Enhanced Mobile GC Vit Architecture for Efficient Image Classification with Application to Plant Disease Detection

Mohamed Jawher Bahrouni[1], Faouzi Benzarti[2], Mohamed Touati[3], Sadok Ben Yahia[4]
Laboratory of Signal, Image and Information Technologies (LR-SITI)-National Engineering School of Tunis (ENIT), University of Tunis El Manar, Tunis, Tunisia[1, 2]
Lab-STICC/UMR CNRS 6285, University of Brest, Brest, France[3]
The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Sonderborg, Denmark[4]

***Abstract*—Efficient and accurate automated diagnosis of plant diseases remains a challenge for deployment on resource-constrained edge devices. While hybrid vision transformers like GCViT balance accuracy and efficiency, they often lose critical high-frequency details such as fine lesion textures and leaf margins that are essential for fine-grained disease classification. To address this gap, we propose the Enhanced High-Frequencies Global Context Visual Transformer (EHF-GCViT), a novel hybrid architecture designed to explicitly enhance high-frequency feature retention within a lightweight framework. The core innovations of EHF-GCViT include: first, a customized, lightweight convolutional refinement block based on depthwise separable operations that acts as a learnable pre-processor to preserve discriminative spatial details before tokenization; second, a gated convolutional block that replaces the final transformer stage, reducing the model memory footprint from 46.36 MB to 34.48 MB; and third, an adaptive normalization strategy to stabilize the training of the integrated heterogeneous layers. Extensive experiments on the PlantVillage tomato disease dataset demonstrate that EHF-GCViT achieves superior performance, surpassing the baseline GCViT, standard Vision Transformers (ViT), and CNN benchmarks (e.g., ResNet) in accuracy, precision, recall, and F1-score. These results validate that explicitly modeling high-frequency features within a hybrid transformer design provides a more memory-efficient and accurate backbone for practical plant disease detection systems targeting edge deployment.**

Keywords*—Hybrid transformer architecture; convolutional refinement block; gated convolution; edge devices; high-frequency features; tomato leaf disease classification

I. Introduction

Climate change is significantly increasing the incidence, severity, and geographic range of plant diseases. Studies, such as [1], have demonstrated that shifting climate patterns are altering host-pathogen interactions, often accelerating disease cycles and allowing pathogens to invade new agricultural zones. Among the affected crops, the tomato—widely cultivated and of high economic importance—is particularly susceptible to various foliar diseases. These infections can lead to substantial yield losses, as highlighted by the World Processing Tomato Council (WPTC), which projects an 11.5% drop in yields for processing tomatoes in 2025 [2].

In response to these pressing challenges, researchers and policymakers are increasingly turning to artificial intelligence (AI) to improve plant health monitoring. Investment in AI-driven technologies aims to develop scalable, automated systems capable of diagnosing plant diseases directly from leaf images.

Convolutional Neural Networks (CNNs) are highly effective at capturing local spatial patterns through convolutional operations. However, their limited receptive field restricts their capacity for modeling long-range dependencies, which is critical for understanding global context in complex agricultural environments. In contrast, Vision Transformers (ViTs) excel at modeling global relationships via self-attention mechanisms but are often reliant on large datasets and substantial computational resources, making them less suitable for many real-world, resource-constrained applications. This limitation has spurred the emergence of hybrid architectures, which combine the local feature extraction strengths of CNNs with the global representation capabilities of Transformers. This design offers a balanced compromise between precision, efficiency, and generalization. By integrating convolutional inductive biases with transformer-based contextual learning, hybrid models achieve superior accuracy and robustness—a critical advantage in domains characterized by subtle visual variations like plant disease detection. Therefore, developing efficient hybrid backbones that can adapt to challenges like varying lighting, leaf orientation, and disease symptom presentation is essential. These innovative frameworks not only promise superior results in specialized domains like agriculture but also establish a foundational benchmark for future progress in broader object detection research.

In this paper, we introduce a novel hybrid model, Hybrid HF GCVIT, designed to enhance the high frequencies for better classification. In this architecture, we introduce a new lightweight convolution block that is intended to be placed between the original GCVIT [3] blocks, enhancing the detection of high frequencies within the extracted features and thereby improving accuracy. We present a new normalization strategy to improve the quality of training and the accuracy of our model. This synergy yields a model that is both accurate

and computationally efficient, making it well-suited for real-time applications and potential deployment on low-power hardware platforms. We built GCViT with modularity in mind, unlike many prior models that optimize solely for classification. Its architecture is intended not only to perform tomato disease classification with high precision but also to serve as a robust backbone for more complex tasks, including object detection, segmentation, and anomaly localization in agricultural contexts. Experimental evaluations on publicly available tomato disease datasets validate the proposed model's superior performance in terms of accuracy, generalization, and computational efficiency of GCViT, surpassing that of state-of-the-art CNNs, ViTs, and hybrid networks. These results suggest that GCViT offers a practical and scalable solution for smart farming applications, especially where on-device intelligence is considered.

II. Related Work

A. CNN-Based Models

Convolutional Neural Networks (CNNs) have established themselves as a foundational approach for plant disease classification, showing their strong capabilities in extracting hierarchical visual features from images. A significant body of research focuses on customizing CNN architectures to enhance accuracy and efficiency for agricultural applications.

Early work demonstrated the potential of CNNs with high parameter counts and data augmentation. For instance, Guerrero-Ibañez and Reyes-Muñoz [4] presented a high-performance CNN model trained on a dataset enriched with real and GAN-generated synthetic images, achieving 99.64% validation accuracy for tomato leaf diseases. While demonstrating high performance, this approach relies heavily on extensive data augmentation rather than intrinsic architectural innovation, and its computational complexity limits suitability for resource-constrained edge devices.

Subsequent research addressed challenges like dataset imbalance and subtle inter-class differences through architectural modifications. Zhang et al. [5] proposed a lightweight dual-attention model (LDAMNet), integrating a dual-attention convolution block to achieve 98.71% average accuracy. However, while LDAMNet reduces model complexity, it operates within the conventional CNN paradigm and lacks the capacity for global context modeling, which is crucial for understanding long-range spatial dependencies in disease patterns.

The pursuit of deployable models has driven the adoption of efficient architectures and transfer learning. Sultan [6] employed a modified Xception architecture with deep transfer learning, achieving 98% accuracy for multiple crops while optimizing for edge deployment. Transfer learning, however, is constrained by features learned from generic datasets like ImageNet, which may not optimally capture disease-specific, high-frequency details such as fine lesion textures.

Radočaj et al. [7] introduced the novel IncMB module, integrating an Inception structure with the Mish activation function into MobileNetV2, achieving 97.78% accuracy and confirming potential for mobile platforms.

Similarly, Vivek Anandh et al. [8] showed a transfer-learned MobileNetV2 model could attain 99.49% accuracy for tomato leaf disease classification on IoT platforms. While these models achieve efficiency through depthwise separable convolutions, they prioritize overall parameter reduction over the explicit preservation of high-frequency spatial details, which can result in the loss of discriminative fine-grained features essential for distinguishing visually similar diseases.

This principle of lightweight design extends to highly optimized, purpose-built CNNs. FL-ToLeD [9] leverages soft attention and depthwise separable convolutions to achieve 99.04% accuracy with only 221,594 parameters. Critically, its attention mechanism operates on downsampled feature maps, inherently discarding high-frequency details before they can be weighed a fundamental limitation for fine-grained classification.

Ashurov et al. [10] integrated Squeeze-and-Excitation (SE) blocks into a Depthwise CNN, achieving 98% accuracy. Although SE blocks enhance channel-wise feature recalibration, they do not explicitly preserve high-frequency spatial information, which can be suppressed if not statistically dominant across channels.

The success of efficient CNNs has proven generalizable across crops, with specialized models achieving high accuracy for rice (99.81%) [11], potato (99.3%) [12], and multiple vegetables (97.12%) [13]. While demonstrating task adaptability, these models are typically optimized for single-crop scenarios and lack architectural mechanisms for explicit high-frequency feature enhancement.

To address data scarcity, data-centric approaches have emerged. Ramadan et al. [14] used CycleGAN for synthetic image generation to enable a MobileNetV2 model to achieve perfect wheat disease classification accuracy.

Joseph et al. [15] constructed real-time multi-crop datasets and proposed the MRW-CNN, achieving 97.04%–98.08% accuracy. These methods improve performance through augmentation rather than architectural innovation for feature preservation, and their reliance on synthetic data adds pre-processing complexity ill-suited for streamlined edge deployment.

Critical Gap in CNN-Based Approaches: While CNNs excel at local feature extraction and have been successfully optimized for edge deployment, they fundamentally struggle with modeling long-range dependencies and global context. More critically, aggressive model compression techniques (e.g., depthwise convolutions) often sacrifice high-frequency feature retention, leading to reduced discriminative power for fine-grained disease classification where subtle texture differences are paramount. No existing CNN approach explicitly models high-frequency spatial features as a core, learnable architectural component prior to feature abstraction.

B. Vision Transformer (VIT)-Based Models

Vision Transformers (ViTs) address the global context limitation of CNNs by capturing long-range spatial relationships across entire images through self-attention mechanisms.

Karimanzira et al. [16] developed a ViT with cascaded group attention (ViT-CGA), achieving top-tier accuracy and robustness, enhanced with Explainable AI (XAI) and an LLM for interpretability. However, pure ViT architectures typically require large-scale datasets and substantial computational resources for effective training, making them impractical for scenarios with limited agricultural data and strict edge-device constraints.

Efforts have been made to adapt ViTs for practical deployment. Barman et al. [17] proposed ViT-SmartAgri, a CNN-free Vision Transformer designed for smartphones, attaining 90.99% accuracy for tomato leaf conditions. While demonstrating mobile potential, this accuracy falls below contemporary CNN benchmarks, suggesting that pure ViTs may sacrifice critical local feature sensitivity, particularly the high-frequency edge and texture information crucial for precise disease boundary detection.

For object detection in complex environments, Wang and Liu [18] introduced TomatoDet, a real-time model based on a Swin Transformer variant, achieving a high mAP of 92.3% for localizing tomato diseases. Although Swin Transformers introduce local windowing to reduce computation, they maintain substantial memory footprints and lack explicit mechanisms to preserve high-frequency spatial details during the patch embedding process, which can blur fine lesion boundaries.

Critical Gap in ViT-Based Approaches: While ViTs excel at global context modeling, they face critical limitations for edge-deployed plant disease detection: 1) high computational and memory requirements, 2) data hunger requiring extensive training sets, and 3) inherent loss of high-frequency spatial information during patch tokenization, where fine-grained disease textures are smoothed or discarded.

C. Hybrid CNN-VIT-Based Models

Hybrid architectures have emerged to synergistically combine the local feature sensitivity of CNNs with the global context modeling of ViTs, aiming to balance both capabilities for complex visual classification.

Sinamenye et al. [19] developed a hybrid EfficientNetV2B3+ViT model for potato disease detection, achieving 85.06% accuracy—an 11.43% improvement over prior methods. However, this architecture employs a simple feature concatenation without addressing the fundamental issue of high-frequency information loss during the ViT tokenization stage, and its modest accuracy suggests suboptimal feature integration.

Alwan and Alturfi [20] introduced a sophisticated framework fusing EfficientNet-B8, a ViT, and a Knowledge Graph (KG), achieving 99.3% accuracy across 38 disease categories. While highly accurate, this approach relies on ensemble complexity and external knowledge integration rather than architectural efficiency, resulting in a model unsuitable for edge deployment due to its substantial computational overhead.

Li et al. [21] proposed the Convolution Self-Guided Transformer (CSGT), merging CNN feature extraction with transformer context modeling to achieve 95.8%–96.9% accuracy across multiple crops. Although demonstrating cross-dataset robustness, CSGT does not explicitly address high-frequency feature preservation, and its self-guided mechanism adds architectural complexity that may hinder optimization for memory-constrained devices.

Sun et al. [22] combines EfficientNetV2 with a Swin Transformer and Coordinate Attention, achieving 99.70% accuracy for tomato disease classification. Despite excellent accuracy, the model retains the full transformer stack in later stages, resulting in a large memory footprint (>45MB) that exceeds practical limits for edge devices. Furthermore, it lacks an explicit mechanism to enhance high-frequency features before tokenization.

Chen et al. [23] proposed a comprehensive framework combining a DenseNet-based Vision Transformer (DVT) with a CycleGAN (CyTrGAN) for augmentation, achieving up to 99.45% accuracy with strong robustness. While effectively addressing class imbalance through data augmentation, DVT does not architecturally prioritize the retention of high-frequency spatial details, and its reliance on GAN-based preprocessing increases deployment complexity.

Critical Research Gap: Existing hybrid CNN-ViT architectures demonstrate that combining local and global feature modeling improves plant disease classification. However, a unified solution remains elusive due to three interconnected limitations:

1) High-frequency feature loss: Current hybrids do not explicitly preserve fine-grained spatial details (e.g., lesion edges, texture patterns) during the CNN-to-ViT transition, as patch embedding and tokenization inherently smooth this critical high-frequency information.

2) Memory inefficiency: Most models retain full transformer stages, resulting in memory footprints (40-50MB+) that preclude deployment on resource-constrained edge devices common in agriculture.

3) Training instability: The integration of heterogeneous CNN and transformer layers often introduces optimization challenges due to conflicting normalization requirements and inconsistent gradient flow.

4) Our contribution: To directly address these gaps, we propose EHF-GCViT (Edge-optimized High-Frequency Gated Convolutional Vision Transformer).

Our novel architecture:

- Introduces a Lightweight High-Frequency Refinement block before the transformer stage to explicitly enhance and preserve discriminative fine-grained features,
- Replaces the final transformer stages with a Gated Convolutional Neck, significantly reducing the memory footprint while maintaining global context awareness, and
- Employs Adaptive Layer Normalization to stabilize the training of heterogeneous architectural components. This design provides a balanced solution that achieves

state-of-the-art accuracy with a memory profile explicitly optimized for real-world edge deployment.

III. Proposed Methodology

A. Asymmetric Depthwise Convolution Block

In this work, we propose a novel Asymmetric Depthwise Convolution Block (ADCB) designed to enhance the detection of high-frequency features, thereby improving the quality and sharpness of the extracted features.

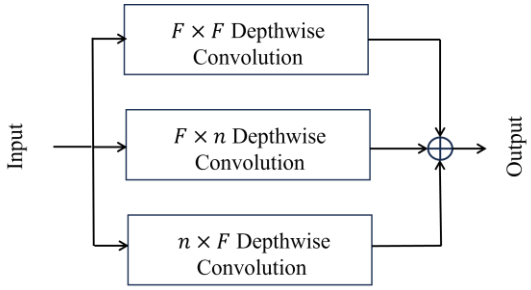

Fig. 1. The asymmetric depth-wise convolution block.

As shown in Fig. 1, the proposed block can be viewed as a lightweight variant of the Asymmetric Convolution Block (ACB) [24] where classical convolution layers are replaced with depthwise convolution layers to significantly reduce computational cost during feature extraction. While maintaining the conceptual structure of the original ACB, our block consists of a symmetric depthwise convolution with kernel size F × F to capture core spatial features and two asymmetric depthwise convolutions with kernel sizes, F × F , F × n and n × F where n < F unlike the original ACB, which restricts n=1, we introduce flexibility in the choice of n, allowing the creation of new asymmetric filters that can be tuned to emphasize directional high-frequency components with greater adaptability. The outputs of the three branches are summed elementwise to produce the block's final output. Like the original ACB, the overall structure can be treated as a single depthwise convolution layer with enhanced directional sensitivity.

B. High-Frequency Depthwise Separable Convolution Block

The high-frequency depthwise separable convolution block shown in Fig. 2, is a new variant of the depthwise separable convolution block, which is the fundamental block of the Mobilenet architecture [25]. We replace the classical depthwise layer with our asymmetric depthwise convolution block to sharpen the extracted features while retaining the 1x1 pointwise convolution for channel wise feature integration. This modification preserves the efficient structure of MobileNet while significantly improving its ability to capture high-frequency and directionally aware spatial features.

We use a residual connection from the block's input to its output to preserve essential base features while selectively refining details. This design enables the block to act as a feature enhancement unit, correcting and sharpening representations rather than replacing them outright. The residual path ensures stability during training and helps the refinement remain focused on informative residual patterns. The resulting block balances efficiency and accuracy, offering enhanced descriptive power without significantly increasing computational complexity, making it especially suitable for resource constrained applications such as mobile and embedded systems.

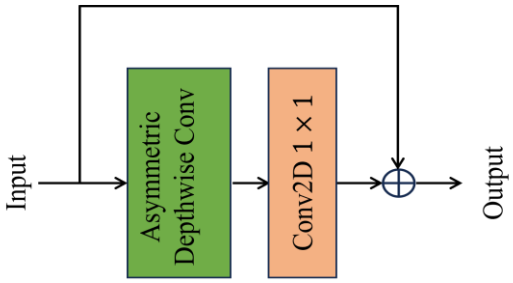

Fig. 2. The high-frequency depth-wise separable convolution block.

C. Normalization Strategy

In the proposed architecture, we apply Group Normalization (GN) following depthwise convolutions and Layer Normalization (LN) following 1×1 pointwise convolutions, in accordance with the functional properties of each normalization technique and the convolutional operations they follow. Depthwise convolutions operate on each input channel independently, making them inherently sparse in inter-channel interaction. GN, which normalizes within groups of channels rather than across the batch dimension, is well-suited for this context. Unlike Batch Normalization (BN), Group Normalization (GN) maintains performance with small batch sizes and better preserves the spatial structure of feature maps, making it more compatible with the independent-channel nature of depthwise operations. Moreover, GN avoids the over-smoothing behavior that Layer Normalization might induce in spatially distributed features, as LN treats each feature vector. Conversely, 1x1 convolutions perform intensive channel mixing, aggregating information across all feature channels at each spatial location. In this setting, LN proves advantageous because it normalizes across the full set of channels for each sample, offering sample-level consistency and effectively stabilizing the dense transformations introduced by pointwise convolutions. This dual strategy using GN with depthwise layers and LN with pointwise layers ensures that normalization remains functionally aligned with the nature of the operations it follows, improving convergence stability and feature quality without disrupting spatial integrity.

D. Gelu Channel Attention Fused MBC

In this work, we propose a novel modification of the previously introduced LRCA Fused MBC block [26] referred to as the GLCA MBC block. The term GLCA denotes the integration of the GELU activation function in place of the Leaky ReLU, applied consistently within both the channel attention module and all constituents' layers of the block. Additionally, we replace the standard convolutional layer with a dilated (atrous) convolution, which enlarges the receptive field without increasing the number of parameters, thereby enabling the network to capture broader contextual information while preserving spatial resolution.

This replacement of conventional 5×5 convolutions with dilated convolutions significantly reduces the computational

burden while maintaining a high level of classification precision. Through extensive experimentation, we demonstrate that this configuration is particularly effective for plant disease classification, offering a more computationally efficient alternative to large kernel convolutions by achieving an equivalent receptive field using smaller, dilated kernels.

Furthermore, we retain the same normalization strategy as the HF depthwise separable convolution block, applying Group Normalization (GN) after classical convolution layers and Layer Normalization (LN) after 1 × 1 pointwise convolutions. This targeted use of normalization improves stability and enhances learning dynamics across the heterogeneous components of the block. Fig. 3 illustrates the architectural design of the proposed GLCA Fused MBC block.

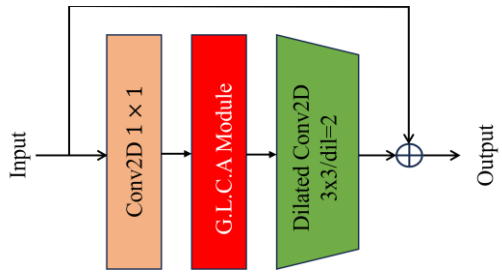

Fig. 3. GLCA fused MBC block.

E. Enhanced High-Frequencies GCVIT Architecture

As previously stated, our architecture is a modified version of the GCViTXXTiny model. We retain the initial patch embedding block and the three hierarchical levels of the original GCViT design. However, we introduce a key enhancement: a refinement layer is inserted between each pretrained pair of transformer stages as shown in Fig. 4.

Each refinement layer maintains the same input and output dimensions, enabling the use of residual connections to preserve and enrich the learned feature representations. This design helps enhance intermediate features without disrupting the original flow of information. The asymmetric filter dimensions were selected as F=5 and n=2 since through experimental optimization this combination produces the most favorable results.

Additionally, we replace the final transformer level of GCViT with a GLCA Fused MBC block, which significantly reduces the number of parameters while maintaining competitive accuracy. This modification leads to a 25.62% reduction in memory usage, making the model lightweight and efficient for deployment on resource constrained devices. Despite these changes, experimental results demonstrate that modified architecture preserves strong performance across evaluation metrics.

Fig. 4 shows the enhanced high-frequency hybrid architecture (with the original GCVIT frozen layers in blue, refinement layers in yellow, GLCA Fused MBC in green, global average pooling in red and Conv 1x1 in pink).

Table I provides a detailed description of the proposed architecture, including the input and output sizes of each layer, the stride values, and the number of groups used in Group Normalization (GNorm).

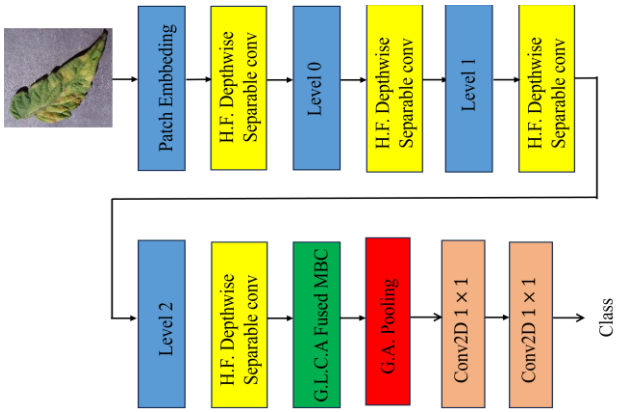

Fig. 4. The enhanced high-frequency hybrid architecture (with the original GCVIT frozen layers in blue, refinement layers in yellow, GLCA Fused MBC in green, global average pooling in red and Conv 1x1(classifier) in pink).

TABLE I. NETWORK ARCHITECTURE LAYERS WITH INPUT/OUTPUT SHAPES AND PARAMETERS

Block	Input	Output	G	
Patch Embedding	$224^2 \times 3$	$56^2 \times 64$	-	2
Dropout	$56^2 \times 64$	$56^2 \times 64$	-	-
H.F D.W.S.Conv0	$56^2 \times 64$	$56^2 \times 64$	8	1
Level 0	$56^2 \times 64$	$28^2 \times 128$	-	2
H.F D.W.S.Conv1	$28^2 \times 128$	$28^2 \times 128$	16	1
Level 1	$28^2 \times 128$	$14^2 \times 256$	-	2
H.F D.W.S.Conv2	$14^2 \times 256$	$14^2 \times 256$	32	1
Level 2	$14^2 \times 256$	$7^2 \times 512$	-	2
H.F D.W.S.Conv3	$7^2 \times 512$	$7^2 \times 512$	32	1
GL.CA Fused MBC	$7^2 \times 512$	$7^2 \times 512$	32	1
Global Average Pooling	$7^2 \times 512$	$1^2 \times 512$	-	-
Conv2D 1×1	$1^2 \times 512$	$1^2 \times 1028$	-	1
Conv2D 1×1	$1^2 \times 1028$	$1^2 \times Class$	-	1

G=Group size and S=Stride

IV. RESULTS

In this section, we describe the experimental setup and the different experiments conducted to rigorously evaluate our model, as well as to compare our results with those obtained from state-of-the-art architectures. In this section, we describe the experimental setup and the different experiments conducted to rigorously evaluate our model, as well as to compare our results with those obtained from state-of-the-art architectures.

A. Dataset

To evaluate the performance of our proposed system, we utilized a dataset of tomato leaf diseases extracted from the PlantVillage [27] database, made available publicly through the repository [28] .This dataset comprises 14,531 tomato leaf images, covering both healthy specimens and those affected by various diseases. Table II shows that the images are categorized into ten distinct classes, which include nine disease types: Bacterial Spot, Early Blight, Late Blight, Leaf Mold, Septoria Leaf Spot, Target Spot, Tomato Mosaic Virus,

Tomato Yellow Leaf Curl Virus, and Two-Spotted Spider Mite, as well as one healthy class. Each image generally depicts a single leaf captured against a uniform background, providing a clean and consistent input for classification tasks. Due to its scale and diversity, this dataset has become a widely adopted benchmark for assessing the effectiveness of machine learning and deep learning models in plant disease diagnosis. The version of the dataset we used includes five predefined folds intended for cross-validation. Each fold is split into 80% for training and 20% for testing, allowing for robust and systematic evaluation across multiple runs. In our experiments, we employed all five folds for cross-validation tests. For ablation studies and comparative analysis with state-of-the-art models, we used only the first fold to ensure consistency and efficiency. During the training phase, we further allocated 10% of the training subset from each fold for validation, while the remaining 90% was used for model training.

TABLE II. NUMBER OF SAMPLES PER CLASS IN THE DATASET

Class	Number of samples per class
Bacterial spot	1702
Early blight	800
Healthy	1272
Late blight	1528
Leaf Mold	762
Septoria leaf spot227	1417
Target Spot227	1124
Tomato mosaic virus227	299
Tomato Yellow Leaf Curl Virus227	4286
Two-spotted spider mite227	1341
Total number of images	14531

B. Metrics

In our work, we opted to use confusion matrix, Accuracy, Precision, Recall, and F1-score as statistics to evaluate the performance of our model. The confusion matrix is a matrix that allows us to measure the performance of a system during a classification task by providing information about the comparison between original classes and predicted classes. This information is classified into four categories:

True positive (TP): When the instance is positive, and the model classifies it as positive.

True negative (TN): When the instance is negative, and the model classifies it as negative.

False positive (FP): When the instance is negative, and the model classifies it as positive.

False negative (FN): When the instance is positive, and the model classifies it as negative.

Accuracy is the basic metric used in most evaluations, and it shows the number of predictions over all predictions. It is calculated by splitting the number of correct predictions by the total number of predictions.

$$Accuracy = \frac{Tp+TN}{TP+FP+TN+FN} \quad (1)$$

Precision shows the number of true positive predictions compared to the total number of all positive predictions.

$$Precision = \frac{Tp}{TP+FP} \quad (2)$$

Recall measures the number of positive cases compared to the total number of all the predictions classified as positive.

$$Recall = \frac{Tp}{TP+FN} \quad (3)$$

F1-score is a measure that we calculate as a harmonic mean or the weighted average of precision and recall.

$$F1-score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (4)$$

C. Training Strategy

To leverage the representational power of pre-trained models while adapting them to our specific classification task, we adopted a transfer learning strategy. The core layers and blocks of the original GCViT architecture were frozen, meaning their weights remained unchanged during training to preserve features learned from large-scale datasets. We integrated our custom blocks HF Depthwise Separable Convolution, GLLCA, and Fused MBC, along with a classification head specifically designed for our target task. Only these newly added components were trained during the training process. This selective training approach reduces computational cost and helps prevent overfitting, particularly when working with limited data. Optimization was performed using the Weighted Adam (W.Adam) algorithm, an enhanced version of Adam that incorporates weight-aware mechanisms to improve both convergence speed and generalization performance.

D. Ablation Study

To assess the individual contributions of key design choices in our hybrid vision model, we conducted a structured ablation study. This process involved systematically removing specific components from our hybrid architecture to measure their direct impact on performance. We focused on evaluating the role of the HF depth-wise separable convolution block (Model A) and the GLCA-Fused MBC (Model B) in improving classification accuracy for plant disease detection.

From the training perspective as shown in Table III, all variants achieved near-perfect accuracy, with losses approaching zero. Model A and the Hybrid GCViT attained 99.92% and 99.94% training accuracy, respectively, accompanied by equally high validation accuracy (99.91% and 99.74%), indicating excellent learning capacity with no signs of overfitting. Although Model B experienced a minor drop in validation accuracy (99.31%) compared to its training accuracy (99.92%), its performance remained competitive, especially given its reduced model size (33.29 MB).

Table IV presents the testing performance on the testing data set and shows that Model A achieved the highest accuracy (99.55%) and F1 score (99.48%), closely followed by the Hybrid GCViT (99.52% accuracy, 99.47% F1). Despite being slightly smaller than Model A, the Hybrid model maintained

nearly identical performance, suggesting that the combination of both enhancements led to a well-balanced architecture. Overall, this ablation study confirms that each modification contributes positively, with the hybrid configuration offering the best trade-off between accuracy, generalization, and model efficiency.

TABLE III. TRAINING STATISTICS OF THE ABLATION STUDY EXPERIMENTS

Model	Train. Accuracy (%)	Train. Loss	Val. Accuracy (%)	Val. Loss	Size (mb)
Original GC VIT	99.29	0.02	99.05	0.05	46.36
Model A	99.92	0.00	99.91	0.01	48.55
Model B	99.92	0.00	99.31	0.04	33.29
EHF-GCViT	**99.94**	**0.00**	**99.74**	**0.03**	**34.48**

Train. = training and Val. = validation

TABLE IV. TESTING STATISTICS OF THE ABLATION STUDY EXPERIMENTS

Model	Test. Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Original GC VIT	98.49	98.08	98.25	98.15
Model A	99.55	99.45	99.52	99.48
Model B	99.45	99.30	99.28	99.29
EHF-GCViT	**99.52**	**99.44**	**99.50**	**99.47**

Train. = training and Val. = validation

E. Cross-Validation Test

To ensure robust and unbiased evaluation of the proposed model, a 5-fold cross-validation protocol was adopted using the folds provided in the Mendeley dataset for tomato disease. This dataset includes predefined training, validation, and testing splits for each fold, enabling consistent and reproducible assessments across multiple experiments. By rotating through the five folds, each model is trained and tested on different subsets of data, which helps mitigate the risk of overfitting and provides a more reliable estimate of generalization performance. This approach is particularly valuable in plant disease classification, where dataset imbalance and subtle inter-class variations can otherwise distort performance metrics if a single train-test split is used. The use of standardized folds also facilitates fair comparisons between models under identical evaluation conditions.

The performance of the proposed model was assessed using 5-fold cross-validation , and the evaluation metrics demonstrate consistently high classification quality across all folds. Table V shows that accuracy ranged from 99.21% to 99.66%, with a mean above 99.4%, indicating the model's strong overall ability to correctly classify instances. Precision values remained between 99.00% and 99.49%, suggesting that the model maintains a low rate of false positives across the different test splits. Similarly, recall values ranged from 98.92% to 99.65%, confirming that the model effectively captures true positive cases with minimal false negatives. The F1-score, which balances precision and recall, was consistently high, above 98.9% in all folds, highlighting the robustness and reliability of the model's predictions. The low variability among the folds further suggests that the model generalizes well to unseen data and is not overly sensitive to the composition of the training and validation subsets.

TABLE V. CROSS-VALIDATION TESTING PERFORMANCES

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Fold 1	99.52	99.44	99.50	99.47
Fold 2	99.21	99.00	98.92	98.95
Fold 3	99.66	99.49	99.65	99.57
Fold 4	99.55	99.42	99.47	99.44
Fold 5	99.24	99.04	99.17	99.10

F. Comparison with Other Models

To evaluate the performance of our proposed architecture, we conducted a comparative analysis against a diverse set of state-of-the-art models, encompassing convolutional neural networks (CNNs), vision transformers (ViTs), and hybrid architectures. We chose these models for comparison based on their availability in the Keras API and their proven stability and performance in image classification tasks. The deep learning community widely adopts these models, which provide reliable implementations with pre-trained weights, making them suitable for fair and reproducible benchmarking. The models are described as follows:

1) ConvNeXt Base [18]: ConvNeXt is a modernized convolutional architecture that integrates design principles from vision transformers into traditional CNN frameworks. Key enhancements include the use of large kernel sizes, inverted bottlenecks, and Layer Normalization, aligning CNNs more closely with transformer-based models.

2) EfficientNetV2-M [29]: EfficientNetV2-M is part of the EfficientNetV2 family, which introduces a combination of MBConv and Fused-MBConv blocks to improve training speed and parameter efficiency. Architecture employs progressive learning strategies, adjusting image size and regularization during training to achieve better performance.

3) GCViT XXTiny: GCViT (Global Context Vision Transformer) is a vision transformer designed to capture global context efficiently. The XXTiny variant is an ultra-lightweight model, making it suitable for resource-constrained environments. Our proposed hybrid model draws inspiration from this variant.

4) GCViT Base: The GCViT Base model expands upon the XXTiny variant, offering a more substantial architecture that balances performance and efficiency. It utilizes hierarchical attention mechanisms to effectively model global context in visual data.

5) ResNet100: ResNet100 is a deep residual network that employs skip connections to mitigate the vanishing gradient problem, enabling the training of very deep networks. It serves as a strong baseline for evaluating the performance of deep CNNs.

6) MobileNetV3-Large: MobileNetV3-Large is an efficient architecture optimized for mobile and embedded

applications. It incorporates neural architecture search (NAS), squeeze-and-excitation (SE) blocks, and novel activation functions like h-swish to achieve a balance between accuracy and latency.

7) ViT Base [30]: The Vision Transformer (ViT) Base model applies trans-former architectures, originally developed for natural language processing, to image classification tasks. It divides images into patches and processes them using self-attention mechanisms, demonstrating the viability of transformer-based models in computer vision.

Table VI summarizes the training performance of all evaluated models, showing the statistics of all evaluated models during the training process.

TABLE VI. PERFORMANCES OF ALL MODELS DURING THE TRAINING PROCESS

Model	Training Accuracy (%)	Training Loss	Validation Accuracy (%)	Validation Loss
ConvNeXt Base	100.0	0.0	98.88	0.079
EfficientNetV2-M	99.32	0.023	97.42	0.106
GCViT XXTiny	99.27	0.021	99.14	0.050
GCViT Base	99.51	0.018	99.48	0.055
ResNet100	99.98	0.001	98.02	0.156
MobileNetV3 Large	99.99	0.002	98.36	0.094
ViT Base	100.0	0.0	98.8	0.075
EHF-GCViT	**99.94**	**0.004**	**99.74**	**0.027**

Fig. 5 shows the evolution of training and validation accuracy during the training process.

Fig. 6 presents the evolution of training and validation loss during the same process.

Table VII summarizes the testing performance of all evaluated models, highlighting a comparison between the proposed EHF hybrid GCVIT and selected state-of-the-art approaches.

The confusion matrix presented in Fig. 7 illustrates the classification performance of our model across 10 classes of tomato leaf diseases.

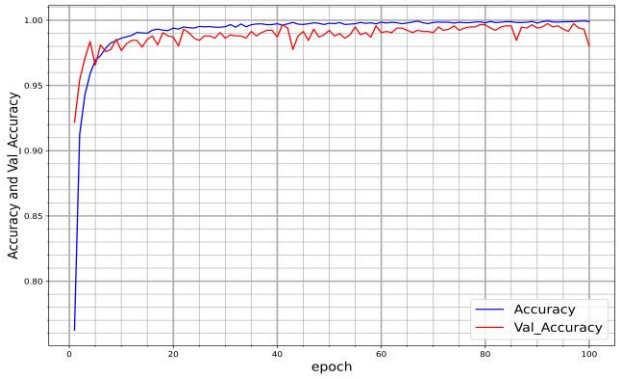

Fig. 5. Evolution of training and validation accuracy over epochs.

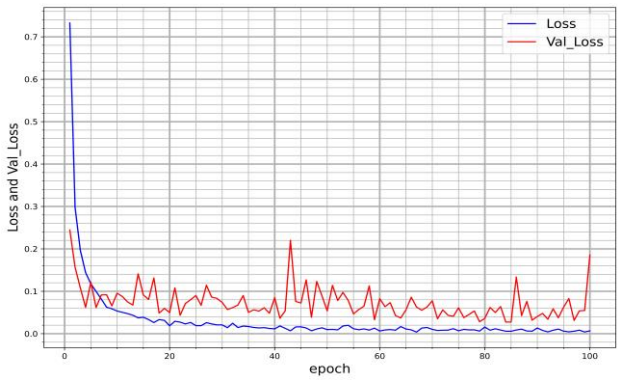

Fig. 6. Evolution of training and validation loss over epoch.

TABLE VII. PERFORMANCES OF ALL MODELS DURING THE TESTING PHASE

Model	Testing Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ConvNeXt Base	96.81	96.14	94.52	95.11
EfficientNetV2-M	96.67	95.64	95.94	95.73
GCViT XXTiny	98.5	98.08	98.03	98.05
GCViT Base	99.18	98.69	99.12	98.88
ResNet100	97.25	96.38	96.87	96.60
MobileNetV3 Large	96.87	95.87	96.04	95.93
ViT Base	98.45	97.91	98.27	98.08
EHF-GCViT	**99.52**	**99.44**	**99.50**	**99.47**

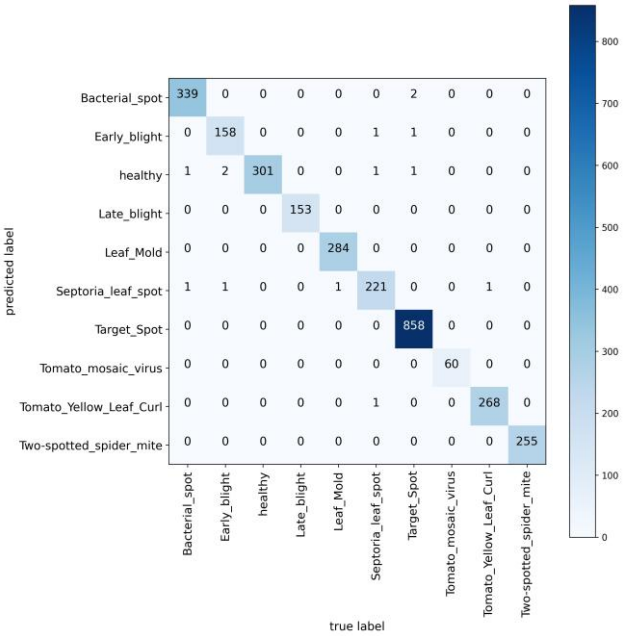

Fig. 7. Confusion matrix of plant tomato diseases classification results using EHF hybrid GCVIT architecture.

Most classes achieved near-perfect classification, with minimal confusion between categories. Notably, Target Spot (858), Two-spotted spider mite (255), Tomato Yellow Leaf Curl (268), and Late blight (153) were classified with perfect

or near-perfect accuracy, demonstrating the model's robustness on these categories. Mi-nor misclassifications occurred for a few classes, such as healthy leaf, which was occasionally confused with Bacterial spot and Septoria leaf spot, and Early blight, which had slight confusion with Septoria leaf spot. Overall, the confusion matrix confirms strong discriminative capability across all classes, with misclassifications being both sparse and low in magnitude.

V. Discussion

This research delivers an in-depth assessment of eight vit and deep learning architectures for plant disease classification, with a focus on both predictive accuracy and computational efficiency. The training, validation, and testing outcomes offer valuable perspectives on each model's learning behavior, generalization capability, and practicality for deployment in real-world settings.

The results show that our proposed method (Hybrid GCViT) consistently demonstrated the strongest overall performance. It achieved the highest validation accuracy (99.74%) and test accuracy (99.52%), along with excellent precision (99.44%), recall (99.50%), and F1 score (99.47%). Despite its lightweight size (34.48 MB) and ranking second to MobileNet V3 Large in memory efficiency, it surpassed all other models across key metrics. This suggests that the proposed hybridization of GCViT with convolutional components effectively captures both local and global features, enhancing both learn-ing and generalization while remaining computationally efficient.

The GCViT family, particularly GCViT Base and GCViT XXTiny, also demonstrated excellent performance. GCViT Base achieved a validation accuracy of 99.48% and a test accuracy of 99.18%, with high precision and recall. GCViT-XXTiny, although more compact, maintained a competitive test accuracy of 98.5%, highlighting the scalability of GCViT-based designs for both high-end and resource-constrained applications.

Within the classical convolutional and hybrid transformer models, ViT Base and ConvNeXt Base achieved perfect training accuracy (100%). However, their performance on the validation and test datasets was slightly lower. This discrepancy suggests a tendency toward overfitting, particularly in ConvNeXt Base, which exhibited a noticeable gap between training and test accuracy (100% vs 96.81%). These findings underscore the importance of applying effective regularization strategies when deploying highly expressive model architectures.

ResNet100 and EfficientNetV2-M showed strong performance, with test accuracy of 97.25% and 96.67%, respectively. However, they were outperformed by the trans-former-based models, particularly in terms of F1 score and model compactness. Similarly, MobileNetV3 Large, known for its efficiency, maintained a relatively high-test accuracy of 96.87%but was outclassed by more recent architectures in both precision and recall.

Another key observation is the relationship between model size and performance. Notably, Hybrid GCViT, despite being the smallest model, delivered the best accuracy. This indicates that newer architectural designs can achieve superior performance without significantly increasing computational demands, which is critical for real-time or edge-based agricultural applications.

In summary, the results highlight the superior performance of transformer based and hybrid transformer-convolutional architectures in classifying diseases in tomato plants. The proposed Hybrid GCViT excels in accuracy and generalization and offers practical advantages in terms of deployment feasibility, making it a strong candidate for integration into intelligent agricultural systems.

VI. Conclusion and Future Work

This work introduces EHF-GCViT, a novel lightweight hybrid architecture for tomato leaf disease classification that addresses a key limitation of existing hybrid vision transformers: the loss of high-frequency spatial information during feature abstraction, along with high memory demands that hinder deployment on edge devices. By explicitly enhancing fine-grained feature preservation within the GCViT framework, the proposed approach achieves a more effective balance between accuracy and computational efficiency.

The proposed architecture integrates three complementary design strategies: (i) a customized lightweight convolutional refinement mechanism that improves the extraction of high-frequency discriminative details prior to transformer processing; (ii) a gated convolutional replacement for the final transformer stage that significantly reduces model memory consumption from 46.36 MB to 34.48 MB; and (iii) an adaptive normalization strategy that stabilizes the training of heterogeneous convolutional and transformer layers. Together, these design choices enable efficient feature refinement without compromising generalization capability.

Extensive experiments conducted on the PlantVillage tomato disease dataset demonstrate that EHF-GCViT consistently outperforms the baseline GCViT, standard Vision Transformers, and representative CNN-based architectures. The proposed model achieves a classification accuracy of 99.52%, confirming its suitability for resource-constrained environments. Ablation studies further validate the contribution of each architectural component and highlight the effectiveness of the proposed design trade-offs.

Beyond performance gains, this study emphasizes the importance of task-driven architectural design in hybrid transformer models. The results show that explicitly modeling high-frequency information rather than increasing model depth or complexity can lead to more accurate and efficient solutions for practical agricultural AI applications.

While the current evaluation demonstrates strong performance on the PlantVillage dataset, several directions warrant further investigation. Validation across diverse real-world field conditions with varying lighting, occlusions, and backgrounds would strengthen evidence of practical robustness. Additionally, extending the approach to multiple crop types and evaluating deployment performance on actual edge hardware platforms would provide valuable insights into the architecture's generalizability and operational feasibility. Moreover, the effectiveness of the proposed high-frequency

refinement may be influenced by image resolution and noise characteristics, which warrants further investigation under diverse sensing conditions. These considerations represent natural next steps in transitioning the model from controlled evaluation to production agricultural systems.

Future work will focus on extending EHF-GCViT as a general-purpose backbone for additional computer vision tasks, such as object detection and semantic segmentation, as well as systematic evaluation under real-world deployment scenarios. Owing to its modular and lightweight design, the proposed architecture holds promise for broader adoption in intelligent agricultural systems and other edge-based visual perception applications.

REFERENCES

[1] N. Abdullahi, K. F. Shamuwa, B. A. Tahir, I. Abubakar, and D. Yekeh, "Climate change and its effects on plant pathogen evolution: systematic review," Dutse Journal of Pure and Applied Sciences, vol. 11, no. 1d, pp. 248-256, 2025.

[2] W. P. T. Council. "WPTC Expects 40.5 Million Tonnes in 2025." https://www.tomatonews.com/en/conference/wptc-expects-405-million-tonnes-in-2025_2_2561.html (accessed.

[3] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global context vision transformers," in International Conference on Machine Learning, 2023: PMLR, pp. 12633-12646.

[4] A. Guerrero-Ibañez and A. Reyes-Muñoz, "Monitoring tomato leaf disease through convolutional neural networks," Electronics, vol. 12, no. 1, p. 229, 2023.

[5] E. Zhang, N. Zhang, F. Li, and C. Lv, "A lightweight dual-attention network for tomato leaf disease identification," Frontiers in Plant Science, vol. 15, p. 1420584, 2024.

[6] T. Sultan et al., "LeafDNet: Transforming Leaf Disease Diagnosis Through Deep Transfer Learning," Plant Direct, vol. 9, no. 2, p. e70047, 2025.

[7] P. Radočaj, D. Radočaj, and G. Martinović, "Image-based leaf disease recognition using transfer deep learning with a novel versatile optimization module," Big Data and Cognitive Computing, vol. 8, no. 6, p. 52, 2024.

[8] K. Vivek Anandh, A. Sivasubramanian, V. Sowmya, and V. Ravi, "Multiclass Classification of Tomato Leaf Diseases Using Convolutional Neural Networks and Transfer Learning," Journal of Phytopathology, vol. 172, no. 6, p. e13423, 2024.

[9] M. H. Alnamoly, A. A. Hady, S. M. Abd El-Kader, and I. El-Henawy, "FL-ToLeD: An improved lightweight attention convolutional neural network model for tomato leaf diseases classification for low-end devices," IEEE Access, 2024.

[10] A. Y. Ashurov et al., "Enhancing plant disease detection through deep learning: a Depthwise CNN with squeeze and excitation integration and residual skip connections," Frontiers in Plant Science, vol. 15, p. 1505857, 2025.

[11] M. H. Bijoy et al., "Towards sustainable agriculture: a novel approach for rice leaf disease detection using dCNN and enhanced dataset," IEEE Access, 2024.

[12] A. Saha, S. M. Musharraf, A. Dey, H. Roy, and D. Bhattacharjee, "Potato Leaf Disease Detection using CNN-A Lightweight Approach," 2025.

[13] M. a. A. Al-Shannaq, S. AL-Khateeb, A. A.-R. K. Bsoul, and A. A. Saifan, "Identification of Plant Diseases in Jordan Using Convolutional Neural Networks," Electronics, vol. 13, no. 24, p. 4942, 2024.

[14] S. T. Y. Ramadan et al., "Improving wheat leaf disease classification: Evaluating augmentation strategies and CNN-based models with limited dataset," IEEE Access, vol. 12, pp. 69853-69874, 2024.

[15] D. S. Joseph, P. M. Pawar, and K. Chakradeo, "Real-time plant disease dataset development and detection of plant disease using deep learning," Ieee Access, vol. 12, pp. 16310-16333, 2024.

[16] D. Karimanzira, "Context-Aware Tomato Leaf Disease Detection Using Deep Learning in an Operational Framework," Electronics, vol. 14, no. 4, p. 661, 2025.

[17] U. Barman et al., "Vit-SmartAgri: vision transformer and smartphone-based plant disease detection for smart agriculture," Agronomy, vol. 14, no. 2, p. 327, 2024.

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976-11986.

[19] J. H. Sinamenye, A. Chatterjee, and R. Shrestha, "Potato plant disease detection: leveraging hybrid deep learning models," BMC Plant Biology, vol. 25, no. 1, pp. 1-15, 2025.

[20] W. H. Alwan and S. M. Alturfi, "Multi-Stage Vision Transformer and Knowledge Graph Fusion for Enhanced Plant Disease Classification," Computer Systems Science and Engineering, vol. 49, no. 1, pp. 419-434, 2025, doi: 10.32604/csse.2025.064195.

[21] H. Li, N. Li, W. Wang, C. Yang, N. Chen, and F. Deng, "Convolution Self-Guided Transformer for Diagnosis and Recognition of Crop Disease in Different Environments," IEEE Access, 2024.

[22] Y. Sun, L. Ning, B. Zhao, and J. Yan, "Tomato Leaf Disease Classification by Combining EfficientNetv2 and a Swin Transformer," Applied Sciences, vol. 14, no. 17, p. 7472, 2024.

[23] 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, 2024.

[24] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1911-1920.

[25] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[26] M. J. Bahrouni and F. Benzarti, "Combined Mobilenet for Plant Diseases Classification," in 2024 IEEE International Conference on Artificial Intelligence & Green Energy (ICAIGE), 2024: IEEE, pp. 1-6.

[27] D. P. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing," arXiv preprint arXiv:1511.08060, pp. 1-13, 2015.

[28] M.-L. Huang and Y.-H. Chang. Dataset of Tomato Leaves, Mendeley Data, 2020-05-27, doi: 10.17632/ngdgg79rzb.1.

[29] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in International conference on machine learning, 2021: PMLR, pp. 10096-10106.

[30] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.