LiteViT: Leveraging the Power of Transformers for Edge AI in Crop Disease Classification

Sai Mahesh Mudavat
Department of Computer Science and Engineering
University of North Texas
Denton, TX, USA
SaiMaheshMudavat@my.unt.edu

Alakananda Mitra
Nebraska Water Center
University of Nebraska-Lincoln
Lincoln, NE, USA
amitra6@unl.edu

Saraju P. Mohanty
Department of Computer Science and Engineering
University of North Texas
Denton, TX, USA
saraju.mohanty@unt.edu

Elias Kougianos
Department of Electrical Engineering
University of North Texas
Denton, TX, USA
elias.kougianos@unt.edu

***Abstract*—Plant diseases significantly impact crop yield, which presents a serious challenge to food security. Despite advances in Artificial Intelligence (AI) for improved disease detection, real-world implementation remains limited due to high computational demands. LiteViT bridges this gap through a proposed Knowledge distillation framework that transforms powerful but computationally heavy Vision transformers (ViTs) into a field-ready tool by distilling knowledge of a 300 million parameter ViT large teacher into an lightweight MobileViT-XXS (extra-extra-small) model of size 3.8 MB, achieving 99.3% accuracy, while retaining nearly identical performance compared to the teacher's 99.7%. The framework integrates a multimodal explainability framework that visually interprets model predictions to enhance interpretability. This framework demonstrates accurate and explainable plant disease detection suitable for edge devices, bridging the gap between laboratory and field-level deployment, thereby advancing smart agriculture.**

***Index Terms*—Smart Agriculture; Edge Computing; Vision Transformer; Model Explainability; Knowledge distillation.**

I. Introduction

Plant diseases cause 10%-16% annual crop yield losses globally [1]. Addressing this requires automated, accessible tools for early detection, crucial for minimizing crop losses and reducing excessive pesticide usage [2]. Deep learning models offer high classification accuracy [3], with Convolution Neural Networks (CNN) based models like VGG and ResNet achieving over 95% [4] accuracy in plant disease classification. ViTs now outperform them via global self-attention [5], [6]. In parallel, agriculture is shifting towards Smart farming, powered by Internet of Agriculture Things (IoAT) [7], using edge devices which are constrained by memory and processing limits. Lightweight models (e.g, EfficientNet and MobileNet) [8] reduce computation but lack interpretability, hindering user trust.A compelling solution is offered by Knowledge distillation [9], allowing compact student models to perform as well as a higher-performing teacher model by using its softened teacher's output distribution to capture inter-class relationships effectively. While effective in CNNs [10], Knowledge distillation remains underutilized in transformer-based systems, particularly in real-time drone assisted monitoring scenarios where precision and speed are critical [11], [12].

To address this, a cross-architecture knowledge distillation framework is introduced, which transfers disease classification capabilities of a ViT large patch 32 teacher model [13] to a compact, MobileViT-XXS (extra-extra-small), the smallest variant of the MobileViT model [14], designed for edge deployment. To further improve interpretability, a multimodal explainability module is integrated, which includes feature visualization, attention tracking, and activation mapping, enhancing trust in model predictions [15], [16].

The paper is structured as follows: Section II introduces related research in plant disease detection. Section III presents novel contributions of this research. Section IV proposed the working of the model. Section V gives experimental validation of the proposed solution, and final Section VI provides a conclusion and future research directions.

II. Related works

Early breakthrough in plant disease detection leveraged CNN architectures with models like AlexNet, GoogleNet, applied across multiple crops [17]. To improve inference time and accuracy, transfer learning techniques with deeper architectures such as VGG and ResNet were introduced [3], [18]. However, their large parameter size prompted the development of lightweight CNNs like MobileNet and SqueezeNet for resource-constrained environments [2], [8]. While CNN effectively captures local image features, it struggles with global context modeling. Vision Transformer (ViT), driven by a self-attention mechanism, addresses this limitation and has recently been adopted in edge-based plant disease classification [6]. Hybrid CNN-ViT have been introduced, combining local and global feature extraction capability [19]. To further enhance edge efficiency, knowledge distillation has gained momentum. It enables compact student models to inherit a larger teacher model's representational power while maintaining accuracy, making them suited for edge devices [10], [20].

With increasing model complexity, explainability is now vital for trustworthy AI. Techniques such as Grad-CAM and CBAM are employed to visualize model attention, improving interpretability [6], [16]. Table I presents a comprehensive overview of related works on plant disease detection using deep learning methods.

TABLE I: Relevant Works On Crop Disease Detection

Research Work	**Model**	**Observation**
Mohanty et al. [17]	CNN (AlexNet, GoogLeNet)	High computational model; not edge optimized
Rahman et al. [2]	2-Stage Lightweight CNN	shallow model; limited features
Wang et al. [3]	VGG16	Too large (138M);not suitable for edge
Huang et al. [10]	Knowledge Distillation using YOLOR object detection model	Multistage distillation with accuracy drop in student model
Khabarlak et al. [21]	Feature KD using EfficientNetV2 large and MobileNetV3	Small gap of 1.34% between models; extra 4.33M parameters
Parez et al. [11]	GreenViT	Compact model but memory intensive
Mahmud et al. [16]	Grad-CAM(4 models)	Lacks size-accuracy balance, clear interpretability maps.
LiteViT (Current Work)	**Distilled ViT large + MobileViT-XXS**	**Optimized for edge deployment with clear attention maps visualisations**

III. Novel Contributions

This section discusses the problem statement, novelty, and significance of LiteViT in identifying plant diseases.

A. Problem Statement

Plant diseases threaten agricultural productivity, especially under dynamic environmental conditions. Although transformer-based models capture complex patterns well, their computational demands limit edge deployment. This work presents a lightweight and interpretable framework for real-time tomato leaf disease detection. A ViT-Large model transfers semantic features to a compact MobileViT-XXS via knowledge distillation. Teacher and student models are trained on a tomato leaf dataset [22] for efficient on-device agricultural inference.

B. Novelty and Significance of the Proposed Solution

The novel contributions of LiteViT are as follows:

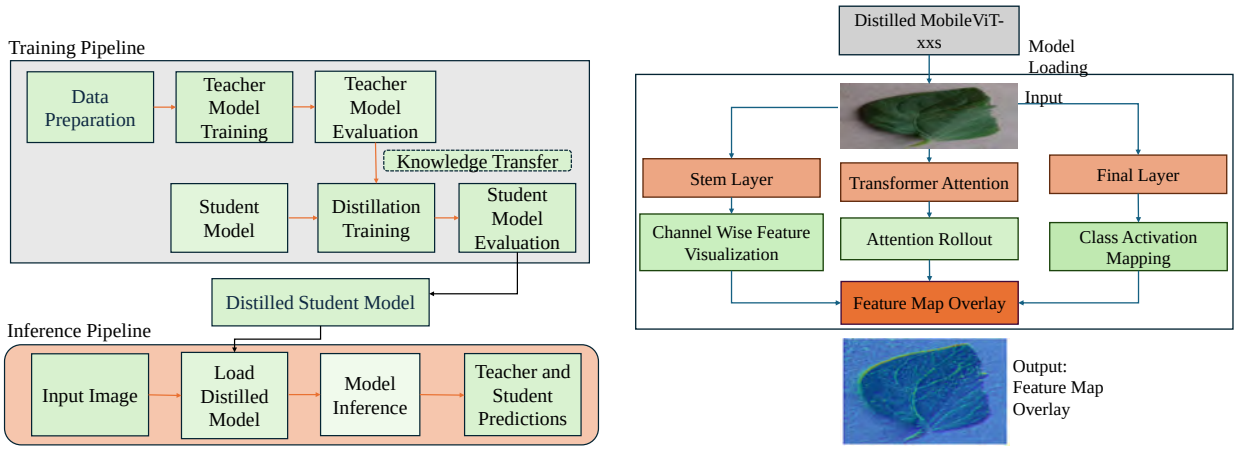
1) A cross-architecture knowledge distillation framework that distills global features from a ViT-Large teacher to a MobileViT-XXS student using a dual-loss objective.
2) A temperature-scaled distillation method that improves the student model's ability to separate visually similar disease classes.
3) Multimodal explainability through channel-wise feature maps, attention rollout, and activation mapping, enabling interpretation of predictions.
4) Edge deployment readiness with compact size of 3.8MB achieved on model with 99.6% reduced parameter count.

IV. Proposed Framework

Accurate and early diagnosis of plant diseases is critical for precision agriculture. Although transformer-based models achieve higher accuracy in image classification, their deployment on the edge remains limited. To address this, we propose a cross-architecture knowledge distillation framework to operate under resource constraints without compromising performance. The approach leverages a high-capacity Vision Transformer (ViT large patch 32) as a teacher model and a lightweight MobileViT-XXS as a student model. The pipeline is divided into two stages: training and inference. Figure 1 illustrates the LiteViT architecture. Figure 1(a) shows the distillation process from teacher to student model. Figure 1(b) depicts the explainability module, highlighting key regions using attention rollout and CAM activation maps.

A. Training Phase

All input tomato-leaf images are resized to 256 × 256 pixels, normalized by dataset mean and standard deviation, and augmented with random horizontal flips to simulate real-world variability. The training phase involves two stages: [1] the ViT-Large teacher (patch size 32) is trained to convergence on the 80 % training split using cross-entropy loss; [2] the MobileViT-XXS student is distilled from the teacher by minimizing a



(a) Knowledge Distillation from ViT-Large to MobileViT-XXS

(b) Explainability Module Highlighting Disease Regions via Attention and CAM.

Fig. 1: LiteViT Framework: Cross-Architecture Distillation and Explainability Pipeline for Plant Disease Detection

joint loss of cross-entropy on ground-truth labels and Kullback–Leibler divergence to the teacher's softened outputs. The remaining 20% of the data is evenly split for validation and testing. The teacher model ViT large patch32 with 300 million parameters is fine-tuned on the dataset. Each image is divided into non-overlapping 32×32 pixel patches, embedded into 1024-dimensional tokens, and passed through 24 transformer blocks with 16 attention heads per block. The model optimization uses a learning rate of 0.0001, cross-entropy loss, and a batch size 32 for five epochs, with accuracy and loss metrics tracked throughout.The teacher model reached stable accuracy within five epochs.

Once the teacher converges, then the student model gets trained. This compact model combines CNN layers for local feature extraction with transformer layers for global context modeling using depthwise separable convolutions and inverted bottleneck layers for computational efficiency. Training was performed in mini-batches for both models during each iteration. The teacher outputs logits z_t using its multi-head self-attention mechanism, while the student produces logits z_s through its compact architecture. The student model is optimized via a dual loss strategy, where hard loss is performed with cross-entropy loss with ground truth labels of disease, and soft loss with KL (Kullback-Leibler) divergence between softened teacher output and student output for dual loss optimization. While the teacher remains in evaluation mode, the student model updates its parameters through backpropagation with the Adam optimizer. The total distillation loss is calculated as shown in equation 1.

$$\mathcal{L}_{\text{distill}} = \alpha \cdot \mathcal{L}_{\text{CE}}(y, z_s) + (1-\alpha) \cdot \tau^2 \cdot \text{KL}\left(\sigma\left(\frac{z_s}{\tau}\right), \sigma\left(\frac{z_t}{\tau}\right)\right) \tag{1}$$

Here, $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss computed between the student model's output and the true labels. In contrast, KL refers to the KL divergence between the softened output distributions of the student and teacher models. The softmax function σ is applied to the logits z_s student and z_t teacher, with temperature parameter τ used to soften these outputs, capturing subtle inter-class relationships. The coefficient α balances the contribution of soft and hard losses. The student model is also trained for five epochs, during which it steadily converges to optimal accuracy. The detailed step-by-step procedure of the training process is outlined in Algorithm 1.

B. Inference Phase

The inference pipeline, as explained in Algorithm 2, loads the teacher model and distilled student model, preprocesses the input leaf image by resizing to 256×256 pixels and applying standard normalization, then performs simultaneous forward passes in evaluation mode. The output logits are converted to probabilities via softmax, and the top predicted class is identified. Predictions are overlaid on the image and saved for visualization, and the top-5 probabilities of classes are used to assess confidence and compare student-teacher alignment. This demonstrates that the compact student model maintains comparable accuracy with the teacher model while enhancing inference speed and efficiency. Performance of models is shown in Table II.

TABLE II: Training and Validation Accuracy of Teacher and Student Models

Epoch	Teacher Train Acc (%)	Teacher Val Acc (%)	Student Train Acc (%)	Student Val Acc (%)
1	95.72	99.07	78.09	93.32
2	99.02	99.02	94.57	97.98
3	98.90	98.80	96.91	98.64
4	99.53	98.91	97.89	99.02
5	99.18	**99.73**	98.26	**99.32**

Algorithm 1 Vision Transformer-based Knowledge Distillation

Input: Dataset D, temperature T, factor α
Output: Trained student model $\mathcal{M}_S$
Resize images to 256×256 pixels, split D into $D_{\text{train}}, D_{\text{val}}$
Teacher Training Phase
Load pretrained ViT-Large as $\mathcal{M}_T$
for each epoch in E_T **do**
  for each batch (x, y) in D_{train} **do**
    $y_T \leftarrow \mathcal{M}_T(x)$, $\mathcal{L} \leftarrow \text{CE}(y_T, y)$
    Update weights of $\mathcal{M}_T$
  end for
end for
Save $\mathcal{M}_T$
Student Training Phase with Distillation
Load pretrained MobileViT-XXS as $\mathcal{M}_S$
for each epoch in E_S **do**
  for each batch (x, y) in D_{train} **do**
    $y_T \leftarrow \mathcal{M}_T(x)$, $y_S \leftarrow \mathcal{M}_S(x)$ // Get outputs
    $\mathcal{L} \leftarrow \alpha \cdot \text{CE}(y_S, y) + (1 - \alpha) \cdot T^2 \cdot \text{KL}(\sigma(y_T/T), \sigma(y_S/T))$
    Update weights of $\mathcal{M}_S$
  end for
end for
Save $\mathcal{M}_S$

Algorithm 2 Inference and Prediction Pipeline

Input: Models $\mathcal{M}_T$, $\mathcal{M}_S$, image x
Output: Predicted classes and top-K probabilities
$x \leftarrow \text{preprocess}(x)$ // Resize and normalize
$y_S \leftarrow \mathcal{M}_S(x)$, $y_T \leftarrow \mathcal{M}_T(x)$ // Run inference using teacher and student models
$p_S \leftarrow \sigma(y_S)$, $p_T \leftarrow \sigma(y_T)$ // Get probabilities
$c_S \leftarrow \arg\max(p_S)$, $c_T \leftarrow \arg\max(p_T)$ // Predicted Probabilities
Output: c_S, c_T, and top-K values from p_S, p_T

V. Experimental Validation

To assess the effectiveness of the LiteViT framework, we conducted comprehensive experiments to ensure accuracy and interpretability. The proposed research was implemented using PyTorch TIMM (pytorch Image models) with GPU acclerated training, evaluated on a tomato leaf disease dataset containing 22,930 images across 10 classes (9 diseased, 1 healthy) which covers bacterial spot, yellow curl, early blight, leaf mold, septoria leafspot, target spot, late blight, mosaic virus, target spot, healthy. Both models were trained for five epochs, with the teacher achieving 99.73% and the student 99.32% , indicating successful knowledge transfer with minimal performance loss. Confusion matrices for teacher and student models revealed consistent per-class accuracy across all disease categories. The student model achieved an Area Under the Receiver Operating Characteristic Curve (ROC-AUC) of 1.00, indicating near-perfect classification performance and demonstrating the model's effectiveness. The classification results are shown in Figures 2 and 3 with over 95% correct predictions and high class separation. Figure 4 shows that all disease classes achieve AUCs above 0.96, demonstrating that distilled MobileViT-XXS maintains high sensitivity and specificity across thresholds. Overall, results with performance comparison with other works are in Table III.

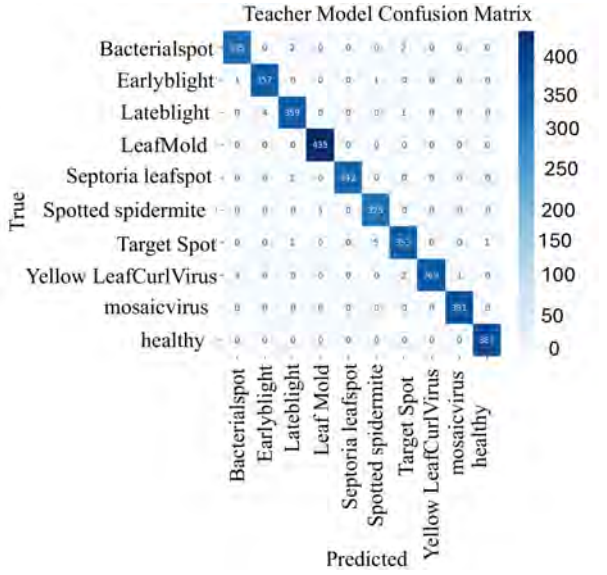


Fig. 2: Confusion Matrix of the Teacher Model on Validation Set

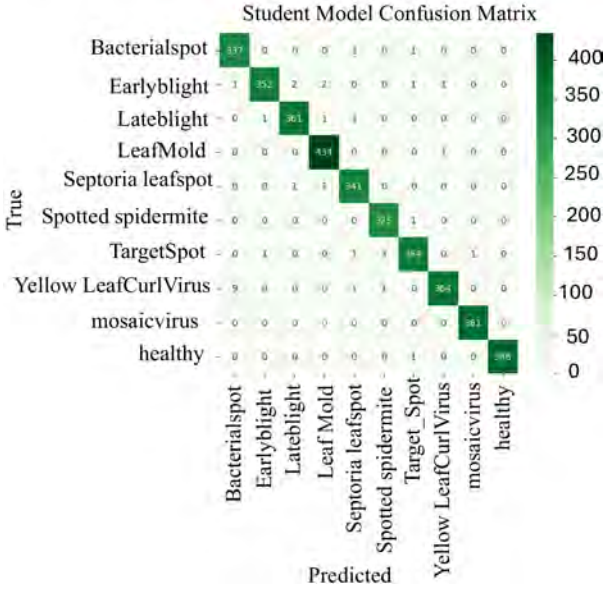


Fig. 3: Confusion Matrix of the Student Model on Validation Set

Pytorch Hooks were used to extract feature maps and attention weights from the student model's early convolutional layers to improve interpretability. The ex-

plainability module integrates three key visualization techniques:[1] Channel-wise feature maps from the stem layer visualize 16 activation channels in a grid to highlight low-level features. [2] Attention rollout aggregates attention weights across layers by sequential matrix multiplication, to trace hierarchical focus. [3] Class activation maps (CAM) overlay heatmaps on input images to highlight (e.g., leaf edges, spot textures) influencing predictions. Figure 5 provides attention maps output where the small yellow color dots show the affected areas. With an inference time of 12.26 milliseconds (vs 25.07 milliseconds for teacher) achieved on model with 99.6% fewer parameters with the size of 3.8 MB (0.95M) is highly efficient and suited for edge deployment on devices like Raspberry pi, supporting real-time, interpretable plant disease detection.

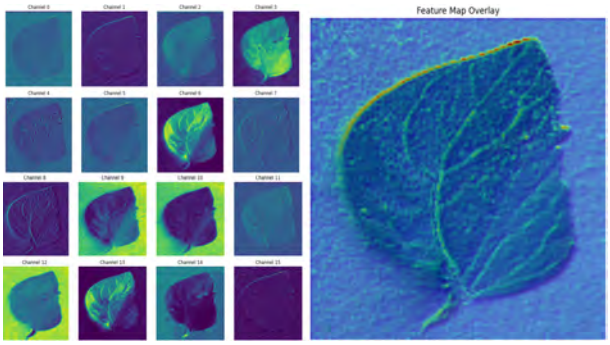


Fig. 5: Class Activation Maps Showing Disease-affected Leaf Regions

TABLE III: Comparison of Results with Current Research

Work	Models	Accuracy	Parameters(M-millions)	Size
Mohanty et al.[17]	AlexNet, GoogLeNet	99.35%	60M, 5M	240MB, 20MB
Wang et al.[3]	VGG16, shallow CNN	90.4%, 79.3%	138M, 5M	552MB, 20MB
Rahman et al.[2]	Lightweight CNN	93.3%	0.8M	3.2MB
Huang et al.[10]	Knowledge Distillation YOLOR	60.3%, 60.4%, 54.2% mAP@0.5	37M, 20.5M, 18.2M	141.6MB 78.4MB, 72.4MB
Li et al.[6]	PMVT (Lightweight ViT)	93.6%	0.98M	3.9MB
Parez et al.[11]	ViT (GreenViT)	100%	21.65M	247MB
Mahmud et al.[16]	CNN (EfficientNetB3)	99.3%	11.1M	44.38MB
LiteViT (Current Work)	**Knowledge Distilled ViT+ MobileViT**	**99.3%**	**0.95M**	**3.8MB**

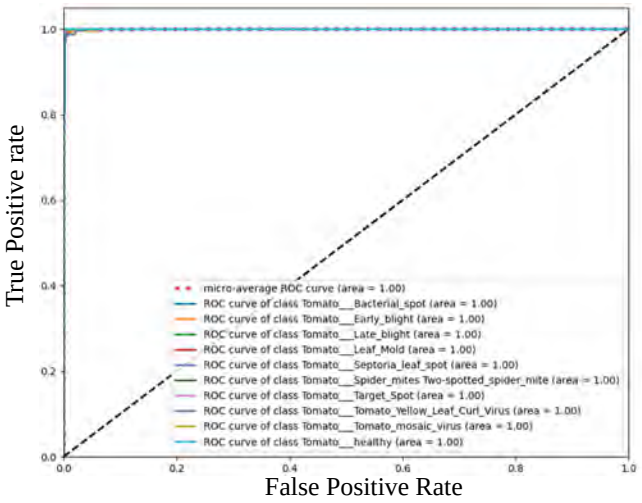


Fig. 4: Multi Class ROC-AUC Scores for disease classification

VI. Conclusion

This research presents a lightweight and interpretable framework for plant disease detection using a cross-architecture knowledge distillation strategy. The system achieves high accuracy by transferring diagnostic intelligence from a large Vision Transformer to a compact MobileViT-XXS model while supporting real-time inference on resource-constrained edge devices. The integration of adaptive learning with temperature-scaled distillation and multimodal feature visualization enhances interpretability by focusing on disease-relevant features and suppressing noise. Despite strong results, the limited environmental diversity in training data may hinder generalization. Future work will include exploring domain aware adaptation, segmentation-based localization, and active learning pipelines to improve robustness, adaptability and real-world performance.

References

[1] S. Savary, L. Willocquet, S. Pethybridge, P. Esker, N. McRoberts, and A. Nelson, "The global burden of pathogens and pests on major food crops,"

Nature Ecology & Evolution, vol. 3, pp. 430–439, Mar. 2019.

[2] C. R. Rahman, P. S. Arko, M. E. Ali, M. A. I. Khan, S. H. Apon, F. Nowrin, and A. Wasif, "Identification and recognition of rice diseases and pests using convolutional neural networks," *Biosystems Engineering*, vol. 194, pp. 112–120, Jun. 2020.

[3] G. Wang, Y. Sun, and J. Wang, "Automatic image-based plant disease severity estimation using deep learning," *Computational Intelligence and Neuroscience*, vol. 2017, no. 2917536, pp. 1–8, Jul. 2017.

[4] M. Arsenovic, M. Karanovic, S. Sladojevic, A. Anderla, and D. Stefanovic, "Solving current limitations of deep learning based approaches for plant disease detection," *Symmetry*, vol. 11, no. 7, p. 939, 2019.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[6] G. Li, Y. Wang, Q. Zhao, P. Yuan, and B. Chang, "Pmvt: A lightweight vision transformer for plant disease identification on mobile devices," *Frontiers in Plant Science*, vol. 14, Art.1256773, Sep. 2023.

[7] A. Mitra, S. L. Vangipuram, A. K. Bapatla, V. K. Bathalapalli, S. P. Mohanty, E. Kougianos, and C. Ray, "Smart agriculture: A comprehensive overview," *SN Computer Science*, vol. 5, no. 8, Art. 969, 2024.

[8] Y. Zhou, S. Chen, Y. Wang, and W. Huan, "Review of research on lightweight convolutional neural networks," in *IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, 2020, pp. 1713–1720.

[9] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[10] Q. Huang, X. Wu, Q. Wang, X. Dong, Y. Qin, X. Wu, Y. Gao, and G. Hao, "Knowledge distillation facilitates the lightweight and efficient plant diseases detection model," *Plant Phenomics*, vol. 5, pp. 1–12, 2023.

[11] S. Parez, N. Dilshad, N. S. Alghamdi, T. M. Alanazi, and J. W. Lee, "Visual intelligence in precision agriculture: Exploring plant disease detection via efficient vision transformers," *Sensors*, vol. 23, no. 15, Art. 6949, 2023.

[12] K. K. Kethineni, S. P. Mohanty, E. Kougianos, S. Bhowmick, and L. Rachakonda, "Spraycraft: Graph-based route optimization for variable rate precision spraying," *arXiv preprint arXiv:2412.12176*, 2024.

[13] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020.

[14] S. Mehta and M. Rastegari, "Mobilevit: Lightweight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2022.

[15] K. Wei, B. Chen, J. Zhang, S. Fan, K. Wu, G. Liu, and D. Chen, "Explainable deep learning study for leaf disease classification," *Agronomy*, vol. 12, no. 5, Art. 1035, 2022.

[16] T. Mahmud, K. Barua, A. Barua, N. Basnin, S. Das, M. Hossain, and K. Andersson, "Explainable ai for tomato leaf disease detection: Insights into model interpretability," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, Dec. 2023, pp. 1–6.

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

[18] V. Kumar, H. Arora, Harsh, and J. Sisodia, "Resnet-based approach for detection and classification of plant leaf diseases," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 495–502.

[19] A. Tabbakh and S. S. Barpanda, "A deep features extraction model based on the transfer learning model and vision transformer "tlmvit" for plant disease classification," *IEEE Access*, vol. 11, pp. 45 377–45 392, 2023.

[20] R. Liu, K. Yang, A. Roitberg, J. Zhang, K. Peng, H. Liu, Y. Wang, and R. Stiefelhagen, "Transkd: Transformer knowledge distillation for efficient semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 20 933–20 949, Dec. 2024.

[21] K. Khabarlak, I. Laktionov, and G. Diachenko, "Feature knowledge distillation using group convolutions for efficient plant pest recognition," in *Proc. International Conference on Agricultural Informatics*, Sep. 2024, pp. 102–110.

[22] Noulam, *Tomato - a large dataset of tomato leaf diseases*, https : / / www . kaggle . com / datasets / noulam/tomato, Accessed: 2025-04-17, 2024.