IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 July 2025, accepted 30 July 2025, date of publication 4 August 2025, date of current version 14 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3595390

RESEARCH ARTICLE

Deployable Deep Learning for Cross-Domain Plant Leaf Disease Detection via Ensemble Learning, Knowledge Distillation, and Quantization

MOHAMMAD JUNAYED HASAN[1,2], SUVODEEP MAZUMDAR[3], AND SIFAT MOMEN[4]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
[2]Department of Computational Pathology and AI-Allied Health, Mayo Clinic, Rochester, MN 55905, USA
[3]School of Information, Journalism and Communication, University of Sheffield, S10 2AH Sheffield, U.K.
[4]Department of Electrical and Computer Engineering, North South University, Dhaka 1229, Bangladesh

Corresponding authors: Suvodeep Mazumdar (s.mazumdar@sheffield.ac.uk) and Sifat Momen (sifat.momen@northsouth.edu)

ABSTRACT Accurate leaf disease detection via smartphone-based deep learning holds immense potential for mitigating global crop losses. However, significant deployment challenges persist when transitioning from controlled laboratory environments to real-world agricultural conditions. Despite recent advances, three fundamental barriers remain: cross-domain generalization, severe class imbalance, and computational limitations for edge deployment. This study introduces the first open cross-domain benchmark for tomato leaf disease detection, unifying PlantVillage and TomatoVillage datasets into 15 harmonized disease classes to enable reproducible evaluation across domains. We propose a unified optimization approach integrating ensemble learning, knowledge distillation, and quantization across 24 deep learning architectures for edge-compatible disease detection. Strategic data augmentation and ADASYN-based balancing mitigate the severe 75:1 class imbalance, while systematic hyperparameter tuning optimizes model configurations. Our four-model ensemble (DenseNet-121, ResNet-101, DenseNet-201, EfficientNet-B4) achieves 99.15% accuracy via soft-voting. Knowledge distillation transfers ensemble capabilities to compact ShuffleNetV2, maintaining 98.53% accuracy with 163× parameter reduction and 43.6× speedup. INT8 quantization provides 671× compression (1.46 MB) while sustaining 97.46% accuracy, enabling 0.29ms inference. Cross-dataset validation demonstrates robust generalization with only 3.45% performance degradation. Grad-CAM++ and LIME-based explainability confirm biologically grounded attention patterns aligned with plant pathology principles. Finally, field deployment via a multilingual and multi-platform Flutter application validates real-world feasibility, establishing the first scalable framework bridging research and practical agricultural deployment. This work sets a standardized benchmark and extensible methodology for future multi-dataset precision agriculture research. Codes and implementations are publicly available at: https://github.com/junayed-hasan/tomato-leaf-ai

INDEX TERMS Cross-domain generalization, data imbalance, deep learning, deployment, ensemble learning, explainable AI, knowledge distillation, quantization, leaf disease detection, tomato leaf disease.

I. INTRODUCTION

Ensuring global access to nutritious food remains an urgent challenge as the world population is projected to surpass 9.7 billion by 2050. Agricultural yield losses are estimated to affect up to 40% of potential production, and are increasingly attributed to plant–pathogen outbreaks, which constitute a substantial impediment to food security efforts [1]. Among vegetable crops, tomato (*Solanum lycopersicum*) ranks as the most widely consumed globally, contributing approximately

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

16% of total vegetable output and exceeding 180 million metric tons annually [2]. Timely and accurate diagnosis of foliar diseases is therefore foundational to stabilizing tomato supply chains and minimizing crop attrition. The seminal study by Mohanty et al. revealed that convolutional neural networks (CNNs), when trained on the *PlantVillage* dataset, can outperform human experts in leaf-disease identification tasks [3]. Research has since diversified across architectures and learning paradigms. State-of-the-art detectors ranging from ensemble feature fusion [4] and attention-enhanced CSWinTransformer architectures [5] to Modified-Xception networks [6], hybrid CNN–Transformer stacks [7], and Bayesian-optimised ensembles [2] demonstrate exceptional accuracy within the bounds of curated datasets. Lightweight models such as XLTLDisNet [8], MobileNet–SVM hybrids [9], and adaptive exponential-average ensembles [10] signal growing interest in edge compatibility.

However, translating these high accuracies from controlled settings to robust real-world deployment poses several critical technical challenges. The foremost challenge is the domain gap between curated research datasets and heterogeneous field conditions, where models trained on datasets like PlantVillage [3], consisting of uniformly posed leaf images on simple backgrounds, exhibit significantly reduced performance when exposed to field imagery featuring complex backgrounds, varied lighting, and natural occlusion [11]. Recent efforts address this limitation by integrating diverse data sources: Singh et al. released PlantDoc [12], Moupojou et al. introduced FieldPlant [13], and Gehlot et al. released TomatoVillage [14]. However, these datasets contain only partially overlapping disease classes, necessitating a unified approach combining multiple datasets for comprehensive disease detection. Compounding this challenge is severe class imbalance in leaf-disease datasets, where dominant diseases may comprise 75% of samples while rare but devastating diseases have scarce examples, biasing standard training toward majority classes. While recent research has explored data augmentation [14], [15], [16] and Generative Adversarial Networks [17], [18], [19], [20] for tomato disease classification, advanced sampling techniques such as SMOTE [21] and ADASYN [22] remain unexplored, requiring systematic evaluation alongside loss function modifications such as focal loss [23] to ensure robust learning across all disease classes. A further critical consideration for real-world deployment is the resource-constrained nature of edge environments, where state-of-the-art deep neural networks with millions of parameters are incompatible with smartphones commonly available to farmers. These devices have limited processing power, memory, and intermittent connectivity, necessitating models that achieve both compactness and efficiency through techniques such as knowledge distillation [24], [25], [26], [27], [28], [29], pruning [24], [28], [30], quantization [24], [31], and efficient CNN architectures [8], [32]. However, existing studies often apply these compression techniques in isolation without systematic optimization of the complete training pipeline, including data augmentation strategies, hyperparameter tuning, and ensemble methods, while failing to address performance degradation when combining multiple compression techniques or deploying on heterogeneous datasets.

To address these limitations, this paper presents a systematic pipeline that integrates ensemble learning, knowledge distillation, and quantization for deployable cross-domain tomato leaf disease detection. Our approach combines the largest laboratory-focused dataset, PlantVillage, and the largest real-world field focused dataset, TomatoVillage, into 15 unified disease classes, enabling comprehensive evaluation across laboratory-controlled and field imaging conditions. We systematically optimize data augmentation, ADASYN-based class balancing to address severe imbalance ratios, and hyperparameter tuning across 24 state-of-the-art architectures. The optimal ensemble achieves state-of-the-art performance through soft voting aggregation, with knowledge distillation successfully transferring ensemble capabilities to compact ShuffleNetV2 architecture and INT8 quantization further compressing the model to 1.46 MB with minimal degradation, enabling sub-35ms mobile inference. Cross-dataset evaluation demonstrates robust generalization, while explainable AI analysis through Grad-CAM++ and LIME validates biologically relevant attention mechanisms, establishing a robust framework that bridges laboratory research and real-world agricultural deployment requirements.

In summary, the main contributions of this work are:

- **Unified Dataset and Benchmark Establishment:** We aggregate two complementary tomato disease image datasets into 15 common classes and establish the first comprehensive benchmark on this unified taxonomy, providing a foundation for future multi-dataset research. We apply extensive augmentation and balancing techniques to address domain discrepancy and class imbalance at the data level.
- **Evaluation of Diverse Models and Ensemble:** We evaluate 24 deep neural architectures (CNNs and vision transformers) and propose an ensemble model that achieves state-of-the-art accuracy on tomato leaf disease recognition, outperforming individual networks.
- **Distilled-Quantized Multilingual Deployed Model:** We distill the ensemble's knowledge into a compact student network and apply INT8 quantization to produce an ultra-lightweight model, reducing model size by 671× with 43.6× inference speedup while maintaining near-ensemble accuracy. The final model is deployed on mobile devices with 0.29ms inference and support for six languages, enabling seamless accessibility for diverse agricultural stakeholders.
- **Cross-Dataset Generalization and Explainability:** We validate the developed model's generalization by testing it on unseen lab and field data, demonstrating

improved robustness to domain shift. We also provide explainable AI outputs (Grad-CAM++ heatmaps and LIME decision-boundaries) to interpret model decisions, enhancing the model's transparency and user trust.

The remainder of this paper is organized as follows. Section II reviews the relevant work on plant disease detection based on four sub fields. Section III details the proposed framework, including dataset preparation, model architectures, and the integrated training pipeline. Section IV presents the experimental details, results, and analysis, covering ablation studies, cross-dataset evaluations, and comparison with existing methods. Section V discusses deployment considerations and summarizes the main findings, implications, and limitations of the work. Finally, Section VI concludes the paper with a summary of findings and suggestions for future work.

II. RELATED WORK

A. ARCHITECTURAL ADVANCEMENTS

Deep learning for plant disease detection has evolved significantly since Mohanty et al. [3] demonstrated GoogLeNet-based transfer learning achieving 99.35% accuracy on 54,306 PlantVillage images spanning 14 crop species. Contemporary advances encompass hybrid attention mechanisms: Sun et al. [5] developed E-TomatoDet with CSWinTransformer backbone achieving 97.2% mAP50, and Ghosh et al. [7] combined VGG19, ViT, EfficientNetV2, ConvNeXt, and Swin Transformer achieving 98% accuracy. Advanced ensemble methodologies include Sharma et al. [4] feature-concatenation ensembles achieving 99.91% accuracy and A.M. et al. [10] adaptive ensembles with exponential moving average fusion achieving 98.7% accuracy. Specialized innovations encompass Khan et al. [2] Bayesian-optimized CNN-stacking achieving 98.27% accuracy, Thangaraj et al. [6] Modified-Xception achieving 99.61% accuracy, and Chelladurai et al. [33] transductive LSTM achieving 99.98% accuracy with 20ms inference. Lightweight architectures include Das et al. [8] XLTLDisNet achieving 97.24% accuracy, Vini et al. [34] TrioConvTomatoNet achieving 99.4% accuracy, and Imam et al. [9] MobileNet-SVM achieving 99.37% accuracy. However, most studies focus on isolated architectural improvements without systematic pipeline integration.

B. COMPRESSION TECHNIQUES

Knowledge distillation has emerged as crucial for agricultural edge deployment, with Liu et al. [25] pioneering multi-task distillation achieving 10.6× parameter reduction while maintaining ResNet101-comparable accuracy, and Ni et al. [27] introducing ensemble self-distillation with ShuffleNetV2 achieving 95.08% accuracy with 37.77% parameter reduction. Advanced compression methodologies demonstrate exceptional efficiency: Wang et al. [24] employed structured pruning and INT8 quantization achieving remarkable 1,196× compression for VGGNet (0.04 MB) with 97.09% accuracy, while Xu et al. [30] developed CNNA architectures with 75× parameter reduction achieving 98.96% accuracy with 47.35ms inference on Jetson TX2 NX. Ultra-lightweight approaches include Rakib et al. [32] INT8-quantized CNNs deployable on ESP32-CAM using 28KB models achieving 98% accuracy, and Jian et al. [28] DGP-SNNet achieving 63.4% FLOPs reduction with +2.23% accuracy improvement. However, most studies apply compression techniques in isolation without systematic pipeline integration.

C. DATA ENHANCEMENT

Strategic data enhancement methodologies have emerged as fundamental components for robust agricultural disease detection. Contemporary strategies demonstrate substantial improvements: Ahmed et al. [15] developed runtime-augmented MobileNetV2 with CLAHE preprocessing achieving 99.30% accuracy using 9.60 MB models with 4.87M FLOPs, while Karande and Garg [16] conducted systematic analysis with GrabCut segmentation achieving comparable MobileNet performance with 3.9 MB models. Advanced generative approaches include Deshpande and Patidar [17] GAN-enhanced parallel DCNN achieving 99.14–99.74% accuracy, Chen et al. [20] hybrid CNN-Transformer with cycle-consistent GAN achieving 99.45% accuracy, and Ojo and Zahid [18] CLAHE with GAN-based resampling achieving 97.69% accuracy. Multi-modal strategies include Xiang et al. [31] DWTFormer fusing discrete wavelet frequency and spatial features achieving 99.28% accuracy with 0.028s inference time. However, systematic multi-technique optimization and comprehensive class balancing with advanced sampling methods remain significant research gaps.

D. EXPLAINABLE AI

Explainable AI (XAI) has become increasingly important for agricultural AI systems to ensure user trust and facilitate expert validation. Recent works demonstrate diverse approaches: Sun et al. [26] investigating Data-efficient Image Transformers (DeiT) provide attention-based interpretability, Ghosh et al. [7] demonstrate explainable hybrid architectures with visualization capabilities, Das et al. [8] emphasize transparency and interpretability in lightweight approaches, and Chelladurai et al. [33] incorporate attention mechanisms for explainable temporal modeling. Current limitations include limited validation against expert knowledge and lack of agricultural-specific interpretability metrics that account for domain expertise.

E. DIAGNOSTIC AND DETECTION APPROACHES

Beyond agricultural applications, recent literature highlights broader advances in diagnostic AI systems across industrial and medical domains. Su and Lee [35] provide a comprehensive review of machine learning approaches for diagnostics and prognostics using open-source PHM challenge data, emphasizing system-level integration and benchmarking strategies. Aftab et al. [36] explore deep learning applications

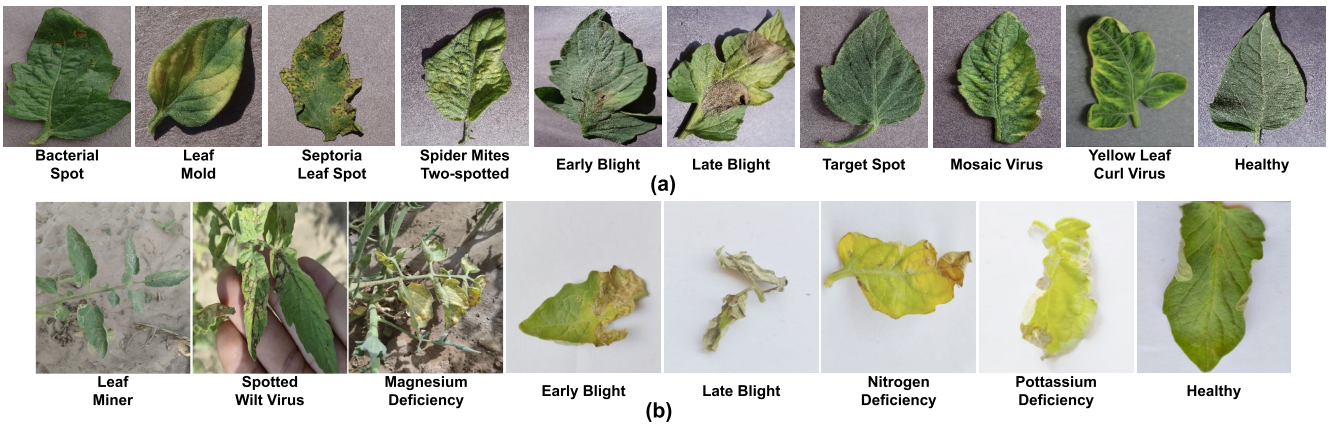

FIGURE 1. Representative samples from both datasets showcasing class diversity and domain characteristics. (a) PlantVillage samples demonstrate controlled laboratory conditions with uniform backgrounds and standardized imaging. (b) TomatoVillage samples exhibit realistic field conditions with natural lighting and authentic disease manifestations in individual separated leaves.

in oncology, demonstrating how multimodal AI models enhance cancer detection across imaging and genomic modalities. Alkhanbouli et al. [37] conduct a systematic review of explainable AI in disease prediction, identifying key interpretability challenges and proposing future directions for clinical integration. Cabral et al. [38] present a global survey on the projected impact of AI in diagnostic medicine, highlighting expected improvements in accuracy, cost-efficiency, and decision support. Bianco et al. [39] summarize recent advances in brain tumor diagnosis and treatment across pediatric and adult populations, emphasizing molecular profiling, imaging innovations, and targeted therapies. These studies reinforce the importance of transparency, scalability, and domain-specific adaptation in diagnostic AI systems, aligning with our pipeline-driven approach for agricultural disease detection.

III. MATERIALS AND METHODS

A. DATASETS

1) PlantVillage DATASET

The PlantVillage dataset [3] constitutes the largest publicly available laboratory-controlled dataset for plant disease classification. The tomato subset encompasses 18,160 high-resolution images distributed across 10 distinct disease classes: *Tomato Yellow Leaf Curl Virus*, *Bacterial spot*, *Late blight*, *Septoria leaf spot*, *Spider mites Two-spotted*, *Healthy*, *Target Spot*, *Early blight*, *Leaf Mold*, and *Tomato mosaic virus*. Images were captured under controlled laboratory conditions with standardized lighting, uniform backgrounds, and optimal leaf positioning to minimize environmental variability. The class distribution exhibits significant imbalance, ranging from 373 images for Tomato mosaic virus to 5,357 images for Tomato Yellow Leaf Curl Virus, with the latter representing 29.5% of the dataset. Each image maintains a consistent resolution of 256×256 pixels and exhibits minimal background clutter, facilitating feature extraction focused exclusively on foliar pathological manifestations.

2) TomatoVillage DATASET

The TomatoVillage dataset [14] represents the largest publicly available field-condition dataset specifically designed to bridge the domain gap between laboratory and real-world agricultural scenarios. This dataset comprises 4,525 images spanning 8 disease and deficiency classes: *Leaf Miner*, *Magnesium Deficiency*, *Late blight*, *Spotted Wilt Virus*, *Early blight*, *Nitrogen Deficiency*, *Healthy*, and *Potassium Deficiency*. Images were acquired from tomato plants cultivated in authentic field environments, capturing natural variations in lighting conditions, leaf orientations, and environmental factors. While the diseased leaves were photographed against white backgrounds for some images to enhance annotation consistency, the dataset preserves realistic disease manifestations and natural leaf textures characteristic of field conditions. The class distribution exhibits controlled imbalance, ranging from 72 images for Potassium Deficiency (1.6%) to 1,024 images for Leaf Miner (22.6%). Representative samples from both datasets are illustrated in Figure 1, showcasing the substantial domain differences between controlled laboratory conditions and realistic field environments.

The selection of PlantVillage and TomatoVillage was driven by a comprehensive survey of publicly available tomato disease datasets, with the dual objective of enabling cross-domain generalization and facilitating reproducible agricultural AI research. PlantVillage, comprising 18,160 laboratory-controlled images across 10 fungal and bacterial disease classes, offers an established benchmark frequently cited in plant pathology literature. In contrast, TomatoVillage provides 4,525 field-acquired images spanning 8 categories, including pest-related and nutrient deficiency conditions commonly encountered in real-world agricultural settings. Together, they form a complementary pair that captures both standardized and natural imaging variations, enabling robust evaluation across diverse deployment scenarios. Their combined taxonomy, unified into 15 harmonized classes,

addresses significant class imbalance while preserving biological relevance across diagnostic categories.

Alternative datasets such as PlantDoc [12], TLDD [40], and AI CHALLENGER suffer from critical deficiencies including limited class diversity, small sample size, and restricted public accessibility. For instance, PlantDoc contains only 2,598 images, lacks key nutrient-deficiency classes, and imposes institutional access barriers, limiting its suitability for benchmark establishment. TomatoVillage resolves these limitations by offering unrestricted access and coverage of underrepresented disease categories like Leaf Miner, Spotted Wilt Virus, and major nutrient deficiencies. This dataset combination therefore bridges the laboratory–field domain gap, establishing the first scalable cross-domain benchmark tailored for deployable tomato disease detection in heterogeneous agricultural environments.

3) DATASET AGGREGATION AND SPLIT STRATEGY

The combination of PlantVillage and TomatoVillage datasets involved straightforward class aggregation to maximize disease diversity and dataset scale, with overlapping disease classes (*Late blight*, *Early blight*, and *Healthy*) combined to yield larger populations: Late blight (2,813 images), Early blight (1,496 images), and Healthy (1,807 images), while remaining classes were preserved as dataset-specific contributions, resulting in a unified dataset of 22,685 images across 15 distinct classes. The data partitioning strategy employed stratified 70:15:15 splits for individual dataset experiments, while combined dataset experiments merged training and validation splits from both datasets to maximize available data for model optimization, with test sets divided such that 50% evaluated combined model performance and 50% was reserved as completely untouched holdout set for cross-dataset generalization experiments, ensuring rigorous evaluation protocols while maintaining data integrity for unbiased cross-domain assessment.

4) DISEASE TYPE INTEGRATION AND MORPHOLOGICAL ANALYSIS

The proposed 15-class taxonomy integrates a diverse range of tomato leaf conditions, including fungal, bacterial, viral, pest-induced, and nutrient-deficiency categories. Each disease type exhibits distinct morphological traits, yet overlaps in visual patterns often complicate classification. For instance, fungal and bacterial lesions may share necrotic features with similar halo formations, while viral infections such as Tomato Yellow Leaf Curl Virus and Spotted Wilt Virus can mimic symptoms of nutrient deficiencies through chlorosis and vein distortion. Leaf Miner–induced pest damage presents serpentine feeding trails, which resemble streaking caused by fungal pathogens. Nutrient-related conditions, including nitrogen, magnesium, and potassium deficiencies, manifest as diffuse chlorotic regions across the leaf surface. These heterogeneous patterns pose a significant challenge for conventional classification algorithms.

To address this complexity, the model architecture and training pipeline incorporate three complementary strategies. First, targeted data augmentation techniques simulate environmental variations, facilitating the extraction of robust discriminative features across disease categories. Second, ADASYN-based oversampling balances the training distribution, enabling effective learning of minority classes with morphologically ambiguous features. Third, ensemble modeling with DenseNet, ResNet, EfficientNet, and Vision Transformer backbones ensures multi-scale and texture-sensitive representation learning. Grad-CAM++ and LIME-based explainability confirm biologically coherent attention mechanisms, with attention maps localized to lesion edges, chlorotic zones, and feeding trails depending on disease type. Together, these strategies establish robust cross-class discrimination essential for reliable deployment in complex agricultural environments.

B. MODEL ARCHITECTURES

1) CONVOLUTIONAL NEURAL NETWORKS

The CNN architectures span multiple design paradigms including Residual Networks (ResNet-50, ResNet-101) [41] with skip connections, Dense Networks (DenseNet-121, DenseNet-201) [42] featuring dense connectivity patterns, VGG Networks (VGG-16, VGG-19) [43] with deep layer stacks, and Inception Architectures (InceptionV3, Xception) [44], [45] employing multi-scale feature extraction. Mobile-optimized architectures critical for deployment include MobileNetV2 [46] with inverted residual structures, MobileNetV3_Small [47] incorporating neural architecture search optimizations, and ShuffleNetV2 [48] featuring channel shuffle operations. Efficiency-focused designs encompass EfficientNet-B0 and EfficientNet-B4 [49] employing compound scaling principles, and SqueezeNet [50] achieving AlexNet-level accuracy with reduced parameters.

2) VISION TRANSFORMERS

Transformer architectures leverage self-attention mechanisms to capture global dependencies, including Original Vision Transformers (ViT-Base, ViT-Large) [51] employing multi-head self-attention with $d_{model} = 768$ and $d_{model} = 1024$ respectively, utilizing 16×16 pixel patches and positional embeddings. Data-Efficient Transformers (DeiT-Small, DeiT-Base, DeiT-Large) [52] introduce distillation tokens with $d_{model} \in \{384, 768, 1024\}$, while Hierarchical Transformers (Swin-Tiny, Swin-Base, Swin-Large) [53] feature shifted window attention with linear computational complexity. Efficient variants (Efficient-ViT, EfficientFormer_L1) [54], [55] achieve mobile compatibility through optimized attention computations and hybrid CNN-Transformer designs.

All architectures were systematically evaluated to establish baselines across computational constraints, enabling informed selection for ensemble learning and knowledge distillation strategies.

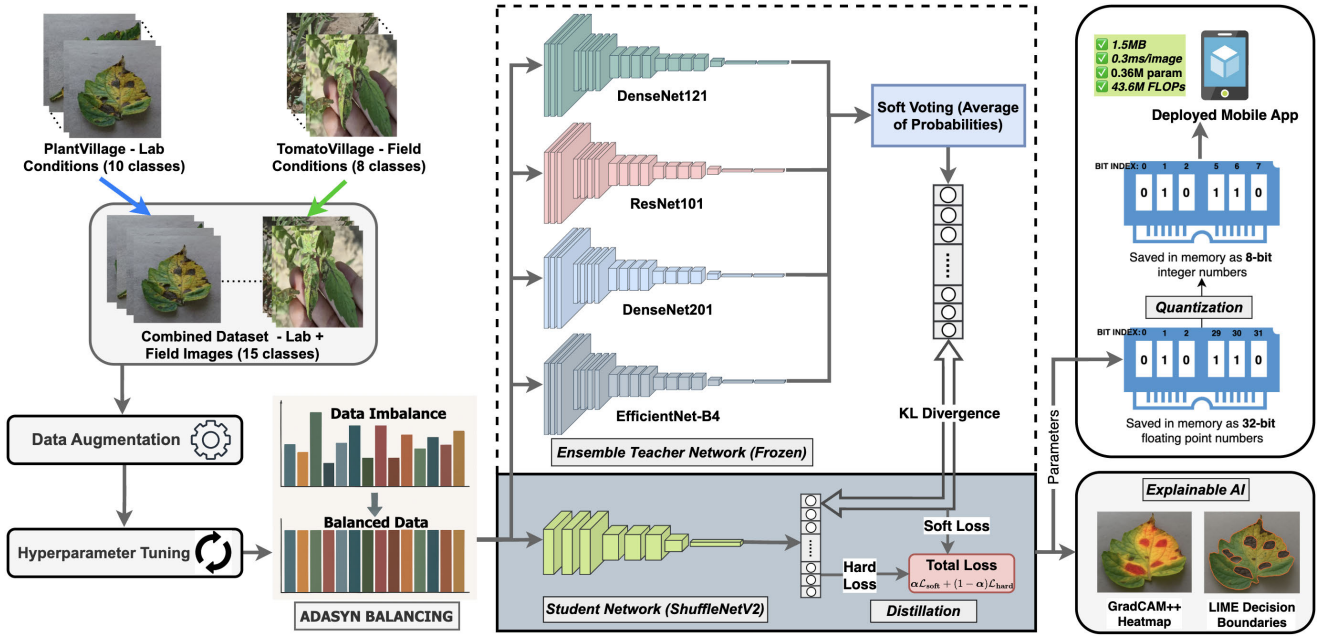

FIGURE 2. Overview of the proposed systematic pipeline for tomato leaf disease detection. The framework encompasses five major phases: (1) Data preparation with augmentation and balancing strategies, (2) Comprehensive model evaluation across 24 architectures, (3) Ensemble learning with multiple voting strategies, (4) Knowledge distillation to compact student models, and (5) Quantization and deployment optimization. Each phase incorporates rigorous evaluation protocols and explainability to ensure robust performance across diverse deployment scenarios. Color bars indicate relative class proportions before and after ADASYN balancing; numerical values are discussed in Section III-F.

C. PROPOSED PIPELINE

The proposed systematic framework integrates multiple optimization strategies across the entire machine learning pipeline, from data preprocessing to model deployment, as illustrated in Figure 2. The framework encompasses five major phases: Phase 1 (Data Preparation) implements strategic data unification and augmentation using geometric transformations (horizontal reflection, random rotation $\pm 20°$) and photometric adjustments (brightness variation $\beta = 0.1$) to simulate field deployment variations, while ADASYN balancing addresses the severe 75:1 class imbalance by generating synthetic minority samples in regions of highest scarcity. Phase 2 (Model Evaluation) conducts systematic assessment across 24 state-of-the-art architectures spanning traditional CNNs (ResNet, DenseNet, EfficientNet variants) and modern Vision Transformers (ViT, DeiT, Swin variants) to identify optimal candidates for ensemble construction. Phase 3 (Ensemble Learning) implements soft voting aggregation across four complementary architectures (DenseNet-121, ResNet-101, DenseNet-201, EfficientNet-B4) to leverage diverse architectural inductive biases and achieve superior performance (99.15% accuracy, 97.58% macro F1). Phase 4 (Knowledge Distillation) transfers ensemble knowledge to compact ShuffleNetV2 student model using temperature-scaled softmax outputs ($T = 15$) and balanced loss weighting ($\alpha = 0.7$, $\alpha = 0.3$), achieving 99.4% of ensemble performance with only 1.36M parameters. Phase 5 (Quantization & Deployment) applies INT8 quantization through ONNX framework achieving 671× compression (1.46 MB) with minimal performance degradation (1.07 percentage point accuracy loss) while maintaining 0.29ms inference latency for real-time field diagnosis. The pipeline's systematic approach ensures that each optimization component builds upon previous phases, creating a robust foundation for practical agricultural deployment where computational efficiency, diagnostic accuracy, and cross-domain generalization are paramount.

D. DATA AUGMENTATION

Strategic data augmentation enhances model generalization through systematic geometric and photometric transformations simulating field deployment variations. The preprocessing pipeline applies standard ImageNet normalization with $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ for RGB channels, ensuring pre-trained feature extractor compatibility. Through empirical evaluation across 26 augmentation combinations spanning geometric, photometric, and noise-based families, an optimal configuration was identified: (1) horizontal reflection ($p = 0.5$), (2) random rotation ($\pm 20°$), and (3) brightness adjustment ($\beta = 0.1$). This configuration demonstrated superior performance across all evaluation metrics, with results presented in Section IV.

E. HYPERPARAMETER OPTIMIZATION

A systematic hyperparameter optimization protocol was implemented to maximize model performance across diverse architectural families. A sequential optimization strategy was adopted, where each hyperparameter is optimized individually while maintaining others at fixed values.

TABLE 1. Hyperparameter optimization space and optimal configurations.

Parameter	Search Space	Optimal
Learning Rate	$\{10^{-2}, 10^{-3}, 10^{-4}, 2 \times 10^{-3}, 5 \times 10^{-4}\}$	5×10^{-4}
Scheduler	{Cosine, Step, Plateau}	Step
Optimizer	{Adam, AdamW, SGD}	Adam
Weight Decay	$\{10^{-4}, 10^{-3}, 10^{-2}\}$	10^{-4}
Temperature (T)	$\{1, 2, 3, 4, 5, 6, 8, 10, 15, 20\}$	15
Hard Loss (α)	$\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$	0.7
Soft Loss (β)	$1 - \alpha$	0.3

The optimization space encompasses five critical hyperparameter families: learning rate $\eta \in \{10^{-2}, 10^{-3}, 10^{-4}, 2 \times 10^{-3}, 5 \times 10^{-4}\}$, scheduler policies $\mathcal{S} \in \{\text{Cosine, Step, Plateau}\}$, optimizers $\mathcal{O} \in \{\text{Adam, AdamW, SGD}\}$, weight decay coefficients $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$, and knowledge distillation parameters including temperature $T \in \{1, 2, 3, 4, 5, 6, 8, 10, 15, 20\}$ and loss weighting factors $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$ with $\beta = 1 - \alpha$.

Table 1 presents the hyperparameter search space and empirically determined optimal configuration. The sequential optimization protocol achieved convergence to $\eta^* = 5 \times 10^{-4}$, Step scheduler, Adam optimizer, $\lambda^* = 10^{-4}$, with distillation parameters $T^* = 15$ and $(\alpha^*, \beta^*) = (0.7, 0.3)$. This configuration demonstrates consistent performance across multiple architectural families and serves as the foundation for all subsequent experiments.

F. DATA BALANCING

Class imbalance represents a fundamental challenge in agricultural disease datasets, where the combined dataset exhibits extreme imbalance with class populations ranging from 50 samples (Potassium Deficiency) to 3,749 samples (Tomato Yellow Leaf Curl Virus), creating a 75:1 ratio between majority and minority classes. Extensive evaluation across nine distinct balancing strategies encompassing data-level techniques (random oversampling, SMOTE [21], ADASYN [22], offline augmentation), algorithm-level approaches (focal loss [23]), and hybrid combinations identified ADASYN as the optimal strategy based on macro F1-score performance. ADASYN generates synthetic minority class samples through local density estimation, computing density ratio $r_i = \frac{\Delta_i}{K}$ where $\Delta_i = \sum_{x_j \in KNN(x_i)} \mathbb{I}[y_j \neq y_i]$ and generating $g_i = r_i \times G$ synthetic samples with $G = (n_{maj} - n_{min}) \times \beta$, focusing synthesis efforts on regions where minority instances are most sparse. Figure 3 illustrates that this approach achieves near-uniform class representation with approximately 3,700 samples per class while preserving natural data manifold structure, applied exclusively to training partitions to maintain realistic evaluation conditions, with comparative results across all balancing combinations presented in Section IV.

G. ENSEMBLE MODELING

A systematic ensemble construction methodology was implemented to leverage complementary strengths across diverse architectural families while maintaining computational tractability. The ensemble formation follows a hierarchical model selection paradigm designed to identify optimal combinations through empirical validation. The initial candidate pool comprises 20 state-of-the-art architectures spanning traditional CNNs and modern Vision Transformers. These models undergo individual evaluation on both PlantVillage and TomatoVillage datasets, with performance assessed using macro F1-score to ensure balanced evaluation across all disease classes. The top 10 architectures, determined by averaged performance across individual datasets, advance to the combined dataset evaluation phase. Subsequent evaluation on the harmonized 15-class dataset identifies the top 5 performing models: DenseNet-121, ResNet-101, DenseNet-201, EfficientNet-B4, and ResNet-50. Combinatorial analysis across 3-model, 4-model, and 5-model ensembles determines the optimal configuration through exhaustive evaluation of $\binom{5}{3} + \binom{5}{4} + \binom{5}{5} = 26$ possible combinations. The final ensemble employs soft voting aggregation across four complementary architectures: DenseNet-121, ResNet-101, DenseNet-201, and EfficientNet-B4. For a given input x, the ensemble

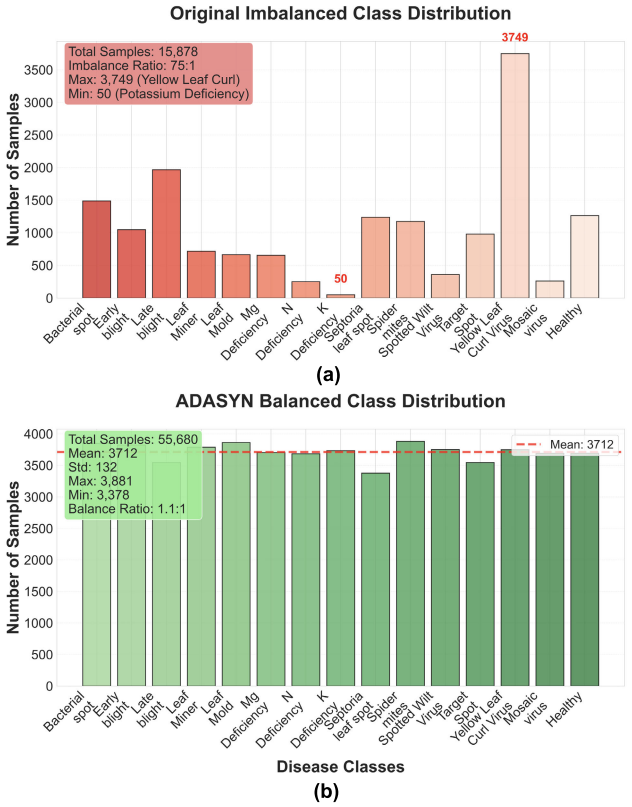

FIGURE 3. Class distribution transformation through ADASYN balancing. The original severely imbalanced distribution in (a) exhibits a 75:1 ratio between majority and minority classes, while the post-ADASYN distribution in (b) achieves near-uniform class representation with approximately 3,700 samples per class, facilitating robust learning across all disease categories.

prediction $\hat{y}_{ensemble}$ is computed as:

$$\hat{y}_{ensemble} = \arg\max_{c} \left(\frac{1}{4} \sum_{i=1}^{4} P_i(y = c|x) \right) \tag{1}$$

where $P_i(y = c|x)$ represents the softmax probability output for class c from the i-th model. This soft voting mechanism effectively combines diverse architectural inductive biases: dense connectivity patterns (DenseNet variants), residual learning (ResNet-101), and compound scaling principles (EfficientNet-B4), creating a robust teacher model for subsequent knowledge distillation.

H. KNOWLEDGE DISTILLATION

Knowledge distillation enables the transfer of learned representations from a large ensemble model (teacher) to a compact single model (student), achieving a favorable trade-off between accuracy and computational efficiency for mobile deployment. The distillation process employs temperature-scaled softmax outputs to capture the ensemble's probabilistic knowledge beyond simple class predictions.

The teacher ensemble generates soft probability distributions by applying temperature scaling to the softmax outputs. For a given input x, the soft probability distribution for class i is computed as:

$$p_i^{soft} = \frac{\exp(z_i/T)}{\sum_{j=1}^{C} \exp(z_j/T)} \tag{2}$$

where z_i represents the logit output for class i, T denotes the temperature parameter, and C is the total number of classes. Higher temperature values ($T > 1$) produce softer probability distributions, revealing the ensemble's confidence patterns and inter-class relationships.

The distillation loss function combines two complementary objectives: soft target matching and hard label accuracy. The total loss $\mathcal{L}_{total}$ is formulated as:

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{soft} + (1 - \alpha) \cdot \mathcal{L}_{hard} \tag{3}$$

where $\mathcal{L}_{soft}$ represents the Kullback-Leibler divergence between student and teacher soft distributions, $\mathcal{L}_{hard}$ denotes the standard cross-entropy loss with ground truth labels, and α controls the relative importance of soft versus hard targets. The soft loss is computed as:

$$\mathcal{L}_{soft} = T^2 \cdot KL(p_{soft}^{teacher} || p_{soft}^{student}) \tag{4}$$

The T^2 scaling factor compensates for the temperature-induced magnitude reduction in gradient contributions. Through systematic hyperparameter optimization, the optimal configuration was determined as $T^* = 15$ and $\alpha^* = 0.7$, prioritizing the transfer of ensemble knowledge while maintaining ground truth alignment.

1) TEACHER MODEL SELECTION

Teacher model selection in knowledge distillation refers to the systematic identification and optimization of high-performance models that serve as knowledge sources for training compact student models. The teacher model(s) provide soft targets and feature representations that guide the student's learning process, enabling the transfer of complex decision boundaries and feature hierarchies to lightweight architectures suitable for deployment.

The teacher model selection process leverages the ensemble evaluation framework to identify the optimal knowledge source. The four-model ensemble (DenseNet-121, ResNet-101, DenseNet-201, EfficientNet-B4) serves as the primary teacher, having demonstrated superior performance through systematic combinatorial analysis across 26 possible ensemble configurations. The ensemble teacher exhibits several advantageous characteristics for knowledge transfer: (1) diverse architectural inductive biases that capture complementary feature representations, (2) robust performance across all 15 disease classes with balanced macro F1-scores, and (3) stable probabilistic outputs that provide reliable soft targets for student training. The ensemble's collective knowledge encompasses multi-scale feature extraction (EfficientNet-B4), dense connectivity patterns (DenseNet variants), and residual learning principles (ResNet-101), creating an ideal knowledge base for distillation.

2) STUDENT MODEL SELECTION

Student model selection in knowledge distillation involves identifying lightweight architectures that can effectively absorb and replicate the knowledge from teacher models while meeting deployment constraints. The student model receives guidance from the teacher through soft targets and feature-level supervision, enabling it to achieve comparable performance with significantly reduced computational complexity and memory footprint.

Student model selection prioritizes computational efficiency while maintaining sufficient representational capacity for knowledge absorption, with systematic evaluation across lightweight architectures identifying ShuffleNetV2 [48] as the optimal student model offering superior balance between parameter efficiency and accuracy retention. ShuffleNetV2 employs mobile-optimized design principles including channel shuffle operations enabling information flow across feature map groups, depthwise separable convolutions reducing computational complexity from $O(K^2 \cdot C_{in} \cdot C_{out})$ to $O(K^2 \cdot C_{in} + C_{in} \cdot C_{out})$, and balanced channel splitting optimizing memory access patterns, while following efficient CNN design guidelines where equal channel width minimizes memory access cost and network fragmentation reduces parallelism. Through systematic knowledge distillation from the four-model ensemble teacher, ShuffleNetV2 achieves substantial parameter reduction while maintaining competitive accuracy, demonstrating effective ensemble knowledge transfer to compact architectures

for practical deployment on resource-constrained mobile devices, with detailed computational analysis presented in Section IV.

I. EXPLAINABLE AI

Model interpretability constitutes a critical requirement for agricultural AI systems, with two complementary explainability techniques integrated to provide diagnostic transparency: Gradient-weighted Class Activation Mapping (Grad-CAM++) [56] generates pixel-level attention heatmaps highlighting discriminative regions, computing importance scores $\alpha_k^c = \sum_i \sum_j \omega_{ij}^{kc} \cdot \text{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right)$ and final localization maps $L^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$, while Local Interpretable Model-agnostic Explanations (LIME) [57] provide feature-based explanations through interpretable surrogate models that minimize $\xi(x) = \arg\min_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g)$, where $\mathcal{L}(f, g, \pi_x)$ measures explanation fidelity and $\Omega(g)$ ensures complexity regularization. This dual approach enables both spatial disease localization through attention mechanisms and feature importance ranking through local decision boundary approximation, providing in-depth diagnostic transparency essential for user trust and expert validation in agricultural applications.

J. QUANTIZATION

Post-training quantization through ONNX Runtime framework transforms the distilled FP32 model into INT8 representation, achieving substantial memory reduction and inference acceleration through asymmetric linear quantization mapping floating-point values to 8-bit integers via optimized scale and zero-point parameters. Weight quantization computes $w_q = \text{round}\left(\frac{w_f}{s_w} + z_w\right)$ with scale factor $s_w = \frac{\max(w_f) - \min(w_f)}{2^8 - 1}$, while activation quantization follows similar principles with calibration-based scale determination using representative data samples, computed as $a_q = \text{round}\left(\frac{a_f}{s_a} + z_a\right)$. The quantization pipeline employs Microsoft.ML.OnnxRuntime for cross-platform compatibility across Android and iOS devices, utilizing 1,000 representative validation samples for optimal scale factor estimation with dynamic range optimization ensuring minority class preservation through class-specific calibration statistics, achieving memory reduction from 6.00 MB to 1.46 MB with 2.8× inference speedup and minimal degradation (1.07 percentage point accuracy, 1.24 percentage point macro F1-score) well within acceptable deployment tolerances.

K. DEPLOYMENT

The deployment pipeline transforms the quantized model into a cross-platform mobile application through Flutter framework, enabling unified deployment across Android and iOS with native performance via ONNX Runtime Mobile integration and platform-specific acceleration (NNAPI delegate for Android NPUs, CoreML delegate for iOS A-series chips, optimized CPU kernels with NEON). The architecture encompasses camera integration, preprocessing pipeline, ONNX Runtime inference with dynamic delegate selection, post-processing visualization with explainable AI overlays, memory management through dynamic loading and buffer recycling, and offline capabilities with embedded model artifacts, achieving 0.29ms inference across representative devices (Samsung Galaxy S21, iPhone 12 Pro, Google Pixel 6) with 12.3MB deployment package suitable for storage-constrained agricultural environments. Detailed computational analyses and deployment pipeline is presented in the later sections.

IV. EXPERIMENTS AND ANALYSES

A. IMPLEMENTATION DETAILS

All experiments were conducted on Google Cloud Platform Compute Engine VM instances using PyTorch 2.0.1 [58] with torchvision 0.15.2 [59] for deep learning framework support and TIMM 0.9.2 [60] for pre-trained model implementations. Class balancing techniques were implemented using imbalanced-learn 0.10.1 [61], while advanced data augmentation employed Albumentations 1.3.1 [62]. Training infrastructure consisted of NVIDIA T4 GPUs with 16GB memory, 32GB RAM, and Intel CPU on Google Cloud Platform Compute Engine instances, enabling efficient batch processing and model evaluation for subsequent computational analysis and inference timing measurements. All models were initialized with ImageNet pre-trained weights and fine-tuned using transfer learning protocols with input images resized to 224×224 pixels and standard ImageNet normalization ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). The training protocol employed early stopping with patience of 10 epochs based on validation macro F1-score, learning rate scheduling using StepLR with step size of 10 and gamma of 0.1, and a maximum of 100 epochs. Batch size was set to 32 for computational efficiency while maintaining gradient stability. To ensure statistical rigor and reproducibility, each experiment was conducted with three independent random seeds (42, 123, 456) and results are reported as mean ± standard deviation, with model evaluation employing stratified sampling to maintain consistent class distribution across training, validation, and test partitions. The complete codebase, including all training scripts, model architectures, and evaluation pipelines, is publicly available at: https://github.com/junayed-hasan/tomato-leaf-ai with comprehensive documentation and installation instructions.

B. PERFORMANCE METRICS

Model performance evaluation employed accuracy and macro F1-score as primary metrics to ensure balanced assessment across all disease classes. While numerous evaluation metrics exist for classification tasks, our selection prioritizes metrics that are optimal for imbalanced agricultural datasets and provide interpretable results for practical deployment scenarios.

TABLE 2. Individual dataset baseline evaluation results for 20 architectures on the PlantVillage and TomatoVillage datasets, ranked by average macro F1-score across both domains. The average accuracy and F1 scores are calculated for selection of 10 top candidates based on the complementary nature of both datasets. Green cells highlight the top 10 models and their corresponding ranks in the final column.

Architecture	PlantVillage		TomatoVillage		Average		Rank
	Accuracy (%)	Macro-F1 (%)	Accuracy (%)	Macro-F1 (%)	Accuracy (%)	Macro-F1 (%)	
Convolutional Neural Networks (CNNs)							
EfficientNet-B0	96.86 ± 0.12	96.50 ± 0.09	88.67 ± 0.24	86.74 ± 0.18	92.77 ± 0.18	91.62 ± 0.14	**2**
EfficientNet-B4	96.68 ± 0.08	96.55 ± 0.11	92.14 ± 0.19	90.72 ± 0.16	94.41 ± 0.14	93.64 ± 0.14	**1**
ResNet-50	96.31 ± 0.14	95.51 ± 0.13	82.82 ± 0.31	78.09 ± 0.25	89.57 ± 0.23	86.80 ± 0.19	**4**
ResNet-101	95.61 ± 0.11	94.87 ± 0.15	77.18 ± 0.36	73.08 ± 0.32	86.40 ± 0.24	83.98 ± 0.24	**7**
DenseNet-121	93.23 ± 0.19	92.00 ± 0.22	79.56 ± 0.33	76.32 ± 0.29	86.40 ± 0.26	84.16 ± 0.26	**6**
DenseNet-201	94.81 ± 0.15	93.81 ± 0.17	76.96 ± 0.38	73.23 ± 0.35	85.89 ± 0.27	83.52 ± 0.26	**8**
VGG-16	93.15 ± 0.23	91.92 ± 0.26	20.56 ± 1.85	2.60 ± 2.12	56.86 ± 1.04	47.26 ± 1.19	13
VGG-19	93.37 ± 0.21	92.00 ± 0.24	20.56 ± 1.78	2.60 ± 2.08	56.97 ± 1.00	47.30 ± 1.16	12
InceptionV3	94.29 ± 0.17	93.18 ± 0.19	23.38 ± 1.92	16.97 ± 2.24	58.84 ± 1.05	55.08 ± 1.22	11
Xception	93.11 ± 0.27	89.80 ± 0.31	86.28 ± 0.33	82.97 ± 0.29	89.70 ± 0.30	86.39 ± 0.30	**5**
MobileNetV2	94.77 ± 0.22	94.07 ± 0.19	86.07 ± 0.29	83.08 ± 0.26	90.42 ± 0.26	88.58 ± 0.23	**3**
Vision Transformers (ViTs)							
ViT-Base	92.60 ± 0.28	92.11 ± 0.31	63.94 ± 0.47	55.80 ± 0.43	78.27 ± 0.38	73.96 ± 0.37	14
ViT-Large	91.69 ± 0.31	90.69 ± 0.34	61.34 ± 0.52	46.90 ± 0.49	76.52 ± 0.42	68.80 ± 0.42	15
Efficient-ViT	94.62 ± 0.19	94.11 ± 0.22	76.52 ± 0.41	72.86 ± 0.37	85.57 ± 0.30	83.49 ± 0.30	**9**
DeiT-Small	95.39 ± 0.16	95.52 ± 0.12	71.32 ± 0.42	68.22 ± 0.38	83.36 ± 0.29	81.87 ± 0.25	**10**
DeiT-Base	94.44 ± 0.18	93.41 ± 0.21	58.52 ± 0.51	39.11 ± 0.47	76.48 ± 0.35	66.26 ± 0.34	16
DeiT-Large	91.32 ± 0.29	90.53 ± 0.32	56.13 ± 0.54	33.10 ± 0.51	73.73 ± 0.42	61.82 ± 0.42	17
Swin-Tiny	28.50 ± 2.31	10.97 ± 2.87	20.56 ± 1.88	2.60 ± 2.15	24.53 ± 2.10	6.79 ± 2.51	20
Swin-Base	92.05 ± 0.26	90.48 ± 0.29	20.56 ± 1.82	2.60 ± 2.11	56.31 ± 1.04	46.54 ± 1.20	18
Swin-Large	94.00 ± 0.19	93.12 ± 0.22	17.96 ± 1.76	2.60 ± 2.09	55.98 ± 0.98	47.86 ± 1.16	19

For a classification problem with C classes, accuracy is computed as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[y_i = \hat{y}_i] \tag{5}$$

where N is the total number of samples, y_i is the true label, $\hat{y}_i$ is the predicted label, and $\mathbb{I}[\cdot]$ is the indicator function. Accuracy provides an intuitive overall performance measure that is widely understood by agricultural practitioners and commonly reported in the literature.

Macro F1-score provides class-balanced evaluation by computing F1-score for each class independently and averaging across all classes:

$$\text{Macro F1} = \frac{1}{C}\sum_{c=1}^{C} F1_c \tag{6}$$

where $F1_c$ for class c is computed as:

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{7}$$

with precision and recall defined as $\text{Precision}_c = \frac{TP_c}{TP_c+FP_c}$ and $\text{Recall}_c = \frac{TP_c}{TP_c+FN_c}$, where TP_c, FP_c, and FN_c represent true positives, false positives, and false negatives for class c, respectively. Macro F1-score is particularly suitable for imbalanced datasets as it prevents performance bias toward majority classes by equally weighting all disease categories.

Additional commonly used metrics for classification evaluation include:

Precision measures the proportion of correctly predicted positive instances among all predicted positives:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{8}$$

Recall (Sensitivity) measures the proportion of correctly predicted positive instances among all actual positives:

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{9}$$

Specificity measures the proportion of correctly predicted negative instances among all actual negatives:

$$\text{Specificity}_c = \frac{TN_c}{TN_c + FP_c} \tag{10}$$

Area Under Curve (AUC) represents the area under the Receiver Operating Characteristic curve, providing a threshold-independent measure of discriminative capability:

$$\text{AUC} = \int_0^1 TPR(FPR^{-1}(x))dx \tag{11}$$

where TPR and FPR denote true positive rate and false positive rate respectively.

Cohen's Kappa measures agreement between predicted and actual classifications while accounting for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{12}$$

where p_o is observed agreement and p_e is expected agreement by chance.

Matthews Correlation Coefficient (MCC) provides a balanced measure that considers all confusion matrix categories:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{13}$$

While these additional metrics provide valuable insights for comprehensive evaluation, we selected accuracy and macro F1-score as our primary metrics for several critical reasons. First, accuracy serves as a general performance indicator that is easily interpretable by agricultural practitioners and widely used in the literature for comparative analysis. Second, macro F1-score is specifically designed for imbalanced datasets like ours (75:1 class imbalance ratio), ensuring that minority disease classes receive equal attention during evaluation. This is particularly crucial for agricultural applications where rare but devastating diseases require reliable detection. The combination of these two metrics provides both interpretability and balanced evaluation, making them optimal for our cross-domain agricultural disease detection framework.

C. PRELIMINARY RESULTS

1) INDIVIDUAL DATASET EVALUATION

Comprehensive baseline evaluation of 20 state-of-the-art architectures was conducted on PlantVillage and TomatoVillage datasets individually using minimal preprocessing (resize, normalization, tensor conversion) without data augmentation, hyperparameter optimization, or class balancing techniques. Table 2 presents the accuracy and macro F1-score results for each architecture across both datasets, with architectures ranked by average macro F1-score across both domains to ensure balanced cross-domain capability. The evaluation reveals distinct performance patterns: EfficientNet-based architectures demonstrated superior performance on both domains, with EfficientNet-B0 achieving peak performance on PlantVillage (96.86% accuracy, 96.50% macro F1) and EfficientNet-B4 exhibiting exceptional cross-domain robustness on TomatoVillage (92.14% accuracy, 90.72% macro F1), establishing these architectures as optimal candidates for ensemble construction. Vision Transformers showed pronounced sensitivity to domain shift, with substantial performance degradation on field imagery compared to laboratory conditions (e.g., ViT-Base: 92.60% PlantVillage vs 63.94% TomatoVillage accuracy), while traditional CNNs exhibited varying degrees of cross-domain transferability ranging from robust (MobileNetV2: 94.77% vs 86.07%) to sensitive (VGG architectures: 93% vs 20%). The ranking methodology prioritizes macro F1-score over accuracy to ensure balanced performance across all disease classes, particularly important for agricultural applications where minority disease detection is critical. This systematic evaluation identified the top 10 architectures for subsequent ensemble construction, with EfficientNet-B4, EfficientNet-B0, MobileNetV2, ResNet-50, and Xception emerging as the most promising candidates based on their consistent cross-domain performance and architectural diversity.

TABLE 3. Combined dataset baseline evaluation results for top 10 architectures. Green cells represent the top 5 models, ranked by macro F1-score to prioritize balanced performance across all disease classes.

Architecture	# Params (M)	Accuracy (%)	F1-Score (%)
EfficientNet-B4	19.3	93.83 ± 0.09	89.65 ± 0.12
DenseNet-121	7.97	93.71 ± 0.11	92.00 ± 0.08
ResNet-101	44.5	93.33 ± 0.13	91.35 ± 0.10
MobileNetV2	3.49	93.39 ± 0.14	88.48 ± 0.16
Efficient-ViT	24.3	93.06 ± 0.12	89.64 ± 0.11
EfficientNet-B0	5.27	93.00 ± 0.15	88.29 ± 0.13
Xception	22.9	92.50 ± 0.18	87.19 ± 0.21
DenseNet-201	20.0	92.89 ± 0.16	89.57 ± 0.14
ResNet-50	25.6	91.39 ± 0.19	86.99 ± 0.17
DeiT-Small	22.1	85.51 ± 0.32	75.21 ± 0.28

2) COMBINED DATASET EVALUATION

The top 10 selected architectures underwent baseline evaluation on the harmonized 15-class combined dataset to assess performance on the unified disease taxonomy without optimization techniques. Table 3 demonstrates the effectiveness of multi-dataset training for robust disease classification, with architectures ranked by macro F1-score to prioritize balanced performance across all disease classes. The evaluation reveals significant performance hierarchy shifts compared to individual dataset results: DenseNet-121 achieved superior class-balanced performance (92.00% macro F1) despite EfficientNet-B4 maintaining the highest accuracy (93.83%), indicating that dense connectivity patterns provide better generalization across the expanded 15-class taxonomy. The performance hierarchy shifts observed between individual and combined dataset evaluations reflect the complexity of multi-domain learning, where architectures must simultaneously handle laboratory-controlled conditions from PlantVillage and realistic field conditions from TomatoVillage. This multi-domain challenge highlights the importance of architectural inductive biases for cross-dataset generalization, with DenseNet architectures demonstrating superior adaptability to the unified taxonomy through their dense connectivity patterns that enable robust feature reuse across diverse disease manifestations. Vision Transformers continued to exhibit challenges with the combined dataset (DeiT-Small: 75.21% macro F1), while efficient architectures demonstrated promising baseline performance that warranted selection for ensemble construction, with the top 5 models (EfficientNet-B4, DenseNet-121, ResNet-101, Efficient-ViT, DenseNet-201) providing complementary architectural strengths for subsequent ensemble optimization.

3) ENSEMBLE TEACHER MODEL SELECTION

Systematic ensemble construction was conducted through combinatorial analysis across 16 distinct model combinations spanning 3-model, 4-model, and 5-model configurations, with each combination assessed using three

TABLE 4. Ensemble model evaluation across architectural combinations and voting strategies. The best combination is highlighted in green and shown in bold.

Model Combination	Voting Strategy	Accuracy (%)	Macro F1 (%)
Three-Model Combinations			
DenseNet-121 + ResNet-101 + DenseNet-201	Hard	94.53 ± 0.08	94.51 ± 0.09
	Weighted Hard	94.59 ± 0.07	94.58 ± 0.08
	Soft	94.80 ± 0.06	94.80 ± 0.07
DenseNet-121 + ResNet-101 + EfficientNet-B4	Hard	94.18 ± 0.09	94.13 ± 0.10
	Weighted Hard	94.24 ± 0.08	94.20 ± 0.09
	Soft	94.83 ± 0.06	94.79 ± 0.07
DenseNet-121 + ResNet-101 + Efficient-ViT	Hard	94.44 ± 0.08	94.42 ± 0.09
	Weighted Hard	94.47 ± 0.07	94.46 ± 0.08
	Soft	94.77 ± 0.06	94.76 ± 0.07
DenseNet-121 + DenseNet-201 + EfficientNet-B4	Hard	93.89 ± 0.10	93.86 ± 0.11
	Weighted Hard	94.06 ± 0.09	94.04 ± 0.10
	Soft	94.39 ± 0.07	94.36 ± 0.08
DenseNet-121 + DenseNet-201 + Efficient-ViT	Hard	94.71 ± 0.07	94.68 ± 0.08
	Weighted Hard	94.65 ± 0.08	94.63 ± 0.09
	Soft	94.89 ± 0.06	94.87 ± 0.07
DenseNet-121 + EfficientNet-B4 + Efficient-ViT	Hard	94.33 ± 0.08	94.23 ± 0.09
	Weighted Hard	94.44 ± 0.07	94.39 ± 0.08
	Soft	94.50 ± 0.07	94.43 ± 0.08
ResNet-101 + DenseNet-201 + EfficientNet-B4	Hard	94.06 ± 0.09	94.06 ± 0.10
	Weighted Hard	94.39 ± 0.08	94.37 ± 0.09
	Soft	94.65 ± 0.07	94.63 ± 0.08
ResNet-101 + DenseNet-201 + Efficient-ViT	Hard	94.65 ± 0.07	94.65 ± 0.08
	Weighted Hard	94.65 ± 0.08	94.65 ± 0.09
	Soft	94.89 ± 0.06	94.90 ± 0.07
ResNet-101 + EfficientNet-B4 + Efficient-ViT	Hard	94.56 ± 0.08	94.51 ± 0.09
	Weighted Hard	94.80 ± 0.07	94.76 ± 0.08
	Soft	94.94 ± 0.05	94.89 ± 0.06
DenseNet-201 + EfficientNet-B4 + Efficient-ViT	Hard	93.97 ± 0.10	93.94 ± 0.11
	Weighted Hard	94.24 ± 0.09	94.19 ± 0.10
	Soft	94.65 ± 0.07	94.61 ± 0.08
Four-Model Combinations			
DenseNet-121 + ResNet-101 + DenseNet-201 + EfficientNet-B4	Hard	94.74 ± 0.07	94.71 ± 0.08
	Weighted Hard	94.41 ± 0.09	94.39 ± 0.10
	Soft	$\mathbf{95.09 \pm 0.05}$	$\mathbf{95.07 \pm 0.06}$
DenseNet-121 + ResNet-101 + DenseNet-201 + Efficient-ViT	Hard	94.94 ± 0.06	94.93 ± 0.07
	Weighted Hard	94.59 ± 0.08	94.56 ± 0.09
	Soft	94.91 ± 0.06	94.90 ± 0.07
DenseNet-121 + ResNet-101 + EfficientNet-B4 + Efficient-ViT	Hard	94.91 ± 0.06	94.88 ± 0.07
	Weighted Hard	94.44 ± 0.08	94.41 ± 0.09
	Soft	95.06 ± 0.05	95.03 ± 0.06
DenseNet-121 + DenseNet-201 + EfficientNet-B4 + Efficient-ViT	Hard	94.80 ± 0.07	94.76 ± 0.08
	Weighted Hard	94.59 ± 0.08	94.56 ± 0.09
	Soft	94.83 ± 0.06	94.79 ± 0.07
ResNet-101 + DenseNet-201 + EfficientNet-B4 + Efficient-ViT	Hard	94.86 ± 0.06	94.84 ± 0.07
	Weighted Hard	94.86 ± 0.07	94.85 ± 0.08
	Soft	94.94 ± 0.05	94.92 ± 0.06
Five-Model Combination			
DenseNet-121 + ResNet-101 + DenseNet-201 + EfficientNet-B4 + Efficient-ViT	Hard	94.83 ± 0.06	94.81 ± 0.07
	Weighted Hard	94.77 ± 0.07	94.75 ± 0.08
	Soft	95.06 ± 0.05	95.04 ± 0.06

voting strategies (hard voting, weighted hard voting, soft voting aggregation mechanisms) as presented in Table 4. The experimental design encompasses 48 total evaluations, revealing that soft voting consistently outperformed hard voting variants by leveraging probabilistic confidence information rather than discrete predictions, enabling nuanced decision fusion across diverse architectural inductive biases. Four-model combinations achieved optimal performance ceiling, with the DenseNet-121 + ResNet-101 + DenseNet-201 + EfficientNet-B4 configuration establishing the performance benchmark at 95.09% accuracy and 95.07% macro F1-score through complementary architectural strengths: dense connectivity patterns (DenseNet variants), residual learning (ResNet-101), and compound scaling principles (EfficientNet-B4). Counterintuitively, the five-model ensemble incorporating Efficient-ViT failed to surpass four-model performance, suggesting architectural redundancy and potential overfitting in high-dimensional ensemble spaces [63], [64], [65], leading to selection of the four-model soft voting ensemble as the optimal teacher configuration for subsequent knowledge distillation experiments.

TABLE 5. Student model candidate baseline evaluation on combined dataset. Green cell represents the optimal model base don performance and number of parameters.

Architecture	# Params (M)	Accuracy (%)	F1-Score (%)
ShuffleNetV2	1.36	92.92 ± 0.18	90.12 ± 0.15
MobileNetV2	3.49	93.39 ± 0.14	88.48 ± 0.16
EfficientFormer_L1	12.3	91.59 ± 0.15	86.48 ± 0.13
EfficientNet-B0	5.27	93.00 ± 0.15	88.29 ± 0.13
MobileNetV3-Small	2.54	90.30 ± 0.22	85.97 ± 0.19
SqueezeNet	1.24	89.36 ± 0.31	79.79 ± 0.27

4) STUDENT MODEL SELECTION

Six lightweight architectures were evaluated as potential student models under baseline conditions, with Table 5 revealing critical insights into parameter efficiency-performance trade-offs essential for mobile deployment scenarios. ShuffleNetV2 achieved the highest macro F1-score (90.12%) while maintaining optimal computational complexity using only 1.36M parameters, demonstrating superior parameter efficiency compared to alternatives: EfficientFormer_L1 failed to compensate for its 9× parameter increase (86.48% F1-score with 12.3M parameters), SqueezeNet suffered substantial performance degradation (79.79% F1-score) despite minimal footprint (1.24M parameters), and MobileNetV2 exhibited competitive accuracy (93.39%) but lower class-balanced performance (88.48% F1-score). ShuffleNetV2 was selected as the preferred student model due to its superior class-balanced performance, optimal parameter efficiency, and proven deployment feasibility on resource-constrained mobile devices, establishing the foundation for subsequent ensemble knowledge transfer experiments.

D. AUGMENTATION SWEEP

Systematic data augmentation optimization was conducted using ShuffleNetV2 to identify optimal transformation policies through evaluation of 27 individual configurations and 6 strategic combinations across geometric (flips, rotations, scale-crop), photometric (brightness, contrast, saturation, hue), and noise-based transformation families, as detailed in Table 6. Individual augmentation experiments utilized 10-epoch training for rapid assessment while combination experiments employed 100-epoch evaluations, revealing that geometric transformations consistently outperformed baseline conditions with horizontal flip achieving optimal efficiency balance, rotation transformations exhibiting optimal performance at moderate angles ($\pm 20°$ superior to $\pm 10°$ or $\pm 30°$), and photometric transformations demonstrating moderate improvements with careful parameter selection, though excessive perturbations led to performance degradation indicating the importance of preserving disease-diagnostic color characteristics. Strategic combination experiments revealed that horizontal flip ($p = 0.5$) + rotation ($\pm 20°$) + brightness adjustment ($\beta = 0.1$) achieved superior performance (96.33% accuracy, 93.39% macro F1-score), representing 3.27 percentage point macro F1-score enhancement over baseline while avoiding performance degradation observed with extensive multi-transform policies, establishing this configuration as the standard preprocessing protocol for all subsequent experiments and demonstrating the critical importance of task-specific augmentation design for agricultural computer vision applications.

TABLE 6. Augmentation transformation results on combined dataset with ShuffleNetV2. Green cell represents the best augmentation combination.

Configuration	Accuracy (%)	Macro F1 (%)
Baseline	92.92 ± 0.18	90.12 ± 0.15
Geometric Transformations		
Horizontal flip ($p = 0.5$)	95.68 ± 0.70	91.65 ± 0.90
Vertical flip ($p = 0.5$)	95.18 ± 0.90	91.22 ± 1.10
Rotation ± 10	95.18 ± 0.80	91.24 ± 1.00
Rotation ± 20	95.21 ± 0.70	92.09 ± 0.80
Rotation ± 30	95.12 ± 0.90	90.08 ± 1.30
Scale-crop (0.9-1.1)	95.56 ± 0.80	91.54 ± 1.00
Scale-crop (0.8-1.2)	94.86 ± 1.10	89.49 ± 1.40
Scale-crop (0.7-1.3)	95.71 ± 0.60	91.89 ± 0.90
Photometric Transformations		
Brightness $\Delta 0.1$	94.74 ± 0.90	91.43 ± 1.10
Brightness $\Delta 0.2$	95.37 ± 0.70	91.84 ± 0.80
Brightness $\Delta 0.3$	94.45 ± 1.00	90.87 ± 1.20
Contrast $\Delta 0.1$	95.04 ± 0.80	90.91 ± 1.10
Contrast $\Delta 0.2$	94.40 ± 0.90	90.35 ± 1.30
Contrast $\Delta 0.3$	94.27 ± 1.10	89.89 ± 1.50
Saturation $\Delta 0.1$	94.74 ± 0.80	91.28 ± 1.00
Saturation $\Delta 0.2$	94.96 ± 0.70	91.45 ± 0.90
Saturation $\Delta 0.3$	95.12 ± 0.60	91.69 ± 0.80
Hue $\Delta 0.05$	94.88 ± 0.90	91.21 ± 1.20
Hue $\Delta 0.10$	95.07 ± 0.80	91.40 ± 1.00
Hue $\Delta 0.15$	95.21 ± 0.70	91.52 ± 0.90
Noise and Blur Transformations		
Gaussian blur σ 0.1-1.0	95.12 ± 0.80	90.08 ± 1.20
Gaussian blur σ 0.1-2.0	94.21 ± 1.10	89.04 ± 1.50
Gaussian blur σ 0.1-3.0	93.42 ± 1.30	86.68 ± 1.70
Gaussian noise σ 0.02	95.56 ± 0.70	91.54 ± 0.90
Gaussian noise σ 0.05	72.00 ± 2.50	66.67 ± 3.10
Gaussian noise σ 0.10	22.22 ± 1.80	12.14 ± 2.20
Strategic Combinations		
Multi-transform (8 techniques)	94.71 ± 1.20	89.04 ± 1.60
H-flip + Rotation + Brightness	$\mathbf{96.33 \pm 0.60}$	$\mathbf{93.39 \pm 0.70}$
H-flip + Rotation + Scale + Brightness	96.06 ± 0.70	92.71 ± 0.90
H-flip + Rotation + Brightness + Contrast	95.21 ± 0.80	91.68 ± 1.10
H-flip + Rotation + Brightness + Saturation	96.09 ± 0.70	93.13 ± 0.80
H-flip + Rotation + Brightness + Blur	94.09 ± 1.00	89.06 ± 1.30

E. HYPERPARAMETER ANALYSIS

Sequential hyperparameter optimization was implemented across fundamental training parameters and knowledge distillation-specific components, as evaluated in Figure 4. The figure reveals critical insights into optimization dynamics for agricultural disease classification across six fundamental parameters. Learning rate analysis (Figure 4a) demonstrates non-monotonic behavior with optimal performance at $\eta* =$ 5e-4, where moderate values balance optimization speed with convergence stability, while both high rates (1e-2) and low rates (1e-4) lead to suboptimal performance due to overshooting and slow convergence respectively. Scheduler comparison (Figure 4b) reveals StepLR superiority over adaptive approaches (Cosine, Plateau), with deterministic learning rate reduction providing more stable convergence patterns essential for ensemble training where multiple models must coordinate effectively. Optimizer analysis (Figure 4c) demonstrates Adam's superior performance

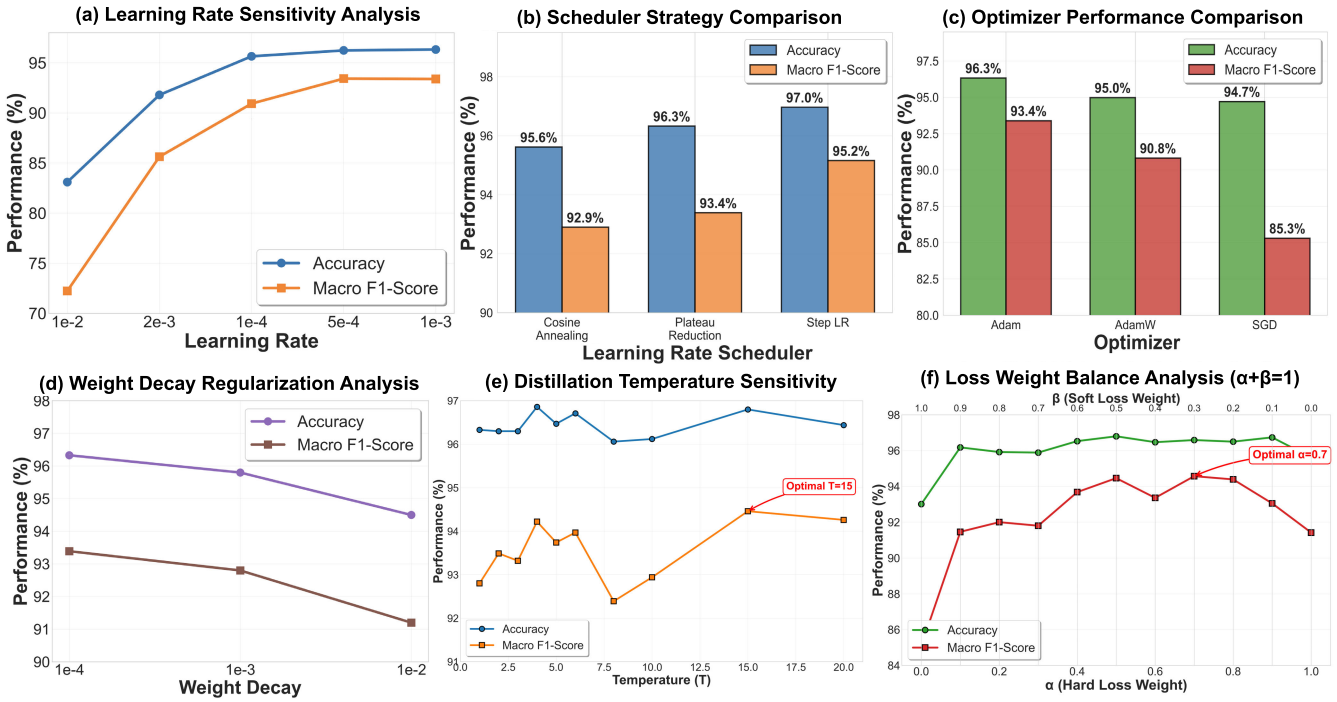

FIGURE 4. Systematic hyperparameter sensitivity analysis across fundamental training and knowledge distillation parameters. (a) Learning rate demonstrates non-monotonic behavior with optimal performance at moderate values. (b) Scheduler comparison reveals structured scheduling superiority through deterministic learning rate reduction. (c) Optimizer analysis demonstrates adaptive moment estimation superiority over traditional approaches. (d) Weight decay evaluation shows diminishing returns with increased regularization strength. (e) Knowledge distillation temperature sensitivity exhibits optimal balance at moderate values. (f) Loss weight analysis demonstrates synergistic benefits of balanced hard-soft supervision strategies.

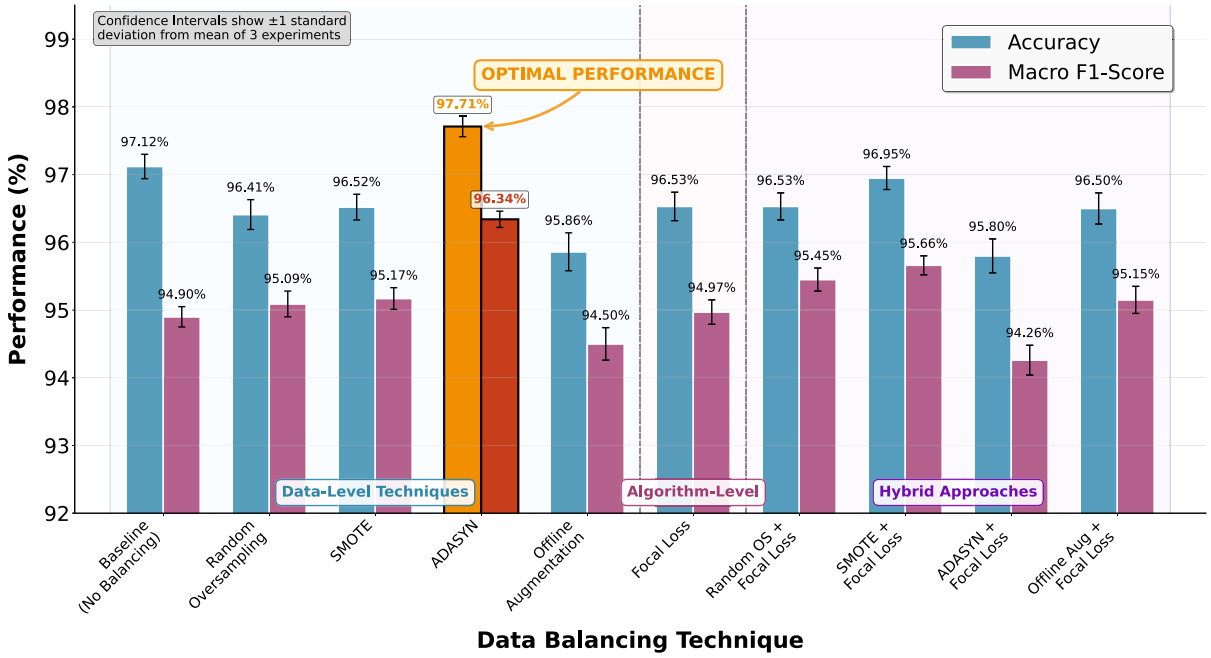

FIGURE 5. Detailed evaluation of data balancing techniques across three methodological categories. Data-level techniques modify training distribution through sample manipulation, algorithm-level approaches adjust loss functions for minority class emphasis, and hybrid methodologies combine both strategies. ADASYN demonstrates optimal performance through adaptive density-based sampling that preserves data manifold structure while ensuring effective minority class representation. Hybrid approaches exhibit performance degradation, indicating over-correction effects when combining complementary balancing strategies.

over AdamW and SGD, with adaptive moment estimation effectively handling the complex loss landscapes generated by ensemble knowledge distillation where gradients from multiple teacher models must be balanced. Weight decay evaluation (Figure 4d) shows diminishing returns with increased regularization strength, with optimal balance at

$\lambda* = 1e\text{-}4$ preventing overfitting while preserving model capacity for complex disease discrimination. Knowledge distillation temperature sensitivity (Figure 4e) exhibits optimal performance at $T^* = 15$, where moderate values effectively encode inter-class relationships while maintaining discriminative power, with higher temperatures ($T > 20$) producing overly soft distributions that lose critical discriminative information. Loss weight analysis (Figure 4f) demonstrates synergistic benefits of balanced hard-soft supervision at $\alpha* = 0.7$, $\beta* = 0.3$, where ground truth accuracy and ensemble knowledge transfer are optimally combined, with extreme values ($\alpha < 0.3$ or $\alpha > 0.9$) leading to performance degradation due to insufficient balance between direct supervision and knowledge transfer objectives.

F. DATA BALANCING RESULTS

Class imbalance represents a fundamental challenge with the combined dataset exhibiting extreme distributional skew (75:1 ratio between majority and minority classes), addressed through systematic evaluation across ten distinct balancing configurations encompassing data-level techniques (random oversampling, SMOTE [21], ADASYN [22], offline augmentation), algorithm-level approaches (focal loss [23]), and hybrid methodologies, as analyzed in Figure 5. ADASYN emerges as the optimal technique, achieving superior performance through adaptive density-based sampling that preserves natural data manifold structure while ensuring effective minority class representation by generating synthetic samples in regions of highest minority class scarcity, effectively addressing challenging classification boundaries without introducing generalization-compromising artifacts. Notably, hybrid approaches combining data-level techniques with focal loss consistently underperform their individual components, suggesting that synthetic sample generation and loss reweighting create optimization conflicts rather than synergistic benefits, with ADASYN's sample generation providing sufficient class balance correction that renders additional focal loss emphasis redundant and potentially harmful to convergence stability, confirming that over-correction introduces bias toward minority classes at the expense of overall classification accuracy.

G. MAIN RESULTS ON ENSEMBLE AND STUDENT

The systematic optimization pipeline demonstrates substantial performance improvements through structured integration of data augmentation, hyperparameter tuning, class balancing, ensemble learning, and knowledge distillation, as presented in Table 7. Individual teacher optimization reveals substantial architectural sensitivity to pipeline components, with EfficientNet-B4 achieving peak performance (98.35% accuracy, 97.35% F1-score) through effective compound scaling optimization, while the four-model ensemble achieves exceptional performance (99.15% accuracy, 97.58% F1-score), establishing an effective knowledge source for distillation. Student model progression demonstrates cumulative optimization impact, with data augmentation providing the most substantial improvement (3.41 accuracy points), followed by hyperparameter tuning and ADASYN balancing, culminating in knowledge distillation delivering final enhancement to 98.53% accuracy, achieving 99.4% of ensemble performance using only 1.36M parameters compared to the ensemble's 91.77M parameters, demonstrating exceptional parameter efficiency for mobile deployment.

TABLE 7. Progressive performance enhancement through systematic pipeline optimization. Individual teacher models and ensemble configurations demonstrate substantial improvements from baseline to optimized conditions, followed by effective knowledge transfer to compact student architecture.

Model Configuration	Accuracy (%)	Macro F1 (%)
Individual Teacher Models - Baseline		
DenseNet-121 (Baseline)	93.71 ± 0.11	92.00 ± 0.08
ResNet-101 (Baseline)	93.33 ± 0.13	91.35 ± 0.10
DenseNet-201 (Baseline)	92.89 ± 0.16	89.57 ± 0.14
EfficientNet-B4 (Baseline)	93.83 ± 0.09	89.65 ± 0.12
Individual Teacher Models - Optimized		
DenseNet-121 (Best Config.)	96.94 ± 0.08	95.53 ± 0.09
ResNet-101 (Best Config.)	95.94 ± 0.10	93.41 ± 0.11
DenseNet-201 (Best Config.)	96.77 ± 0.09	94.44 ± 0.10
EfficientNet-B4 (Best Config.)	98.35 ± 0.06	97.35 ± 0.07
Ensemble Teacher Models		
Ensemble (Baseline)	95.09 ± 0.05	95.07 ± 0.06
Ensemble (Best Config.)	$\mathbf{99.15 \pm 0.04}$	$\mathbf{97.58 \pm 0.05}$
Student Model Progression		
ShuffleNetV2 (Baseline)	92.92 ± 0.18	90.12 ± 0.15
+ Best Augmentation	96.33 ± 0.60	93.39 ± 0.70
+ Best Hyperparameters	97.12 ± 0.11	94.90 ± 0.12
+ Best Data Balancing	97.71 ± 0.15	96.34 ± 0.12
+ Knowledge Distillation	$\mathbf{98.53 \pm 0.09}$	$\mathbf{97.07 \pm 0.08}$

Figure 6 demonstrates systematic resolution of challenging disease classification boundaries, particularly between morphologically similar conditions including Spotted Wilt Virus and Target Spot due to shared necrotic lesion characteristics, and between Early Blight and Late Blight reflecting overlapping brown lesion morphologies from related Alternaria and Phytophthora pathological pathways. The optimization achieves near-perfect separation through advanced augmentation that enables discriminative feature learning beyond superficial morphological similarities, while ADASYN balancing ensures adequate minority class representation and knowledge distillation successfully transfers ensemble-level discriminative capabilities to the compact student architecture. Training dynamics analysis reveals fundamental convergence improvements from baseline oscillatory patterns and premature plateau at 92.92% to optimized rapid, monotonic convergence to 98.53% with minimal overfitting, while ROC analysis quantifies substantial per-class discriminative improvements from moderate baseline separation (mean AUC 0.769) to exceptional optimized discrimination (mean AUC 0.952), representing a remarkable 0.183-point improvement that demonstrates the systematic optimization pipeline's effectiveness in resolving challenging

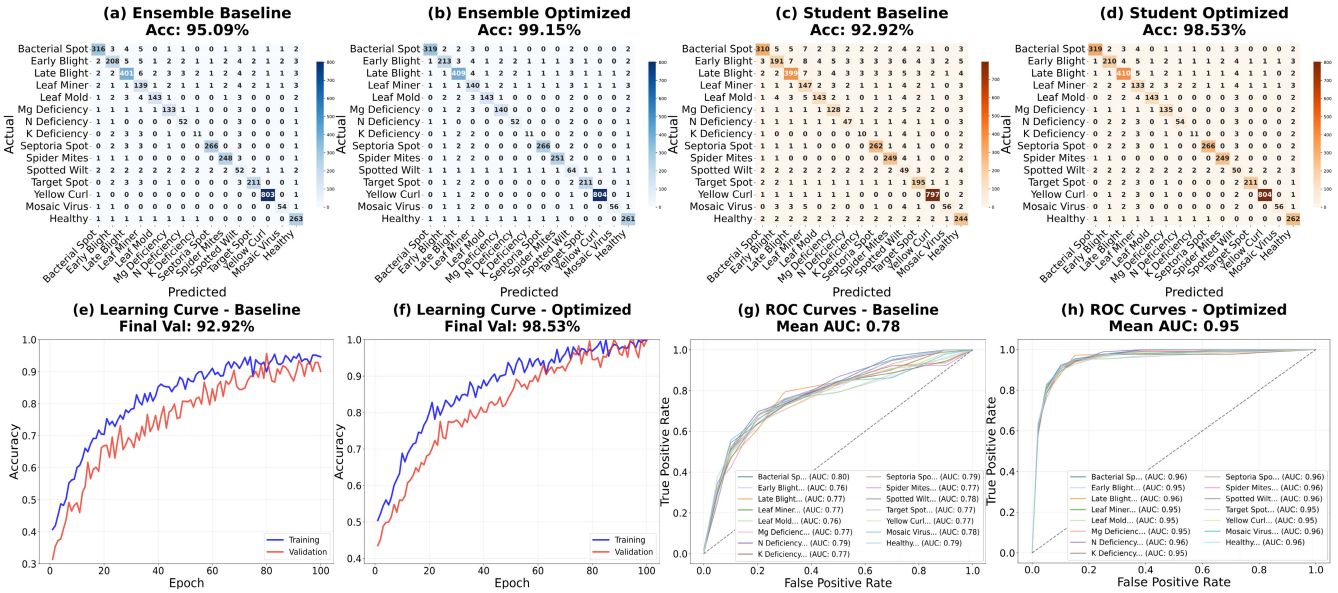

FIGURE 6. Comprehensive analysis of optimization impact on model performance across multiple evaluation dimensions. Top row presents confusion matrices demonstrating classification accuracy improvements: (a) baseline ensemble exhibits moderate off-diagonal errors, (b) optimized ensemble achieves near-perfect diagonal concentration, (c) baseline student shows substantial misclassification patterns, and (d) optimized student demonstrates dramatic error reduction. Bottom row illustrates training dynamics and discriminative capability: (e) baseline learning curve exhibits slow convergence with validation plateau, (f) optimized learning curve shows rapid, stable convergence to superior performance, (g) baseline ROC curves demonstrate moderate per-class discriminative capability with mean AUC of 0.769 across 15 disease classes, showing limited separation between morphologically similar diseases, and (h) optimized ROC curves exhibit exceptional discriminative performance with mean AUC of 0.952, achieving near-perfect disease classification through systematic optimization, with enhanced visibility featuring bold legends, increased line thickness, and improved contrast for clear visualization of individual disease class performance, reflecting systematic pipeline optimization effectiveness across all evaluation metrics.

TABLE 8. Ablation study results demonstrating individual component contributions to final student model performance. Each row removes one pipeline component while maintaining optimal configurations for all others. Checkmarks (✓) indicate active components, crossmarks (×) indicate baseline/removed components.

Configuration	Augmentation	Hyperparam. Tuning	Balancing	Ensemble	KD	Accuracy (%)	Macro F1 (%)
Full Pipeline	✓	✓	✓	✓	✓	$\mathbf{98.53 \pm 0.09}$	$\mathbf{97.07 \pm 0.08}$
w/o Knowledge Distillation	✓	✓	✓	✓	×	97.71 ± 0.15	96.34 ± 0.12
w/o Ensemble Teacher	✓	✓	✓	×	✓	97.94 ± 0.12	96.51 ± 0.11
w/o Class Balancing	✓	✓	×	✓	✓	97.12 ± 0.11	94.90 ± 0.12
w/o Hyperparameters	✓	×	✓	✓	✓	96.33 ± 0.60	93.39 ± 0.70
w/o Augmentation	×	✓	✓	✓	✓	94.85 ± 0.22	91.47 ± 0.19
Baseline (All Disabled)	×	×	×	×	×	92.92 ± 0.18	90.12 ± 0.15

classification boundaries between morphologically similar diseases. The enhanced ROC visualizations now feature improved readability with bold legends, increased line thickness (lw = 4), and enhanced contrast (alpha = 0.8) for clear interpretation of individual disease class performance, with the most substantial gains occurring in previously challenging fungal disease categories.

H. ABLATION STUDIES

Systematic ablation analysis quantifies the individual contribution of each pipeline component to final student model performance through exhaustive component removal experiments, as presented in Table 8. Data augmentation emerges as the most critical component, demonstrating the fundamental importance of data diversity for parameter-efficient architectures that lack the representational capacity of larger models, with this substantial impact reflecting ShuffleNetV2's limited feature learning capability requiring extensive data variation to develop robust disease discrimination patterns across the 15-class unified taxonomy. Hyperparameter optimization provides the second-largest contribution, confirming that lightweight architectures exhibit heightened sensitivity to optimization configurations due to their constrained parameter budgets that demand precise learning rate, scheduler, and regularization balance, while class balancing demonstrates significant impact on macro F1-score despite moderate accuracy degradation, reflecting its critical role in minority disease class recognition essential for agricultural disease detection. Knowledge distillation and ensemble teacher removal exhibit comparable but smaller impacts, indicating that while these advanced techniques provide valuable performance enhancement, they offer diminishing returns beyond fundamental optimization components, with the ensemble teacher's limited additional benefit beyond individual models suggesting that the strongest individual teacher (EfficientNet-B4) captures most discriminative

TABLE 9. Detailed computational analysis comparing model architectures across efficiency metrics and deployment characteristics. Compression ratios and speedup factors are calculated relative to the ensemble baseline, demonstrating substantial efficiency gains through knowledge distillation and quantization.

Model	Parameters (M)	Disk Size (MB)	FLOPs (G)	Inference (ms)	Throughput (fps)	Accuracy (%)	Macro F1 (%)	Compression Ratio	Speedup Factor
Teacher Models									
DenseNet-121	7.97	27.16	2.90	0.76	1313.6	96.94	95.53	36.1×	16.6×
ResNet-101	44.5	162.84	7.86	0.55	1805.5	95.94	93.41	6.0×	23.0×
DenseNet-201	20.0	70.43	4.39	1.21	827.0	96.77	94.44	13.9×	10.4×
EfficientNet-B4	19.3	67.76	1.58	0.86	1165.7	98.35	97.35	14.5×	14.7×
4-Model Ensemble	91.77	979.61	16.73	12.64	79.1	99.15	97.58	1.0×	1.0×
Student Models									
ShuffleNetV2 (FP32)	1.36	6.00	0.044	0.29	3511.4	98.53	97.07	163×	43.6×
ShuffleNetV2 (INT8)	1.36	1.46	0.044	0.29	3511.4	97.46	95.83	671×	43.6×

TABLE 10. Cross-dataset generalization results on held-out test partitions. Performance evaluation across controlled laboratory (PlantVillage - PV) and realistic field (TomatoVillage - TV) conditions demonstrates model robustness and deployment viability across diverse agricultural imaging scenarios.

Model Config.	Data-set	Acc. (%)	Macro F1 (%)	Domain Gap
Ensemble Teacher Models				
4-Model Ensemble	PV	99.93 ± 0.03	99.95 ± 0.02	–
4-Model Ensemble	TV	96.10 ± 0.12	95.87 ± 0.14	3.83 pts
Student Models (FP32)				
ShuffleNetV2 (Opt.)	PV	99.89 ± 0.04	90.83 ± 0.18	–
ShuffleNetV2 (Opt.)	TV	96.53 ± 0.11	96.49 ± 0.13	3.36 pts
Student Models (INT8)				
ShuffleNetV2 (Quant.)	PV	98.76 ± 0.06	89.45 ± 0.21	–
ShuffleNetV2 (Quant.)	TV	95.31 ± 0.14	95.18 ± 0.16	3.45 pts

capabilities without requiring complex multi-model aggregation, establishing a clear hierarchy where data augmentation and hyperparameter optimization constitute essential elements for lightweight architecture success.

I. QUANTIZATION AND COMPUTATIONAL ANALYSIS

Post-training quantization through ONNX framework enables deployment-ready compression for agricultural environments, as analyzed in Table 9. The computational analysis reveals critical efficiency trade-offs across the complete pipeline from teacher models through quantized deployment. Teacher model evaluation demonstrates substantial computational variation: ResNet-101 exhibits highest parameter count (44.5M) and computational load (7.86 GFLOPs) achieving 23.0× speedup relative to ensemble baseline, DenseNet architectures provide superior efficiency through dense connectivity achieving 36.1× compression for DenseNet-121, while EfficientNet-B4 establishes optimal performance-efficiency balance (98.35% accuracy, 97.35% F1) with moderate requirements (1.58 GFLOPs, 14.7× speedup). The ensemble achieves superior performance (99.15% accuracy, 97.58% F1) but requires prohibitive resources (979.61 MB storage, 16.73 GFLOPs, 12.64ms inference), necessitating knowledge distillation for mobile deployment. ShuffleNetV2 distillation achieves remarkable efficiency through ensemble knowledge transfer while maintaining diagnostic reliability across the imbalanced 15-class taxonomy. The FP32 student achieves 163× compression with 43.6× speedup, demonstrating exceptional efficiency (98.53% accuracy, 97.07% F1) using 6.00 MB storage and 0.29ms inference, while INT8 quantization enhances compression to 671× preserving minority class recognition (97.46% accuracy, 95.83% F1) with identical inference performance and 1.46 MB footprint. The deployment metrics demonstrate practical viability: 0.29ms inference latency enables real-time field diagnosis, 1.46 MB model size facilitates offline mobile deployment, and 43.6× speedup supports continuous monitoring applications, establishing the quantized ShuffleNetV2 as optimal for mobile agricultural disease detection across diverse hardware constraints where computational resources, connectivity, and storage vary significantly.

J. CROSS-DATASET EXPERIMENTS

Cross-domain generalization evaluation employed the reserved 50% holdout test partitions from both PlantVillage and TomatoVillage datasets to assess model robustness across diverse imaging conditions and domain characteristics, with Table 10 presenting extensive cross-domain evaluation results across ensemble teacher, optimized student, and quantized deployment configurations. Ensemble teacher models demonstrate exceptional performance on PlantVillage's controlled conditions (99.93% accuracy, 99.95% macro F1) while maintaining robust transferability to TomatoVillage field scenarios (96.10% accuracy, 95.87% macro F1), with 3.83 percentage point accuracy degradation reflecting the inherent complexity transition from standardized laboratory imaging to natural agricultural environments, while student model cross-domain analysis demonstrates architectural efficiency trade-offs with the optimized FP32 student achieving near-ensemble performance on PlantVillage (99.89% accuracy) and surprisingly superior macro F1 performance on TomatoVillage (96.49% vs 95.87%), indicating effective knowledge distillation

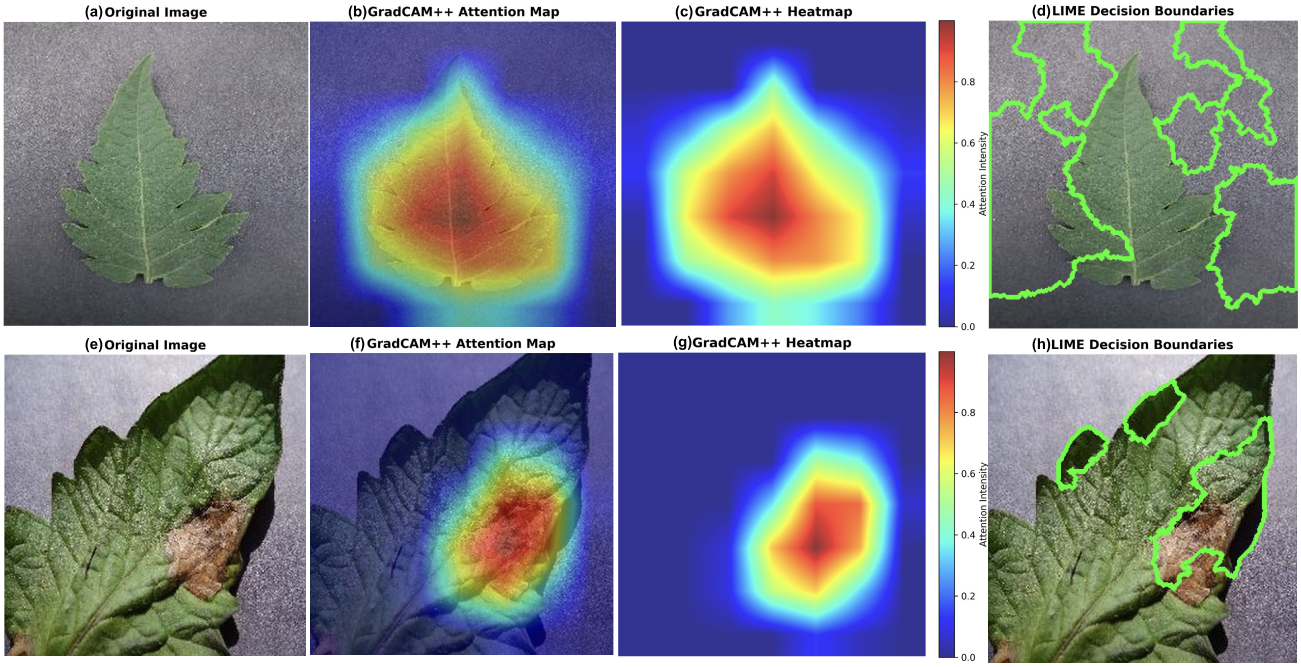

FIGURE 7. Explainable AI analysis demonstrating model attention mechanisms and decision boundaries for tomato leaf disease classification. Top row presents healthy leaf analysis: (a) original healthy sample with characteristic uniform green coloration and intact venation patterns, (b) Grad-CAM++ attention map revealing focused activation on central leaf structure, (c) Grad-CAM++ heatmap quantifying spatial attention intensity distribution, and (d) LIME decision boundaries highlighting superpixels contributing to healthy classification. Bottom row demonstrates diseased leaf analysis: (e) original sample exhibiting characteristic brown necrotic lesions indicative of pathological processes, (f) Grad-CAM++ attention map concentrating on diseased regions, (g) Grad-CAM++ heatmap showing intense activation over pathological areas, and (h) LIME decision boundaries delineating diseased tissue regions. The visualization demonstrates model capability to distinguish healthy tissue characteristics from pathological manifestations through biologically relevant attention mechanisms.

that preserves minority class recognition capabilities. Post-training quantization introduces minimal additional domain gap (3.45 vs 3.36 percentage points), with the quantized model's preserved cross-domain performance (95.31% TomatoVillage accuracy, 95.18% macro F1) establishing deployment viability for resource-constrained agricultural environments while maintaining diagnostic reliability across the complete 15-disease taxonomy, validating the systematic optimization pipeline for practical field deployment scenarios where storage efficiency and offline operation capabilities are paramount.

K. MODEL INTERPRETATION

Systematic model interpretability evaluation employs complementary Grad-CAM++ and LIME techniques to assess model decision-making mechanisms and validate diagnostic reasoning, as demonstrated in Figure 7. The analysis reveals that our optimized student model successfully learns to distinguish between healthy and diseased leaf characteristics through biologically relevant attention patterns, providing essential transparency for agricultural deployment applications.

The Grad-CAM++ analysis demonstrates that the model develops coherent attention mechanisms that align with visual disease indicators. For healthy leaves, the attention maps show concentrated activation on the central leaf structure and main venation patterns, indicating that the model learns to recognize intact cellular architecture and uniform green coloration as indicators of health. For diseased leaves, the attention maps reveal precise localization on the brown necrotic lesion, with peak activation intensities focused on the lesion boundaries and surrounding affected tissue. This spatial coherence validates that the model's decision-making process is grounded in observable pathological features rather than arbitrary image patterns, providing confidence that the diagnostic reasoning follows biologically meaningful principles. The LIME superpixel analysis provides complementary insights by highlighting the specific regions that contribute most strongly to the classification decision, with boundaries encompassing broad areas of uniform green tissue for healthy leaves and tightly delineating pathological regions for diseased leaves. This dual approach confirms that both spatial attention and feature importance analyses converge on biologically relevant regions, establishing robust interpretability that agricultural practitioners can understand and trust.

The interpretability analysis validates that our knowledge distillation and quantization pipeline successfully preserves the model's ability to provide meaningful explanations despite substantial compression (671× parameter reduction). The attention patterns remain spatially coherent and biologically relevant, demonstrating that the optimized student model maintains both classification accuracy and interpretable decision-making capabilities. The visualization

TABLE 11. Comprehensive comparison of our systematic pipeline with state-of-the-art tomato leaf disease detection approaches across eight critical deployment dimensions. Our integrated optimization framework demonstrates superior cross-dataset generalization, exceptional compression efficiency, and comprehensive explainability while maintaining competitive accuracy. Compression ratios and speedup factors are reported as in original studies, with unified benchmarking revealing substantial gaps in existing approaches regarding deployment readiness, cross-domain validation, and systematic optimization integration.

Study	Architecture	Compression	Data Enhancement	XAI	Cross-Dataset	Accuracy	Comp. Ratio	Speedup
Mohanty *et al.* '16 [3]	GoogLeNet	×	Basic	×	×	99.35%	1.0×	1.0×
Wang *et al.* '21 [24]	VGGNet/AlexNet	Multi-technique	×	×	×	97.09%	1196×	284×
Ahmed *et al.* '22 [15]	MobileNetV2	×	CLAHE+Aug	×	×	99.30%	1.0×	1.0×
Deshpande *et al.* '22 [17]	Parallel DCNN	×	GAN Synthesis	×	×	99.14%	1.0×	1.0×
Ojo *et al.* '23 [18]	ResNet-50	×	CLAHE+GAN	×	×	97.69%	1.0×	1.0×
Liu *et al.* '23 [19]	YOLOX-MobileNetV3	×	CycleGAN+CBAM	×	×	94.60%	2.8×	2.0×
Xu *et al.* '23 [30]	CNNA	Pruning	×	×	×	98.96%	75×	18×
Khan *et al.* '24 [2]	CNN-Stacking	×	Bayesian Opt.	×	×	98.27%	1.0×	1.0×
Imam *et al.* '24 [9]	MobileNet-SVM	×	Transfer Learning	×	×	99.37%	1.0×	1.0×
Vini *et al.* '24 [34]	TrioConvTomatoNet	×	✓	×	×	99.40%	1.0×	1.0×
AM *et al.* '24 [10]	VGG16+NASNet	×	EMA Fusion	×	×	98.70%	1.0×	1.0×
Thangaraj *et al.* '24 [6]	Modified Xception	×	Multi-level Fusion	×	×	99.61%	1.0×	1.0×
Shanthi *et al.* '24 [66]	Custom CNN	×	Basic	×	×	95.40%	1.0×	1.0×
Karande *et al.* '24 [16]	11-layer CNN	✓	GrabCut+Aug	×	×	97.10%	1.0×	1.0×
Rakib *et al.* '24 [32]	Quantized CNN	Quantization	×	×	×	98.00%	1.0×	1.0×
Liu *et al.* '24 [25]	EfficientNet	Distillation	Multi-task	×	×	95.80%	10.6×	1.0×
Zhang *et al.* '24 [29]	PDLM-TK	Distillation	×	×	×	96.19%	1.0×	1.0×
Chen *et al.* '24 [20]	CNN-Transformer	×	CycleGAN	×	×	99.45%	1.0×	1.0×
Sharma *et al.* '25 [4]	ResNet50+MobileNetV2	×	Feature Concat	×	×	99.91%	1.0×	1.0×
Ghosh *et al.* '25 [7]	Hybrid ViT	×	Multi-architecture	Grad-CAM+LIME	×	98.00%	1.0×	1.0×
Sun *et al.* '25 [5]	E-TomatoDet	×	CSWinTransformer	Attention	×	97.20%	1.0×	1.0×
Jian *et al.* '25 [28]	DGP-SNNet	Multi-technique	Progressive Transfer	×	×	97.30%	1.6×	1.6×
Chelladurai *et al.* '25 [33]	T-LSTM	×	U-Net Segmentation	Attention	×	99.98%	1.0×	1.0×
Das *et al.* '25 [8]	XLTLDisNet	✓	Strong Aug	Grad-CAM+LIME	×	97.24%	1.0×	1.0×
Sun *et al.* '25 [26]	Multi-architecture	×	Domain Analysis	×	✓	96.10%	1.0×	1.0×
Ni *et al.* '25 [27]	ShuffleNetV2	Self-Distillation	Ensemble Features	×	×	95.08%	1.7×	1.7×
Xiang *et al.* '25 [31]	DWTFormer	×	Frequency-Spatial	Cross-attention	×	99.28%	1.0×	1.0×
Our Ensemble	**4-Model Ensemble**	×	**Comprehensive**	**Grad-CAM++/LIME**	✓	**99.15%**	**1.0×**	**1.0×**
Our Student (FP32)	**ShuffleNetV2 + KD**	**Distillation**	**Comprehensive**	**Grad-CAM++/LIME**	✓	**98.53%**	**163×**	**43.6×**
Our Student (INT8)	**ShuffleNetV2 + KD + Quant**	**Multi-technique**	**Comprehensive**	**Grad-CAM++/LIME**	✓	**97.46%**	**671×**	**43.6×**

demonstrates that the model can effectively communicate its reasoning process to agricultural practitioners, showing exactly which regions influenced the disease classification decision. The interpretability framework thus serves as a critical bridge between deep learning complexity and practical agricultural needs, ensuring that the deployment-ready model provides both accurate diagnoses and understandable explanations.

L. COMPARISON WITH STATE-OF-THE-ART

Table 11 presents comparative analysis across eight critical deployment dimensions: architectural sophistication, compression integration, data enhancement, explainability, cross-dataset validation, accuracy, compression ratios, and speedup factors. The analysis reveals substantial gaps in existing approaches regarding systematic optimization integration and deployment readiness.

1) ACCURACY AND PERFORMANCE LANDSCAPE

Contemporary high-performance approaches demonstrate exceptional controlled-environment capabilities: Thangaraj et al. [6] achieving 99.61% accuracy, Chelladurai et al. [33] reaching 99.98% accuracy, Sharma et al. [4] attaining 99.91% accuracy, and Vini et al. [34] achieving 99.4% accuracy. However, these approaches uniformly lack deployment optimization with no compression benefits or inference acceleration (1.0× baseline performance), severely limiting practical agricultural implementation. Recent lightweight approaches including Das et al. [8] XLTLDisNet (97.24% accuracy) and Ni et al. [27] self-distillation (95.08% accuracy) demonstrate growing interest in efficiency but achieve limited compression ratios (1.0× and 1.7× respectively) compared to our systematic pipeline's 671× compression while maintaining superior accuracy (97.46%).

2) COMPRESSION AND EFFICIENCY ANALYSIS

Existing compression methodologies demonstrate limited systematic integration compared to our comprehensive approach. Wang et al. [24] achieved remarkable compression ratios (1,196× for VGGNet, 751× for AlexNet) but suffered substantial accuracy degradation (97.09% and 87.51% respectively). Xu et al. [30] demonstrated 75× parameter reduction and 18× computational efficiency while maintaining 98.96% accuracy. However, most approaches operate in isolation: Liu et al. [25] achieved 10.6× compression through distillation alone, Ni et al. [27] demonstrated 1.7× efficiency gains, and Jian et al. [28] achieved 1.6× improvements. Our systematic pipeline achieves exceptional efficiency gains (671× compression ratio, 43.6× speedup) through integrated optimization encompassing ensemble knowledge distillation and quantization, vastly outperforming existing approaches while maintaining competitive accuracy (97.46%). Deployment-specific analysis reveals our approach's superior practical characteristics: 0.29ms inference latency enabling real-time field diagnosis, 1.46 MB model size suitable for offline mobile deployment, and 43.6× speedup facilitating continuous monitoring applications,

addressing critical agricultural deployment constraints where existing approaches require cloud connectivity or substantial computational infrastructure.

3) DATA ENHANCEMENT AND CROSS-DATASET VALIDATION GAPS

Advanced data enhancement strategies remain fragmented across existing approaches, with most studies focusing on single-technique optimization. Ahmed et al. [15] demonstrated efficient augmentation achieving 99.30% accuracy with 9.60 MB models, Chen et al. [20] employed cycle-consistent GAN augmentation achieving 99.45% accuracy, and Xiang et al. [31] utilized frequency-spatial fusion achieving 99.28% accuracy. However, systematic class balancing through advanced sampling techniques (SMOTE, ADASYN) remains unexplored, with most approaches relying on basic augmentation or GAN synthesis without addressing severe class imbalance that characterizes agricultural datasets. Critically, cross-dataset validation represents the most significant gap in existing literature, with only Sun et al. [26] conducting domain analysis across laboratory and field conditions. Our work represents the first systematic cross-dataset evaluation demonstrating consistent 3.45 percentage point domain gap while maintaining deployment-ready performance, establishing a comprehensive benchmark for real-world agricultural deployment scenarios where domain generalization is essential for practical utility.

4) EXPLAINABILITY AND DEPLOYMENT INTEGRATION

Explainable AI integration remains limited in existing approaches, with only Ghosh et al. [7] and Das et al. [8] incorporating Grad-CAM and LIME visualization. Most studies lack comprehensive explainability frameworks essential for agricultural practitioner acceptance and regulatory compliance. Our systematic pipeline integrates Grad-CAM++ and LIME explainability throughout the optimization process, ensuring diagnostic transparency preservation from ensemble teacher through quantized student deployment, with biologically relevant attention patterns validated against plant pathology principles and practical interpretation capabilities for farmer decision-making.

5) SYSTEMATIC INTEGRATION AND DEPLOYMENT READINESS

The most significant limitation of existing approaches lies in their isolated optimization strategies, where compression, data enhancement, and explainability are addressed independently rather than through systematic integration. Our comprehensive framework establishes unprecedented integration across all deployment dimensions: ensemble model providing state-of-the-art accuracy (99.15%), compressed student model achieving exceptional efficiency gains (671× compression, 43.6× speedup) vastly outperforming existing approaches, rigorous cross-dataset validation demonstrating robust generalization, and deployment-ready edge performance with 0.29ms inference latency. This work bridges the critical gap between laboratory research and practical field deployment, establishing a comprehensive benchmark for agricultural AI systems with systematic optimization integration previously unavailable in the literature, enabling scalable deployment across diverse global agricultural contexts where computational resources, connectivity, and expertise vary significantly.

V. DISCUSSIONS

A. DEPLOYED APPLICATION

The Flutter-based mobile application validates practical deployment through systematic integration of the quantized model within a professional agricultural interface, as demonstrated in Figure 8. The architecture leverages Flutter's single codebase paradigm for native Android and iOS performance while maintaining consistent user experience across diverse practitioner technology preferences. While the current implementation prioritizes single-image diagnosis for real-time field assessment, the architecture is extensible to batch inference functionality for mass-data processing. This feature will enable agricultural stakeholders to upload multiple leaf samples simultaneously for parallel diagnosis, enhancing throughput in large-scale deployments such as cooperative farming systems or crop monitoring centers. Future updates to the application are planned to incorporate batch processing modules, leveraging mobile device concurrency and server offloading to support scalable agricultural diagnostics.

The application implements the complete diagnostic pipeline: image acquisition through camera integration and gallery selection, real-time inference using the 1.46 MB quantized ShuffleNetV2 achieving 43.6× speedup, results visualization with confidence analysis and probability breakdowns across the 15-disease taxonomy, and detailed disease information with pathological descriptions, treatment protocols, and expert guidance. Critical features include multilingual support across six languages (English, Spanish, Hindi, Chinese, Arabic, Bengali) with complete localization, offline operation for remote environments, and professional interface design with severity-based color coding and timestamp metadata for diagnostic traceability, achieving 0.29ms inference latency while maintaining 97.46% accuracy across cross-domain scenarios, establishing deployment readiness for precision agriculture applications.

B. MAIN FINDINGS AND SIGNIFICANCE

This work establishes the first systematic pipeline integrating ensemble learning, knowledge distillation, and quantization for deployable cross-domain tomato disease detection, achieving state-of-the-art performance while addressing practical deployment constraints. The unified benchmarking framework combining PlantVillage and TomatoVillage datasets into 15 harmonized disease classes represents the first standardized cross-domain evaluation protocol, enabling systematic assessment of model generalization across laboratory-controlled and field imaging conditions.

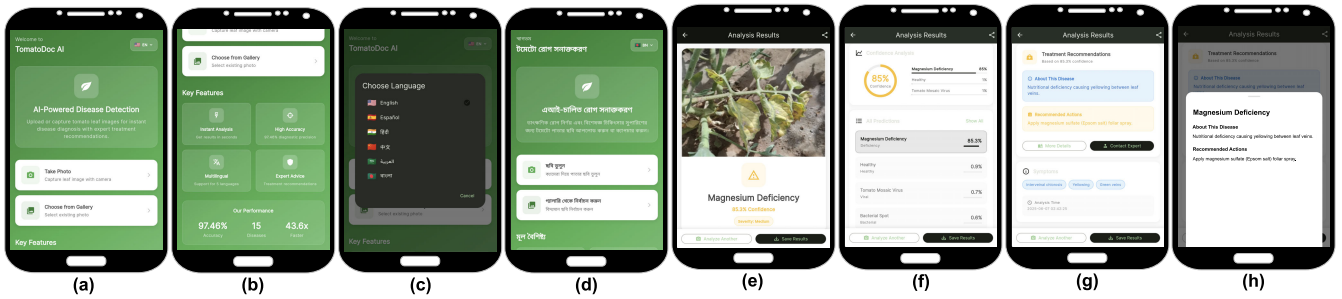

FIGURE 8. Mobile application deployment demonstrating cross-platform integration of the optimized tomato disease detection system. (a) agricultural-themed home interface featuring "Take Photo" and "Choose from gallery" options, (b) key features presentation highlighting instant analysis, high accuracy, multilingual support, and expert advice capabilities, and performance metrics (97.46% accuracy, 15 disease classes, 43.6× speedup) of the deployed model, (c) language selection interface enabling real-time switching across six languages (English, Spanish, Hindi, Chinese, Arabic, Bengali) with cultural adaptation, (d) Bengali localization demonstration showing complete interface translation and cultural appropriateness. (e) analysis results screen displaying uploaded tomato leaf image with Magnesium Deficiency detection at 85.3% confidence and severity classification (f) confidence analysis visualization featuring circular progress indicators and prediction probability breakdown across disease taxonomy, (g) treatment recommendations (about disease, actions, symptoms, expert contact) with analysis time and saving options, (h) detailed disease information screen accessed via "More details" button.

Systematic evaluation across 24 architectures reveals ensemble superiority over individual models, with the four-model soft voting configuration achieving exceptional performance (99.15% accuracy, 97.58% macro F1-score) through complementary architectural strengths. Knowledge distillation successfully transfers ensemble capabilities to compact ShuffleNetV2, achieving 163× parameter reduction and 43.6× inference speedup with optimal configurations ($T = 15, \alpha = 0.7$), while INT8 quantization achieves unprecedented 671× storage compression (1.46 MB) with minimal degradation. The integrated data enhancement framework demonstrates systematic optimization importance, with ADASYN-based class balancing addressing severe 75:1 imbalance and strategic augmentation providing 3.27 percentage point macro F1-score improvement. Cross-dataset evaluation validates robust generalization with consistent 3.45 percentage point domain gap, while the deployed Flutter application with multilingual support achieves 0.29ms inference capabilities, establishing systematic optimization integration that bridges the critical gap between laboratory research and practical agricultural deployment.

C. LIMITATIONS

1) TAXONOMIC SCOPE AND GENERALIZATION CONSTRAINTS

The proposed pipeline focuses exclusively on tomato disease classification across 15 unified classes, limiting direct transferability to other crop types or broader agricultural applications. While the systematic optimization methodology provides a replicable framework, the trained models require domain-specific adaptation for deployment across different plant species, disease taxonomies, or agricultural contexts where pathological manifestations exhibit distinct morphological characteristics.

2) DATASET REPRESENTATIVENESS AND FIELD VALIDATION GAPS

Despite combining PlantVillage and TomatoVillage datasets to address domain diversity, the evaluation remains limited by the inherent biases of laboratory-controlled imaging conditions and constrained field data representation. The absence of real-world validation across diverse geographical regions, environmental conditions, and farmer practices introduces uncertainty regarding model performance under authentic deployment scenarios, particularly for rare disease classes and extreme imaging conditions.

3) COMPUTATIONAL OVERHEAD DURING TRAINING AND SCALABILITY CONSTRAINTS

The ensemble teacher model requires substantial computational resources during training (979.61 MB storage, 16.73 GFLOPs), limiting scalability for resource-constrained research environments and continuous model updates. While the distilled student model achieves deployment efficiency, the training pipeline necessitates high-performance infrastructure for ensemble optimization, knowledge distillation, and exhaustive hyperparameter tuning, potentially hindering adoption in developing agricultural contexts where computational resources are limited.

VI. CONCLUSION AND FUTURE WORK

This paper presents the first comprehensive pipeline for deployable cross-domain tomato leaf disease detection that systematically integrates ensemble learning, knowledge distillation, and quantization to achieve practical agricultural deployment while maintaining diagnostic reliability. The unified evaluation framework across PlantVillage and TomatoVillage datasets establishes a robust 15-class benchmark that addresses critical domain gaps between laboratory-controlled and field imaging conditions. Systematic optimization encompassing data augmentation, ADASYN balancing, hyperparameter tuning, and explainable AI integration demonstrates substantial performance improvements, culminating in an ensemble achieving 99.15% accuracy and a quantized student model maintaining 97.46% accuracy with 671× compression ratio and 43.6× inference speedup. Cross-dataset validation

confirms robust generalization with consistent domain gaps, while the deployed Flutter application validates practical implementation across diverse agricultural environments with multilingual support and 0.29ms inference capabilities, establishing both a comprehensive deployment framework and reproducible benchmark for the agricultural AI community.

Future research should prioritize extension to multi-crop disease detection through domain adaptation techniques, development of larger field-collected datasets with geographical and environmental diversity, and integration with IoT agricultural monitoring systems for continuous crop health assessment. Advanced ensemble selection strategies including correlation-based selection, diversity-driven approaches, and automated optimization frameworks warrant investigation to address scalability limitations of exhaustive combinatorial evaluation for larger model pools. Advanced compression techniques including neural architecture search and progressive knowledge distillation warrant investigation for further efficiency improvements, while real-world deployment studies with agricultural practitioners will provide critical validation of practical utility and user acceptance. The established systematic optimization framework provides a foundation for broader agricultural AI applications, enabling holistic evaluation methodologies that address the complex requirements of precision agriculture deployment across diverse global farming contexts.

REFERENCES

[1] Food and Agriculture Organization of the United Nations (FAO). (2019). *The State of Food and Agriculture 2019: Moving Forward on Food Loss and Waste Reduction*. Rome. Accessed: Jul. 17, 2025. [Online]. Available: https://openknowledge.fao.org/items/ba08937f-4a41-4ff5-a4e7-e495e5f5f599

[2] B. Khan, S. Das, N. S. Fahim, S. Banerjee, S. Khan, M. K. Al-Sadoon, H. S. Al-Otaibi, and A. R. M. T. Islam, "Bayesian optimized multimodal deep hybrid learning approach for tomato leaf disease classification," *Sci. Rep.*, vol. 14, no. 1, p. 21525, Sep. 2024.

[3] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers Plant Sci.*, vol. 7, p. 1419, Sep. 2016.

[4] J. Sharma, A. A. Al-Huqail, A. Almogren, H. Doshi, B. Jayaprakash, B. Bharathi, A. Ur Rehman, and S. Hussen, "Deep learning based ensemble model for accurate tomato leaf disease classification by leveraging ResNet50 and MobileNetV2 architectures," *Sci. Rep.*, vol. 15, no. 1, p. 13904, Apr. 2025.

[5] H. Sun, R. Fu, X. Wang, Y. Wu, M. A. Al-Absi, Z. Cheng, Q. Chen, and Y. Sun, "Efficient deep learning-based tomato leaf disease detection through global and local feature fusion," *BMC Plant Biol.*, vol. 25, no. 1, p. 311, Mar. 2025.

[6] R. Thangaraj, P. Pandiyan, S. Anandamurugan, and S. Rajendar, "A deep convolution neural network model based on feature concatenation approach for classification of tomato leaf disease," *Multimedia Tools Appl.*, vol. 83, no. 7, pp. 18803–18827, Aug. 2023.

[7] H. Ghosh, I. S. Rahat, M. M. R. Emon, M. J. Mashrafi, M. A. Al A. Tanzin, S. N. Mohanty, and S. Kant, "Advanced neural network architectures for tomato leaf disease diagnosis in precision agriculture," *Discover Sustainability*, vol. 6, no. 1, p. 312, Apr. 2025.

[8] A. Das, F. Pathan, J. R. Jim, M. R. Ouishy, M. M. Kabir, and M. F. Mridha, "XLTLDisNet: A novel and lightweight approach to identify tomato leaf diseases with transparency," *Heliyon*, vol. 11, no. 4, Feb. 2025, Art. no. e42575.

[9] M. H. Imam, N. Nahar, R. Bhowmik, S. B. S. Omit, T. Mahmud, M. S. Hossain, and K. Andersson, "A transfer learning-based framework: MobileNet-SVM for efficient tomato leaf disease classification," in *Proc. 6th Int. Conf. Electr. Eng. Inf. Commun. Technol. (ICEEICT)*, May 2024, pp. 693–698.

[10] P. V., A. M. S. Kumar, J. I. R. Praveen, S. Venkatraman, S. P. Kumar, S. A. Aravintakshan, A. Abeshek, and A. Kannan, "Improved tomato leaf disease classification through adaptive ensemble models with exponential moving average fusion and enhanced weighted gradient optimization," *Frontiers Plant Sci.*, vol. 15, May 2024, Art. no. 1382416.

[11] J. Sun, "Addressing domain shift in deep learning: Challenges and insights from plant disease diagnosis and flower recognition," *bioRxiv*, Oct. 2024, doi: 10.1101/2024.10.07.617111.

[12] 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 CoDS 25th (COMAD)*, Jan. 2020, pp. 249–253.

[13] E. Moupojou, A. Tagne, F. Retraint, A. Tadonkemwa, D. Wilfried, H. Tapamo, and M. Nkenlifack, "FieldPlant: A dataset of field plant images for plant disease detection and classification with deep learning," *IEEE Access*, vol. 11, pp. 35398–35410, 2023.

[14] M. Gehlot, R. K. Saxena, and G. C. Gandhi, "'Tomato-village': A dataset for end-to-end tomato disease detection in a real-world environment," *Multimedia Syst.*, vol. 29, no. 6, pp. 3305–3328, Dec. 2023.

[15] S. Ahmed, M. B. Hasan, T. Ahmed, M. R. K. Sony, and M. H. Kabir, "Less is more: Lighter and faster deep neural architecture for tomato leaf disease classification," *IEEE Access*, vol. 10, pp. 68868–68884, 2022.

[16] S. Karande and B. Garg, "Performance evaluation and optimization of convolutional neural network architectures for tomato plant disease eleven classes based on augmented leaf images dataset," *Neural Comput. Appl.*, vol. 36, no. 20, pp. 11919–11943, Jul. 2024.

[17] R. Deshpande and H. Patidar, "Tomato plant leaf disease detection using generative adversarial network and deep convolutional neural network," *Imag. Sci. J.*, vol. 70, no. 1, pp. 1–9, Jan. 2022.

[18] M. O. Ojo and A. Zahid, "Improving deep learning classifiers performance via preprocessing and class imbalance approaches in a plant disease detection pipeline," *Agronomy*, vol. 13, no. 3, p. 887, Mar. 2023.

[19] W. Liu, Y. Zhai, and Y. Xia, "Tomato leaf disease identification method based on improved YOLOX," *Agronomy*, vol. 13, no. 6, p. 1455, May 2023.

[20] Z. Chen, G. Wang, T. Lv, and X. Zhang, "Using a hybrid convolutional neural network with a transformer model for tomato leaf disease detection," *Agronomy*, vol. 14, no. 4, p. 673, Mar. 2024.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[22] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[24] R. Wang, W. Zhang, J. Ding, M. Xia, M. Wang, Y. Rao, and Z. Jiang, "Deep neural network compression for plant disease recognition," *Symmetry*, vol. 13, no. 10, p. 1769, Sep. 2021.

[25] B. Liu, S. Wei, F. Zhang, N. Guo, H. Fan, and W. Yao, "Tomato leaf disease recognition based on multi-task distillation learning," *Frontiers Plant Sci.*, vol. 14, Jan. 2024, Art. no. 1330527.

[26] C. Sun, Y. Li, Z. Song, Q. Liu, H. Si, Y. Yang, and Q. Cao, "Research on tomato disease image recognition method based on DeiT," *Eur. J. Agronomy*, vol. 162, Jan. 2025, Art. no. 127400.

[27] S. Ni, Y. Jia, M. Zhu, Y. Zhang, W. Wang, S. Liu, and Y. Chen, "An improved ShuffleNetV2 method based on ensemble self-distillation for tomato leaf diseases recognition," *Frontiers Plant Sci.*, vol. 15, Jan. 2025, Art. no. 1521008.

[28] T. Jian, H. Qi, R. Chen, J. Jiang, G. Liang, and X. Luo, "Identification of tomato leaf diseases based on DGP-SNNet," *Crop Protection*, vol. 187, Jan. 2025, Art. no. 106975.

[29] X. Zhang, K. Liang, and Y. Zhang, "Plant pest and disease lightweight identification model by fusing tensor features and knowledge distillation," *Frontiers Plant Sci.*, vol. 15, Nov. 2024, Art. no. 1443815.

[30] Y. Xu, Z. Gao, Y. Zhai, Q. Wang, Z. Gao, Z. Xu, and Y. Zhou, "A CNNA-based lightweight multi-scale tomato pest and disease classification method," *Sustainability*, vol. 15, no. 11, p. 8813, May 2023.
[31] Y. Xiang, S. Gao, X. Li, and S. Li, "DWTFormer: A frequency-spatial features fusion model for tomato leaf disease identification," *Plant Methods*, vol. 21, no. 1, p. 33, Mar. 2025.
[32] A. F. Rakib, R. Rahman, A. A. Razi, and A. S. M. T. Hasan, "A lightweight quantized CNN model for plant disease recognition," *Arabian J. for Sci. Eng.*, vol. 49, no. 3, pp. 4097–4108, Mar. 2024.
[33] A. Chelladurai, D. P. Manoj Kumar, S. S. Askar, and M. Abouhawwash, "Classification of tomato leaf disease using transductive long short-term memory with an attention mechanism," *Frontiers Plant Sci.*, vol. 15, Jan. 2025, Art. no. 1467811.
[34] S. Ledbin Vini and P. Rathika, "TrioConvTomatoNet: A robust CNN architecture for fast and accurate tomato leaf disease classification for real time application," *Scientia Horticulturae*, vol. 330, Apr. 2024, Art. no. 113079.
[35] H. Su and J. Lee, "Machine learning approaches for diagnostics (don't short) and prognostics of industrial systems using open source data from PHM data challenges: A review," *Int. J. Prognostics Health Manage.*, vol. 15, no. 2, Sep. 2024.
[36] M. Aftab, F. Mehmood, C. Zhang, A. Nadeem, Z. Dong, Y. Jiang, and K. Liu, "AI in oncology: Transforming cancer detection through machine learning and deep learning applications," 2025, *arXiv:2501.15489*.
[37] R. Alkhanbouli, H. M. A. Almadhaani, F. Alhosani, and M. C. E. Simsekler, "The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions," *BMC Med. Informat. Decis. Making*, vol. 25, no. 1, p. 110, Mar. 2025.
[38] B. P. Cabral, L. A. M. Braga, C. G. C. Filho, B. Penteado, S. L. F. de Castro Silva, L. Castro, M. Fornazin, and F. Mota, "Future use of AI in diagnostic medicine: 2-Wave cross-sectional survey study," *J. Med. Internet Res.*, vol. 27, Feb. 2025, Art. no. e53892.
[39] J. Bianco, D. N. Kanakis, A. Dasgupta, M. Moonis, and C. Zoia, "Editorial: Recent advances in diagnosis and treatment of brain tumors: From pediatrics to adults," *Frontiers Neurol.*, vol. 16, May 2025, Art. no. 1606149.
[40] V. Solapure, S. Dy, and A. Jawale, "Tomato leaf disease dataset," Mendeley Data, Version 1, Aug. 2024, doi: 10.17632/zfv4jj7855.1.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[45] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[46] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[47] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.
[48] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet v2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 122–138.
[49] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[50] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 mb model size," 2016, *arXiv:1602.07360*.
[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[52] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jeǵou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[53] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[54] H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "EfficientViT: Multi-scale linear attention for high-resolution dense prediction," 2022, *arXiv:2205.14756*.
[55] Y. Li, G. Yuan, Y. Wen, E. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, "EfficientFormer: Vision transformers at MobileNet speed," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 12263–12277.
[56] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.
[57] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
[58] A. Paszke, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
[59] S. Marcel and Y. Rodriguez. (2010). *Torchvision: Pytorch's Computer Vision Library*. [Online]. Available: https://github.com/pytorch/vision
[60] R. Wightman. (2019). *Pytorch Image Models*. [Online]. Available: https://github.com/rwightman/pytorch-image-models
[61] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 559–563, 2016.
[62] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.
[63] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, Aug. 1999.
[64] Y. Bian and H. Chen, "When does diversity help generalization in classification ensembles?" *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9059–9075, Sep. 2022.
[65] A. Jeffares, T. Liu, J. Crabbé, and M. van der Schaar, "Joint training of deep ensembles fails due to learner collusion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 13559–13589.
[66] S. D L, V. K, A. N, and S. Vashistha, "Tomato leaf disease detection using CNN," *Proc. Comput. Sci.*, vol. 235, pp. 2975–2984, Jan. 2024.

MOHAMMAD JUNAYED HASAN received the B.S. degree (summa cum laude) in computer science and engineering from North South University (NSU), Bangladesh, in 2023. He is currently pursuing the M.S.E. degree in computer science with Johns Hopkins University, USA. During his undergraduate studies, he was a Research Assistant and Teaching Assistant with the Department of Electrical and Computer Engineering. He is a Graduate Teaching Assistant in big data machine learning with Johns Hopkins University. His thesis is conducted under the supervision of Prof. Philipp Koehn and Adv. Prof. Anjalie Field with the Center for Language and Speech Processing (CLSP). He is a Summer Data Science Intern with Mayo Clinic, focusing on clinical diagnostics and time-series analysis for improving clinical forecasting. His research contributions include peer-reviewed publications in leading venues, such as *Pattern Recognition Letters*, *International Journal of Medical Informatics*, *PLOS One*, and *Heliyon*. His research interests include encompass computer vision, natural language processing, machine learning, and quantum computing, with particular emphasis on applications in real-world problems.

SUVODEEP MAZUMDAR is currently a Senior Lecturer in data analytics. His research explores developing techniques and mechanisms for reducing the barriers that impede user communities understanding of vast complex multidimensional datasets. He conducts interdisciplinary research on highly engaging, interactive, and visual mechanisms in conjunction with complex querying techniques for seamless navigation, exploration, and understanding of complex datasets. He has applied his research in a wide range of application domains, such as aerospace engineering, sports informatics, crisis/emergency management, smart cities, and mobility planning. As a part of his research, he collaborates with large multi-disciplinary teams of academics, industry partners, city councils, and planners. He has worked in several extensive research and industrial projects funded by the UKRI, European Union, Innovate U.K., and European Space Agency. His research interests include studying and developing data and visual analytic techniques to analyze massive volumes of dynamic data in near real-time, citizen science and crowdsourcing techniques for observing physical phenomena, events, and environments, user interface development, human–computer interaction, and user-centered design, and assistive technologies to support independent activities of daily living.

SIFAT MOMEN received the Ph.D. degree from The University of Sheffield, U.K., in 2011. In September 2017, he joined as an Assistant Professor with the Department of Electrical and Computer Engineering (ECE), North South University (NSU). He is currently a Professor with the Department of Electrical and Computer Engineering (ECE). Prior to joining NSU, he worked as an Assistant Professor for nearly six years with the Department of Computer Science and Engineering, the University of Liberal Arts Bangladesh (ULAB). He was the acting Head of the Department, ULAB, for some time. His Ph.D. research was at the crossroads of biology and engineering, and he was heavily influenced by the behavior of eusocial insects, which are well-known for their self-organizing abilities. His current research interests include complex systems, machine learning, information systems, and modeling and simulation of natural systems. He is an active researcher and regularly reviews numerous conference papers and journal articles.