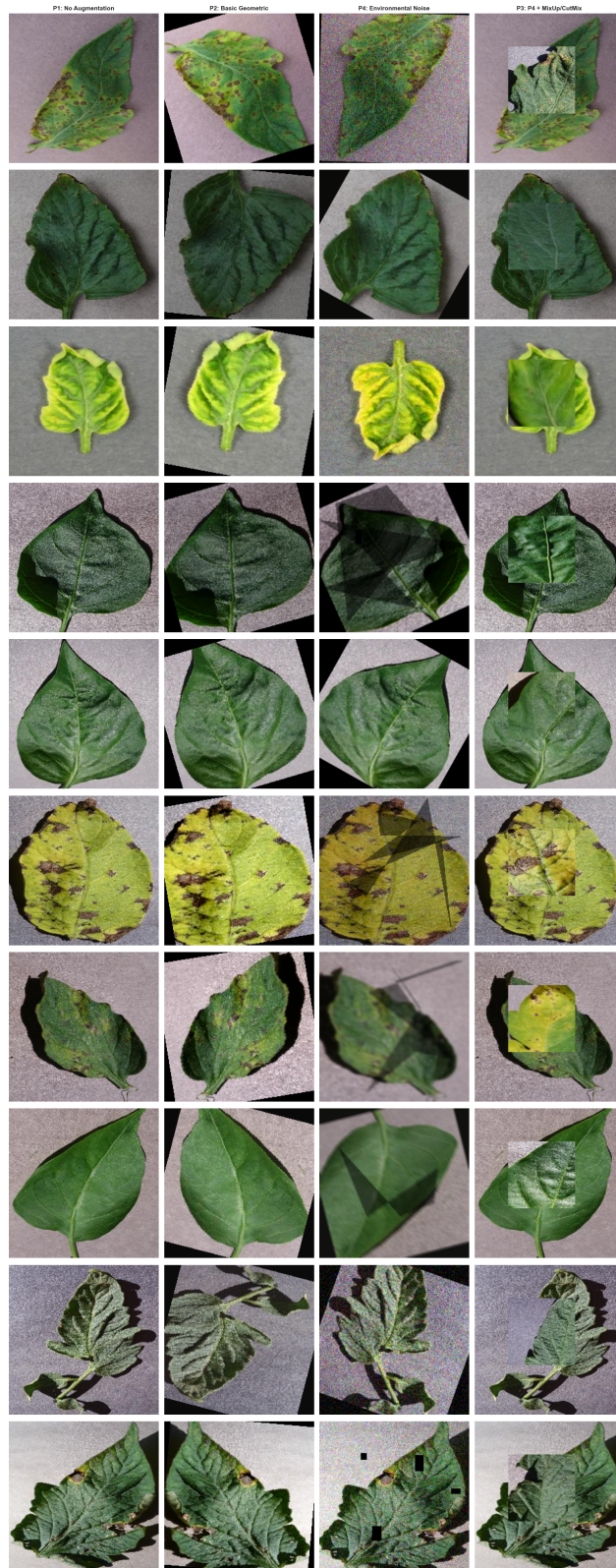


Figure A-2: Additional PlantVillage samples transformed by each training pipeline (10 random samples).

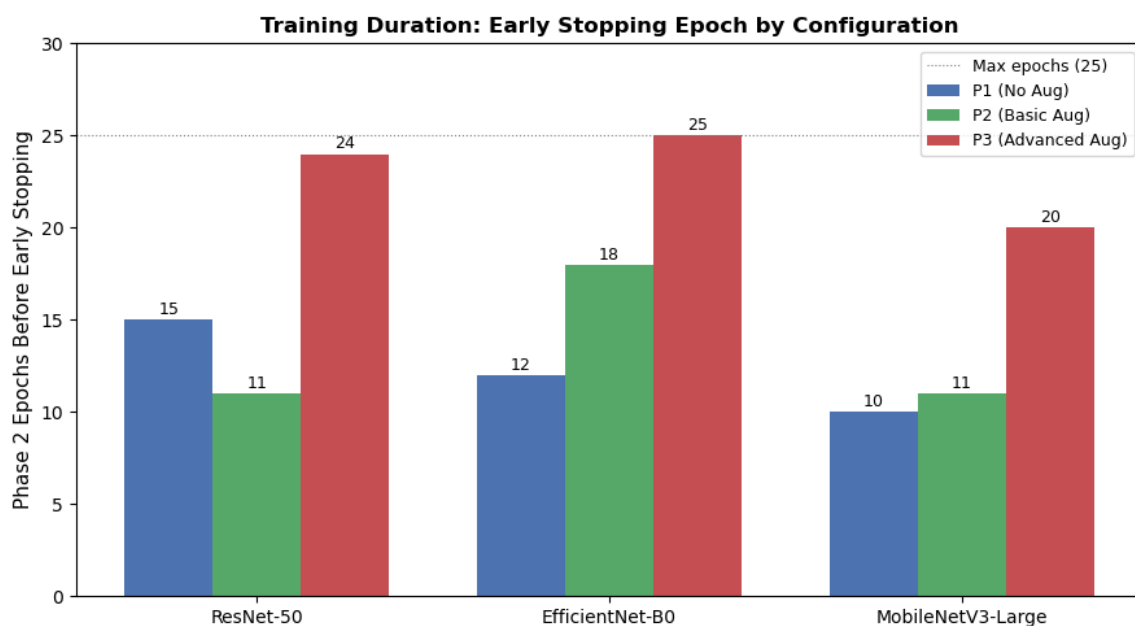


Figure A-3: Early stopping epoch for each model-pipeline combination, trained on PlantVillage source domain.

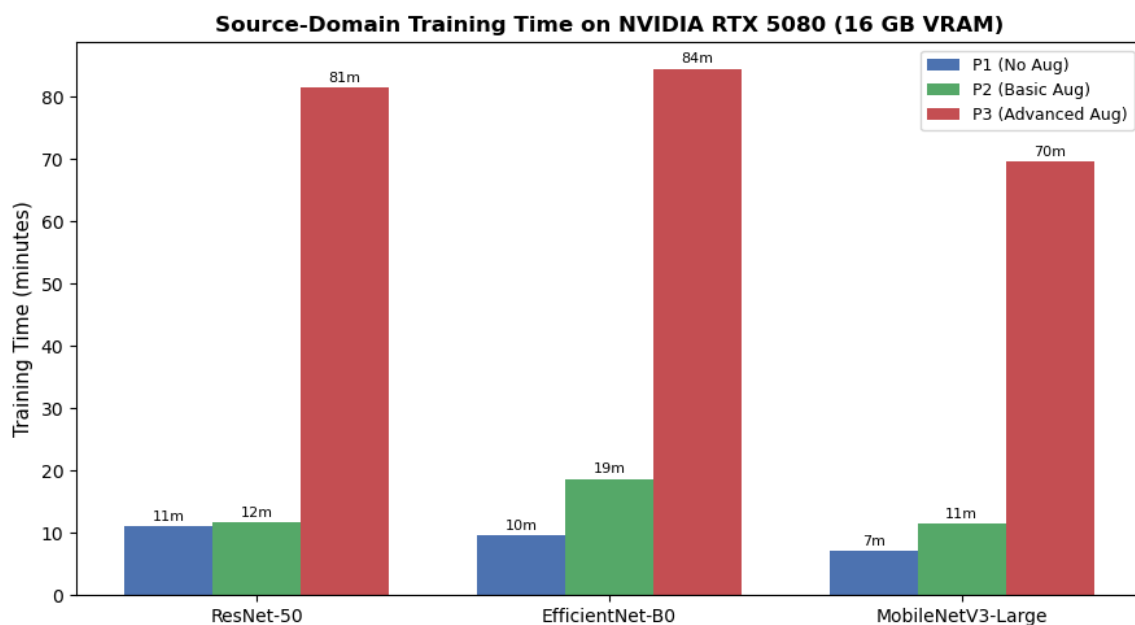


Figure A-4: Training time and early stopping epoch for each model-pipeline combination, trained on PlantVillage source domain.

Bibliography

- [1] Y. Kaya and E. Gürsoy, “A review of deep learning architectures for plant disease detection,” *Turkish Journal of Biology*, vol. 49, no. 5, pp. 459–497, 2025.
- [2] K. Antwi, K. E. Bennin, D. K. P. Asiedu, and B. Tekinerdogan, “On the application of image augmentation for plant disease detection: A systematic literature review,” *Smart Agricultural Technology*, vol. 9, p. 100590, Dec. 2024.
- [3] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, and N. Batra, “PlantDoc: A dataset for visual plant disease detection,” in *Proc. 7th ACM IKDD Conf. on Data Science (CoDS) and 25th COMAD*, Hyderabad, India, 2020, pp. 249–253.
- [4] A. Abbas, S. Jain, M. Gour, and S. Vankudothu, “Tomato plant disease detection using transfer learning with cGAN synthetic images,” *Computers and Electronics in Agriculture*, vol. 187, p. 106279, 2021.
- [5] Ü. Atila, M. Uçar, K. Akyol, and E. Uçar, “Plant leaf disease classification using EfficientNet deep learning model,” *Ecological Informatics*, vol. 61, Art. 101182, 2021.
- [6] M. S. Krishna, P. Machado, R. I. Otuka, S. W. Yahaya, F. N. dos Santos, and I. K. Ihianle, “Plant leaf disease detection using deep learning: a multi-dataset approach,” *J — Multidisciplinary Scientific Journal (MDPI Open Access)*, vol. 8, no. 1, Art. 4, pp. 1–24, Jan. 2025.

- [7] A. Y. Ashurov et al., “Enhancing plant disease detection through deep learning: a depthwise CNN with squeeze-and-excitation integration and residual skip connections,” *Frontiers in Plant Science*, vol. 15, Art. 1505857, 2024.
- [8] A. González-Briones, S. López Florez, P. Chamoso, L. F. Castillo-Ossa, and E. S. Corchado, “Enhancing plant disease detection: incorporating advanced CNN architectures for better accuracy and interpretability,” *International Journal of Computational Intelligence Systems*, vol. 18, Art. 120, 2025.
- [9] M. Jung et al., “Construction of deep learning-based disease detection model in plants,” *Scientific Reports*, vol. 13, Art. 7331, 2023.
- [10] S. M. Hassan, A. K. Maji, M. Jasinski, Z. Leonowicz, and E. Jasinska, “Identification of plant-leaf diseases using CNN and transfer-learning approach,” *Electronics*, vol. 10, no. 12, Art. 1388, 2021.
- [11] M. J. Karim et al., “Enhancing agriculture through real-time grape leaf disease classification via an edge device with a lightweight CNN architecture and Grad-CAM,” *Scientific Reports*, vol. 14, Art. 16022, 2024.
- [12] F. Adnan, M. J. Awan, A. Mahmoud, H. Nobanee, A. Yasin, and M. Altameem, “EfficientNetB3-adaptive augmented deep learning (AADL) for multi-class plant disease classification,” *IEEE Access*, vol. 11, pp. 85426–85440, 2023.
- [13] S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” *Frontiers in Plant Science*, vol. 7, Art. 1419, 2016.
- [14] K. P. Ferentinos, “Deep learning models for plant disease detection and diagnosis,” *Computers and Electronics in Agriculture*, vol. 145, pp. 311–318, 2018.
- [15] J. G. A. Barbedo, “Plant disease identification from individual lesions and spots using deep learning,” *Biosystems Engineering*, vol. 180, pp. 96–107, 2019.

- [16] L.-H. Li and R. Tanone, “Improving robustness using MixUp and CutMix augmentation for corn leaf diseases classification based on ConvMixer architecture,” *Journal of ICT Research and Applications*, vol. 17, no. 2, pp. 167–180, 2023. doi: 10.5614/itbj.ict.res.appl.2023.17.2.3.
- [17] X. Jiang et al., “PlantCaFo: An efficient few-shot plant disease recognition method based on foundation models,” *Computers and Electronics in Agriculture*, 2025. Article ID S2643651525000305.
- [18] J. Chen, D. Zhang, A. Zeb, and Y. A. Nanekaran, “Identification of rice plant diseases using lightweight attention networks,” *Expert Systems with Applications*, vol. 169, p. 114514, 2021. doi: 10.1016/j.eswa.2020.114514.
- [19] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [21] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.
- [22] A. Howard et al., “Searching for MobileNetV3,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1314–1324.
- [23] J. Bergstra and Y. Bengio, “Random Search for Hyper-Parameter Optimization,” *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
- [24] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 60, 2019.

- [25] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random Erasing Data Augmentation,” in *Proc. AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 13001–13008, 2020.
- [26] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” in *Proc. International Conference on Learning Representations (ICLR)*, 2018.
- [27] S. Yun, D. Han, S. Oh, S. Chun, J. Choe, and Y. Yoo, “CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.
- [28] M. Wang and W. Deng, “Deep Visual Domain Adaptation: A Survey,” *Neurocomputing*, vol. 312, pp. 135–153, 2018.
- [29] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, “Domain Generalization: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2023.
- [30] M. A. Al Mamun, M. I. Sarkar, M. Z. Alam, and S. R. I. Karim, “Evaluating The Efficacy Of Hybrid Deep Learning Models In Rice Variety Classification,” *International Journal of Advanced Research in Intelligent Systems Engineering (IJARISE)*, 2024.
- [31] J.-H. Kim, W. Choo, and H. O. Song, “Puzzle Mix: Exploiting Saliency and Local Statistics for Optimal Mixup,” in *Proc. 37th Int. Conf. Machine Learning (ICML)*, vol. 119, pp. 5275–5285, 2020.
- [32] S. Huang, X. Wang, and D. Tao, “SnapMix: Semantically Proportional Mixing for Augmenting Fine-Grained Data,” in *Proc. 35th AAAI Conf. on Artificial Intelligence*, pp. 1628–1636, 2021.
- [33] M. T. Ahad, Y. Li, B. Song, and T. Bhuiyan, “Comparison of CNN-based Deep Learning Architectures for Rice Diseases Classification,” *Artificial Intelligence in Agriculture*, vol. 9, pp. 22–35, 2023. doi: 10.1016/j.aiia.2023.07.001.