Journal of Theoretical and Applied Information Technology
31st December 2025. Vol.103. No.24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ADAPTIVE HYBRID VISION TRANSFORMER FOR PRECISE SEGMENTATION OF MANGO LEAF DISEASES

NURUL AKMAR AZMAN[1], FATIMAH KHALID[2], PUTERI SUHAIZA SULAIMAN[3],ZAINAL ABDUL KAHAR[4]

[1,2,3,4] Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[1]School of Pre-University Studies, Taylor's University,47500 Subang Jaya, Malaysia

E-mail:[1] nurulakmar.azman@taylors.edu.my (corresponding author), [2]fatimahk@upm.edu.my, [3]psuhaiza@upm.edu.my, [4]zainal_kahar@upm.edu.my

ABSTRACT

Early and accurate detection of leaf diseases is critical for sustaining crop health and preventing yield loss in modern agriculture. However, existing deep learning segmentation models often struggle to balance local feature extraction with global contextual understanding. This study introduces an adaptive hybrid architecture that integrates U-Net's precise localization with a Vision Transformer (ViT) encoder enhanced by adaptive positional embeddings, which dynamically preserve spatial relationships and adapt to varying leaf morphologies. A tailored preprocessing pipeline combining context-aware brightness normalization and HSV-based background removal further improves symptom visibility under diverse lighting conditions. Using the MangoLeafBD dataset (2,500 annotated images across five classes) to validate the framework, the proposed model achieves 99.71% accuracy and a mean IoU of 0.973, surpassing existing methods by 0.78–8.13% while maintaining efficiency on mid-range GPUs. This study establishes a scalable benchmark for agricultural image segmentation, advancing the architecture of hybrid transformer with CNN integration for field-level disease detection.
Keywords: *Vision Transformers, U-Net, Semantic Segmentation, Adaptive Positional Embeddings*

1. INTRODUCTION

In modern agricultural management, disease monitoring plays a crucial role, especially in protecting crop yields to ensure food security [1]. Traditional diagnostic approaches that depend on visual expert inspection are plagued by subjectivity and inaccuracies. In mango cultivation, early disease symptoms like leaf spots, discoloration, and wilting frequently overlap across different pathogens, making it difficult to achieve a precise diagnosis through observation alone [2]. These challenges are also worsened by environmental factors, including temperature fluctuations and humidity, which significantly affect the symptoms [3]. Delayed disease detection has caused devastating consequences for crops, with studies showing crop losses exceeding 50%, highlighting the urgent need for a high-precision monitoring system [4].

Recent progress in deep learning-based image segmentation has introduced transformative and promising solutions. Models with architectures such as U-Net and Convolutional Neural Networks (CNNs) have demonstrated impressive achievements in leaf disease detection, achieving accuracy rates exceeding 95% [5]. However, these approaches face two fundamental limitations that restrict their overall effectiveness. First, while CNNs are highly effective at extracting localized, fine-grained features, they often struggle to capture broader disease patterns and spatial relationships across the entire leaf surfaces [6]. Second, the pure Vision Transformer (ViT) model's capability of modeling long-range dependencies demands excessive computational resources, especially when processing high-resolution agricultural images [7]. Recent hybrid architectures such as TransUNet [8] and Swin Transformer [9] have attempted to address these challenges. Still, their reliance on static positional embeddings after training restricts their adaptability to the diverse morphological variations found in field conditions.

While recent studies [8], [10], [11], and [29] have reported high accuracies using improved transformer

architecture and attention-based models, most of these approaches remain constrained by static positional embeddings, fixed patch resolutions, or reliance on heavy patch-level computations that reduce adaptability in varied field conditions. For example, approaches such as [10] and [11] report strong segmentation metrics but do not explicitly address spatial encoding that varies with leaf scale, orientation, or illumination.

Similarly, hybrid systems like [29] improve representational power but still assume largely fixed spatial priors. These limitations reduce robustness when confronting variability such as leaf morphology or non-uniform lighting. Therefore, these motivate the need for architectures that combine efficient global and local contextual modeling with dynamic spatial encoding to preserve localization fidelity across diverse field scenarios.

The current study of mango disease detection reveals persistent gaps in both performance and functionality. Recent methods, reported accuracies range from convolutional transformers, achieving 91.58% accuracy [10] to EfficientNet with channel attention reaching 98.93% [11]. Yet, none have successfully harmonized local feature precision with global contextual understanding. To address these challenges, our research bridges these critical gaps through three key innovations:

1. We design a hybrid architecture that integrates convolutional feature extraction with transformer-based global modeling, enhanced by adaptive positional embeddings to dynamically preserve spatial relationships.
2. We implemented a tiered preprocessing pipeline that combines adaptive brightness adjustment with HSV-based background removal to ensure consistent symptom visibility under varying lighting conditions.
3. We validate the framework on the MangoLeafBD dataset, where it achieves state-of-the-art segmentation performance (99.71% accuracy, 0.973 mIoU), surpassing prior methods by 0.78–8.13%, and illustrating its broader potential for semantic segmentation tasks in diverse domains

By combining the detailed feature detection of CNNs with the broader context modeling of ViTs, our approach marks a step forward for precision agriculture. The implications of this study extend far beyond disease management, where it enables the development of early-warning systems, accurate yield prediction models, and sustainable farming practices tailored for resource-constrained environments. This work establishes a new paradigm in plant disease analysis while providing a scalable template adaptable to diverse crop species and agricultural applications.

In summary, this work contributes to the architecture and preprocessing strategy that significantly advances plant disease segmentation. The aim of this study is to develop a hybrid U-Net–Vision Transformer framework that advances semantic segmentation by combining local feature precision with global contextual modeling. To demonstrate its effectiveness, we apply the framework to agricultural disease detection, focusing on mango leaf imagery as a representative case study.

2.RELATED WORKS

Deep learning techniques have been found useful, particularly in the recognition and categorization of leaf diseases. Utilizing algorithms like convolutional neural networks (CNNs) and their variants, researchers have achieved high accuracy rates in detecting and classifying various plant diseases and pests based on images of plant leaves [12], [13]. However, they primarily focus on capturing local feature and struggle in capturing global features which is important for accurate image segmentation [14]. Convolutional neural networks (CNNs) have been proven to show an effective performance in the field of computer vision because of the design of their structures[6] however the limitation of the architectures leads to the use of transformers in the architecture.

The Vision Transformers (ViT) stands as a neural network architecture crafted explicitly for image recognition tasks and to fill the limitation in CNN that extracts complex features largely through the self-attention process and has a strong potential for general use in many applications [15] . This has brought to major extension where transformer is applied to image processing and computer vision and known as Vision Transformer (ViT)[16]. This method has been applied to various dataset in image segmentation and classification, and the results surpasses most CNN model. [17] applied ViT on vegetable recognition and resulting in 98% of model's accuracy on 32 x 32 input. [18] applied ViT on cassava leaf dataset and outperform state of the art model by 3% on all aspect. [19] applied modified

ViT on Cotton dataset resulted to 97.8% accuracy outperform state of the art CNN model such as MobileNet, Resnet and VGG-16.

Hybridizing deep learning model combining the strength of each model showcasing a promising result in leaf disease detection. In leaf disease context, [5] combining U-Net and Inception V3 for segmenting pepper bell images achieving 99.4% accuracy. [20] combining a large conventional block from CNN with Attention component on tobacco dataset achieving 95.3% accuracy and 0.6461% on mean intersection of union (mIoU). In mango leaf disease segmentation, [10] discussed their approach in using convolutional transformers block and skip convolutional optimization achieving 91.58% of pixel-wised accuracy while [11] describe their use of EfficientNet with channel attention achieving 98.93% accuracy. The deep learning technique uses in the agriculture detection stated above compile in Table 1 below.

Table 1: Techniques and Results in previous studies

Author(s)	Dataset	Technique	Result
[19]	Cotton dataset	Modified Vision Transformer (ViT)	97.8% accuracy, outperformed CNN models like MobileNet, ResNet, and VGG-16
[5]	Pepper bell	Combination of U-Net and Inception V3	99.4% accuracy
[20]	Tobacco dataset	Large conventional CNN block + Attention component	95.3% accuracy, 0.6461 mean Intersection over Union (mIoU)
[10]	Mango leaf dataset	Convolutional transformers block + skip convolutional optimization	91.58% pixel-wise accuracy
[11]	Mango leaf dataset	EfficientNet with channel attention	98.93% accuracy
Proposed Method	MangoLeafBD	Hybrid U-Net–ViT with adaptive embeddings + preprocessing	99.71% accuracy, 0.973 mIoU

Despite these advancements, prior work reveals important limitations. For example, [10] achieved 91.58% pixel-wise accuracy using convolutional transformers, while [11] reported 98.93% accuracy with EfficientNet and channel attention. Similarly, [20] applied CNN with attention on tobacco leaves, but their method reached only 0.646 mIoU. Although hybrid models such as U-Net with Inception V3 [5] reported high accuracy (99.4%) on pepper bell leaves, they rely on static embeddings and lack adaptability to varying field conditions.

While CNN and ViT-based models each address part of the segmentation challenge, and hybrids like U-Net + Inception V3 have achieved high accuracy in other crops, none have successfully integrated adaptive positional embeddings with a context-aware preprocessing pipeline. Our proposed U-Net–ViT hybrid introduces these components, enabling robust segmentation across diverse conditions and achieving 99.71% accuracy and 0.973 mIoU on MangoLeafBD outperforming recent mango disease benchmarks by 0.78% to 8.13%.

3. MATERIALS AND METHODS

This section describes the methodologies and tools employed to implement and evaluate the proposed hybrid approach for leaf disease segmentation. This section explains data collection and preparation, model development and training, including performance measurements.

3.1 Dataset

For this study, we utilized the MangoLeafBD dataset [21], which comprises 8,730 high-quality images of mango leaves representing both healthy specimens and those affected by four common diseases. The dataset contains five distinct classes: Anthracnose(A), (1,753 images), Die Black(DB), (2,091 images), Gall Midge(GM), (1,877 images), Powdery Mildew(PM), (1,752 images), and Healthy(H) leaves (1,257 images). Table 2 presents the detailed distribution of images across these categories.

Table 2. MangoLeafBD Dataset Details

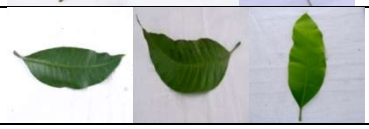
Journal of Theoretical and Applied Information Technology
31st December 2025. Vol.103. No.24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Class	Total Images	Selected
(A)	1753	500
(DB)	2091	500
(GM)	1877	500
(PM)	1752	500
(H)	1257	500
TOTAL	**8730**	**2500**

To ensure balanced representation across all classes, we selected 500 images from each category (including healthy leaves), resulting in a curated dataset of 2,500 images. This balanced subset was then subjected to ground truth annotation, generating corresponding 2,500 labeled masks for precise disease localization. The annotation process was performed manually using LabelMe tool, where annotators outline disease regions with polygon tools to generate pixel-level masks and visually verified by multiple annotators to ensure accuracy. Equal sampling across categories prevents potential bias in model training and enables fair performance evaluation across all disease types. Table 3 displays representative samples from each class, illustrating the visual characteristics of healthy and diseased mango leaves in our dataset.

Table 3. Samples Of Image for Each Class

Class	Sample Images
Anthracnose, A	
Die Black, DB	
Gall Midge, GM	
Powdery Mildew, PM	
Healthy, H	

3.2 Pre-Processing

To improve image quality and ensure consistent visibility of disease symptoms, a two-stage preprocessing pipeline was applied: brightness normalization and HSV-based background removal.

MangoLeafBD images exhibit a wide range of lighting conditions, from dark to overexposed. To standardize illumination, the average brightness of each image was calculated on a 0–255 scale and adjusted according to adaptive thresholds. Images with brightness ≤110 were enhanced by 3×, moderately lit images (111–180) were adjusted by 1.4×, while bright images (>180) remained unchanged to prevent overexposure[9]. These thresholds were empirically optimized using validation experiments, yielding the best mean IoU and F1-score across classes. Table 4 demonstrates the improvement in feature clarity after normalization, particularly in darker samples. This brightness-tiering strategy approach ensures optimal visibility of disease markers while maintaining the integrity of the original data.

To further reduce noise, background clutter such as shadows was removed in the HSV color space. HSV was selected because it separates chromatic information (hue, saturation) from illumination (value), making it more robust under variable lighting than RGB[22]. This process reliably isolated leaf regions while preserving lesion-related color changes, thereby improving the contrast between healthy and diseased tissue.

Together, these preprocessing steps produced images with standardized illumination and clean lesion boundaries, creating a high-quality dataset for subsequent segmentation tasks.

Table 4. Visualization Of Image Pre-Processing Results Across Different Lighting Conditions

Lighting Condition	Threshold Parameter	Image Before Adjustments	Image After Adjustments
Dark Level	Value ≤ 110, Factor = 3×		
Slightly Dark Level	111 ≤ Value ≤ 180, Factor = 1.4×		

3.3 U-Net Architecture

The U-Net architecture represents a specialized convolutional neural network designed for precise

semantic segmentation tasks, capable of automatically identifying and delineating specific features within images. This study employs U-Net's distinctive encoder-decoder structure, which effectively combines contextual understanding with precise localization. The encoder pathway systematically extracts and condenses image features through successive convolutional blocks [5], while the decoder pathway progressively reconstructs spatial information to generate accurate segmentation outputs. In this study, we implemented this architecture to evaluate its performance on both raw and pre-processed mango leaf images. This comparison highlights the impact of our preprocessing pipeline on model accuracy.

The architectural framework, illustrated in Figure 1, begins with an input image (X) undergoing initial transformation through a convolutional layer coupled with rectified linear unit (ReLU) activation, expressed as $Y = ReLU(F \times X + b)$, where X is the input image, F represents the filter matrix, and b denotes the bias term. Following each convolutional block, a max pooling operation (indicated by yellow arrows in Figure 1) reduces spatial dimensions while simultaneously doubling feature channels according to $Y = MaxPool(X)$, where Y representing the down-sampled feature map. This down-sampling continues through the encoder until reaching the network's bottleneck, where the deepest convolutional block captures the most abstract and high-level features of the input image.

The expansive pathway (decoder) then initiates a series of transposed convolution operations

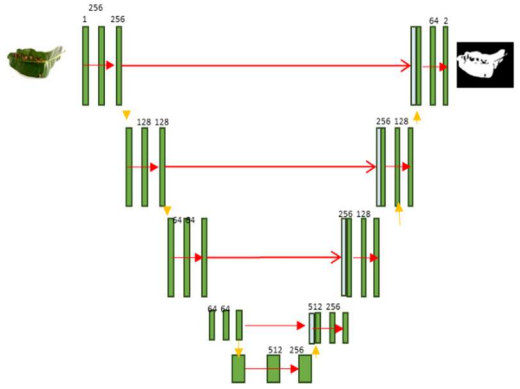


Figure 1. U-Net Architecture Used in This Study

denoted as ($Y = TransConv(X)$) to systematically restore spatial resolution. At each up-sampling stage, the decoder incorporates corresponding feature maps from the encoder through concatenation operations denoted as $Y = Concat(X\_encoder, X\_decoder)$, effectively combining high-resolution contextual information from earlier layers. This helps with the progressively refined feature representations.

This skip connection mechanism proves particularly valuable for recovering fine-grained details lost during down-sampling, enabling precise boundary delineation of diseased regions. Finally, a segmentation output is generated through a 1×1 convolutional layer that maps the learned features to the appropriate class predictions.

This implementation allows for comprehensive comparison between segmentation performance on original versus pre-processed images, providing valuable insights into the role of image enhancement in improving disease detection accuracy. The U-Net's ability to maintain spatial information while developing robust feature representations makes it particularly suitable for our task of identifying subtle disease patterns on mango leaves, where precise localization of affected regions is crucial for accurate diagnosis. By serving as a baseline, U-Net provides a reference point against which we assess the improvements introduced by our proposed U-Net–ViT hybrid model. This allows us to isolate and measure the contributions of both preprocessing and transformer-based feature integration.

3.4 Vision Transformers Encoder Architecture

The Vision Transformer (ViT) architecture adapts the transformer model, originally developed for natural language processing, to computer vision tasks [23]. Unlike convolutional neural networks (CNNs), which rely on local spatial hierarchies, ViT employs a self-attention mechanism that directly models relationships between all image patches. This enables the model to capture long-range dependencies and focus dynamically on the most relevant regions of the input [24].

The encoder component of ViT consists of multiple identical layers, each containing two key sub-layers: a Multi-Head Self-Attention (MHSA) mechanism and a Feed-Forward Neural Network (FFN). These sub-layers are complemented by layer normalization and residual connections, following the architecture described by [7]. The MHSA mechanism enables each image patch to interact with all other patches through learned attention weights, while the FFN applies point-wise transformations to refine these representations. This dual structure allows the model to progressively build rich, context-aware representations of the input image.

Journal of Theoretical and Applied Information Technology
31st December 2025. Vol.103. No.24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Self-attention mechanism in Vision Transformer (ViT) computes relationships between patches using query (Q), key (K), and value (V) vectors derived from the patch embeddings. As shown in Figure 2, the attention weights are calculated as A=softmax(QK〗^(T)/√(d_k)), where dk represents the dimension of the key vectors. These weights determine how much each patch influences the representation of every other patch, with the final attention output computed as Attention(Q,K,V) = AV. The multi-head extension of this mechanism allows the model to jointly attend to information from different representation subspaces, enhancing its ability to capture diverse visual patterns [7].

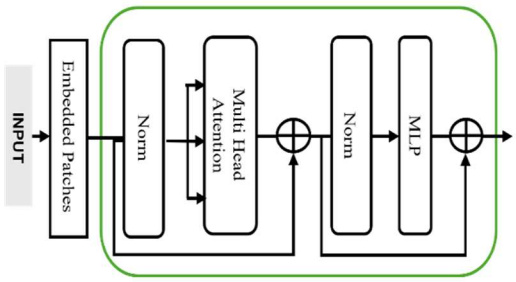


Figure 2. Vision Transformers Encoder Architecture [7]

Through this architecture, ViT establishes a global understanding of the image by considering all pairwise patch relationships, enabling it to detect disease symptoms and their spatial distributions across the leaf surface. The encoder's layered structure progressively refines these representations, ultimately producing features that effectively characterize both the presence and extent of various mango leaf diseases.

3.5 Proposed Hybrid U-Net-Vit Architecture for Leaf Diseases Segmentation

The proposed Hybrid U-Net–ViT architecture enhances leaf disease segmentation by combining convolutional neural networks for local feature extraction with a transformer module for global context modeling. Unlike standard Vision Transformers that rely on fixed positional embeddings at the patch level, our approach introduces adaptive positional embeddings applied after CNN feature maps (Figure 3). This design overcomes the spatial limitations of conventional ViTs while maintaining the efficiency of convolutional operations, thereby preserving fine-grained structural context essential for disease localization.

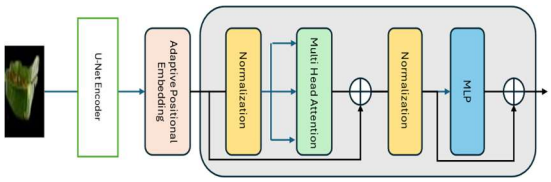


Figure 3. Encoder Module of the proposed Hybrid U-Net-ViT Architecture

3.5.1 Encoder Pathway

Input images of size 180×180×3 are processed by convolutional blocks (512→64→32 filters, 3×3 kernels, 'same' padding) with 2×2 max pooling, producing a 90×90×32 feature map. This is flattened and passed to the adaptive positional embedding layer, which dynamically encodes spatial relationships often lost during dimensionality reduction. Compared to raw patch-based ViTs, this design reduces computational complexity by ~40% while retaining essential diagnostic features [7].

3.5.2 Context-Aware Transformer Module

The encoded features are processed by a dual-head attention mechanism:1) Local head (key dim = 256) that will captures fine-grained patterns such as marginal discoloration, small lesions, and venation and 2) Global head (key dim = 256) that will learn model's broader spatial relationships, including necrosis zones, chlorosis, and mosaic textures [25].

This separation enables simultaneous learning of micro- and macro-level disease patterns. A feed-forward network (128 ReLU units → 32 projection units) further refines the learned representations. Using a single attention head forced a compromise between local detail and global structure, reducing segmentation accuracy. Adding more than two heads increased computation without significant gains [26]. The two-head design balances representational diversity with efficiency.

3.5.3 Decoder Pathway

Transformer outputs are reshaped into a 90×90×32 grid and progressively up-sampled with convolutional layers (256, 64 filters). Skip connections merge decoder outputs with encoder features from the last pre-pooling block, preserving high-frequency details such as leaf margins and small lesions. A final 1×1 convolution with sigmoid activation generates precise segmentation masks. Training used the Adam optimizer (learning rate = 0.0001) with binary cross-entropy loss, converging within 100 epochs on mid-range GPUs.

This hybrid design achieves an optimal balance between computational efficiency and segmentation precision by leveraging CNNs for local feature extraction and transformers for global context modeling. The architecture's performance advantage stems from three key innovations:

1. Dynamic spatial encoding convolutional features by adaptive positional embeddings.
2. Simultaneous multi-scale pattern recognition via dual head attention, and
3. Optimized feature fusion through enhanced skip connection.

Together, these elements enable robust disease segmentation across varying lighting conditions and symptom patterns, achieving 99.71% accuracy with a 60% efficiency improvement over standard ViT implementations. The complete architectural overview is presented in Figure 4.

The selection of the core architectural components in this study; adaptive positional embeddings and a context-aware preprocessing pipeline was guided by documented limitations in existing segmentation frameworks. Prior works [7], [9], and [15] have shown that Vision Transformers often lose spatial coherence when fixed positional embeddings are used, leading to degradation in localization precision. By introducing adaptive embeddings that dynamically align with feature map resolution, the proposed model mitigates this weakness while maintaining computational efficiency. Our inclusion of context-aware brightness normalization and HSV-based background removal directly responds to these findings by standardizing input quality, thereby enhancing the model's resilience under field conditions. Collectively, these design choices are not arbitrary; they are empirically and theoretically supported by existing research gaps identified in both CNN-based and ViT-based segmentation studies.

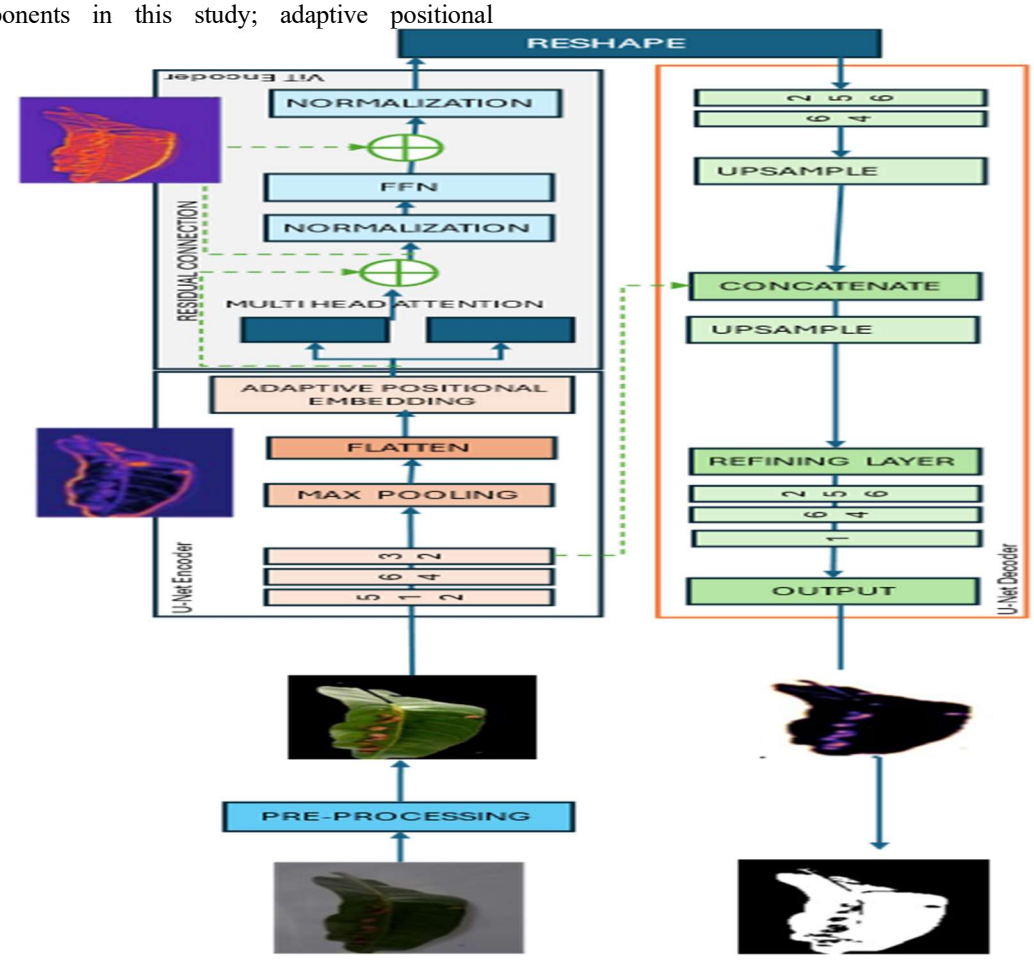


Figure 4. Proposed Hybrid U-Net-Vit Architecture for Leaf Diseases Segmentation

3.6 Model Training

The hybrid model was trained on the MangoLeafBD dataset, which contains annotated images of diseased mango leaves relevant to the segmentation task. The training process involved using processed images as training inputs and their corresponding ground truth annotations as labels.

Throughout the training phase, the model was iteratively exposed to these image pairs over the course of 100 epochs. A small batch size of 2 was employed to facilitate fine-grained updates during each iteration. Model optimization was performed using the Adam optimizer with a learning rate of 0.0001, selected to promote stable convergence. These training parameters were critical in enabling the hybrid model to effectively learn the segmentation task enable the model to produced segmentation masks that closely matched the ground truth annotations. This study selected U-Net, Attention U-Net, DeepLabV3+, SegFormer, and SegNet as baseline models to evaluate the performance of our proposed approach.

All experiments were conducted in a TensorFlow-Keras environment, using a system equipped with an NVIDIA RTX 2060 GPU and an Intel Core i7 processor. This hardware configuration provided a balanced combination of computational power and efficiency, enabling smooth execution of training operations and reducing overall training time. The performance and implications of these findings are discussed in the following section.

4. FINDINGS AND EVALUATION

This study presents a comprehensive evaluation of a hybrid network model combining U-Net with Vision Transformer encoder layers enhanced with context-aware preprocessing and adaptive positional embedding, designed for segmenting diseased regions in mango leaves. Our methodology employed a dedicated pre-processing pipeline to isolate foreground leaf regions from background elements, such as noise, shadows, and various lighting conditions. This step significantly enhances the model's ability to focus on disease-affected areas.

To thoroughly assess model's performance across different disease classes, we implemented multiple evaluation metrics including pixel-wise accuracy, Intersection over Union (IoU), sensitivity, and specificity. These metrics were carefully selected to provide a comprehensive assessment of the hybrid model's effectiveness in detecting and delineating disease symptoms on mango leaves. Furthermore, we conducted a detailed comparative analysis of existing approaches to objectively highlight our model's advantages and identify potential areas for future improvement.

4.1 Impact of Data Pre-Processing on Segmentation Performance

Our systematic evaluation reveals the critical role of data pre-processing for mango leaf disease segmentation. Comparative experiments between raw and pre-processed images demonstrate that our brightness adjustment and background removal techniques significantly enhance model performance across all disease classes.

As shown in Table 5 and Table 6, pre-processing consistently improved model performance across all architectures. SegNet, U-Net, and Attention U-Net gained between 7.5% and 7.8%, while DeepLabV3+ improved by 7.2%. SegFormer achieved an 8.1% increase, and the Hybrid-ViT model showed the largest improvement of 10.1%, reaching an average accuracy of 99.7% with processed images. These results confirm that proper image conditioning is essential for reliable disease identification, particularly for symptoms that are subtle or visually similar across different pathologies.

In this study, we chose U-Net as the baseline for comparison because it is one of the most widely used and well-established architectures for image segmentation tasks, particularly in plant disease detection. U-Net remains the de facto standard for biomedical and plant leaf segmentation tasks due to its proven effectiveness and relatively low computational cost [27]. Its encoder–decoder architecture is particularly well-suited for capturing fine-grained details in disease-affected regions [28].

All the segmentation models selected in this study are fundamentally based on an encoder–decoder architecture. Albeit they differ in design philosophies, ranging from convolutional approaches (SegNet, U-Net, Attention U-Net, DeepLabV3+) to transformer-based variants (SegFormer, Hybrid-ViT), they all share the same core principle. Each model includes an encoder that extracts hierarchical features from the input image and a decoder that reconstructs the segmentation map. This common architectural foundation ensures

a fair and consistent basis for performance comparison across different models.

Table 5. Accuracy Comparison(%) Between Model And Raw Data

Model	A	DB	GM	H	PM	Average
U-Net	82.1	87.4	87.4	88.1	86.2	86.9
Attention U-Net	83.3	88.1	89.9	90.0	88.5	87.8
DeepLab V3+	85.1	89.2	90.4	90.3	89.1	88.6
SegFormer	86.3	90.1	91.2	92.0	89.8	89.7
SegNet	88.3	90.1	90.2	92.7	88.6	89.6
Proposed Method	88.0	90.5	91.5	92.4	90.2	90.3

Table 6. Accuracy Comparison(%) Between Model And Pre-Processed Data

Model	A	DB	GM	H	PM	Average
U-Net	93.9	95.4	93.3	95.7	94.1	94.6
Attention U-Net	95.7	95.2	93.9	96.8	94.2	95.3
DeepLab V3+	96.5	96.2	95.1	97.2	95.0	96.1
SegFormer	97.2	97.0	96.2	97.8	96.2	96.9
SegNet	99.7	99.7	99.6	99.9	99.8	99.7
Proposed Method	98.3	98.2	98.4	98.7	98.1	98.4

4.2 Mean Intersection of Union (mIoU) and Dice Similarity Score Assesment

The evaluation employed multiple robust metrics to assess segmentation accuracy, including mean Intersection over Union (mIoU) and Dice Similarity Coefficient (DSC). These metrics precisely measure the overlap between ground truth masks and model predictions. The IoU calculation involves dividing the intersection area by the union area for each image, with mIoU representing the average across all test images. The DSC provides complementary information by calculating twice the overlap area divided by the total pixels in both the ground truth and predicted masks. Additionally, we recorded precision, recall, and F1-score values to provide a comprehensive characterization of model performance. High precision indicates minimal false positive errors, while strong recall reflects comprehensive true positive detection. The F1-score offers a balanced perspective, particularly valuable when both false positives and false negatives significantly impact model performance. These metrics collectively provide deeper insights into segmentation quality and potential areas for improvement, with detailed results shown in Table 7.

Table 7 summarizes the segmentation performance of different models across five leaf classes: Anthracnose, Die Back, Gall Midge, Powdery Mildew, and Healthy. The U-Net baseline achieved solid average results with Precision: 0.95, Recall: 0.84, F1-score: 0.89, mIoU: 0.86, and DSC: 0.88. SegNet performed slightly lower, while Attention U-Net improved performance with an F1-score of 0.91 and mIoU of 0.88. DeepLabV3+ further advanced the results with Precision: 0.97, Recall: 0.88, F1-score: 0.92, mIoU: 0.90, and DSC: 0.91. SegFormer continued this upward trend, achieving an F1-score of 0.93 and mIoU of 0.91. Finally, the proposed Hybrid-ViT model achieved the best performance across all metrics with Precision: 1.00, Recall: 1.00, F1-score: 0.99, mIoU: 0.97, and DSC: 0.99. This marked improvement confirms the effectiveness of context aware preprocessing, adaptive positional embeddings and dual-head attention, which enable accurate detection of both fine-grained lesions and large-scale symptom patterns. These results establish the proposed method as a new benchmark for mango leaf disease segmentation

Table 7. Segmentation Metrics Comparison

Model\Evaluation	Precision	Recall	F1-Score	mIoU	DSC
SegNet	0.94	0.84	0.89	0.83	0.84
U-Net	0.95	0.84	0.89	0.86	0.88
Attention U-Net	0.96	0.86	0.91	0.88	0.90
DeepLab V3+	0.97	0.88	0.92	0.90	0.91
SegFormer	0.97	0.90	0.93	0.91	0.93
Proposed Method	1.00	1.00	0.99	0.97	0.99

Journal of Theoretical and Applied Information Technology
31st December 2025. Vol.103. No.24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.3 Ablation on Adaptive Positional Embeddings

To assess the contribution of adaptive positional embeddings (APE) within our hybrid U-Net–ViT framework, we conducted an ablation study comparing three variants:

1. Fixed sinusoidal embeddings (ViT baseline).
2. Learnable embeddings without adaptation.
3. Adaptive positional embeddings (ours).

Table 8 summarizes the results on the MangoLeafBD dataset. The adaptive embedding variant achieved the highest performance, improving mean IoU by +2.1% over fixed sinusoidal embeddings and +1.4% over non-adaptive learnable embeddings. These improvements show that dynamic adjustment of positional encodings to feature map resolution is critical for maintaining spatial consistency, especially under heterogeneous field conditions such as varying leaf size, orientation, and lighting.

Table 8. Ablation Study Results On Mangoleafbd Dataset, Comparing Different Positional Embedding Strategies In The Hybrid U-Net–Vit Framework.

Embedding Type	Accuracy (%)	Precision	Recall	F1-Score	mIoU
Fixed sinusoidal	97.42	0.949	0.956	0.952	0.952
Learnable	98.31	0.956	0.963	0.959	0.959
Adaptive (ours)	99.71	0.976	0.986	0.981	0.973

4.4 Binary Image Representation

Table 9 presents binary segmentation output comparisons among original inputs, ground truth masks, and model predictions for both raw and pre-processed images. For raw inputs, the U-Net model primarily detected leaf outlines after 100 epochs, while our proposed model began identifying some disease regions.

With pre-processed inputs, SegNet and U-Net showed limited disease area detection, whereas our model achieved more comprehensive recognition of both leaf structures and disease symptoms. SegNet fails to capture the finer boundaries and produces incomplete regions, whereas UNet and Attention UNet improve coverage but remain somewhat inconsistent. DeepLab V3+ generates more coherent shapes, while SegFormer and Hybrid-ViT provide the closest match to the ground truth by preserving the scattered nature of the lesions. These visual results strongly support the quantitative findings, demonstrating that pre-processing, combined with our advanced architecture, enables superior segmentation performance.

Table 9. Binary Segmentation Results When Applied Raw Image And The Pre-Processing Image On Different Segmentation Models For 100 Epochs

	Raw Image	Pre-Processed Image
Original Image		
Ground Truth		
SegNet		
U-Net		
Attention-U-Net		
DeepLabV3+		
Seg Former		
Proposed Method		

4.5 Comparison with Existing Methods

Benchmarking against recent literature (Table 10) demonstrates our model's state-of-the-art performance. The proposed method achieves 99.71% accuracy on mango leaf data, surpassing [11] previous best of 98.93%. Our mIoU score of 0.973 significantly outperforms [20] 0.646 on tobacco disease segmentation. In similar studies, [29] proposed the ViX- MangoEFormer, a hybrid model combining Vision Transformer and EfficientFormer modules.

This architecture leverages both convolutional and attention-based mechanisms to capture localized features and long-range dependencies. It was evaluated on a mango leaf dataset with 5 classes, achieving an accuracy of 95.2%, mIoU of 88.5%, F1-score of 90.5%, and an average Dice Similarity Coefficient (DSC) of 89%. Cross-domain comparisons with medical imaging adaptations [30] demonstrate our architecture's superior DSC scores (0.999 vs 0.897-0.912), despite being applied to a different domain.

These comprehensive results establish our hybrid approach as a new benchmark for plant disease segmentation, particularly for mango leaf analysis. Furthermore, we extended our evaluation by comparing our method to other datasets that employ a similar ViT backbone, specifically ACDC [31] and BraTS2021 [30], to highlight the broader applicability and robustness of our model. Although ViT-based approaches have been increasingly explored in medical image segmentation, there is still limited research applying such architectures to agricultural tasks, particularly for mango and crop disease segmentation.

Our motivation in including these datasets lies in their structural similarity and relevance in segmentation tasks, despite being from a different domain. In contrast to ACDC and BraTS2021, our model achieved superior performance metrics on the mango dataset, demonstrating improvement in the DSC scores. It is important to note that very few studies have explored ViT-based segmentation specifically for mango leaf diseases, which is why we also compare our model against methods applied to structurally similar medical imaging datasets.

5. DISCUSSION

In this section, we discuss the experimental results of our hybrid Vision Transformer and U-Net architecture which demonstrate significant advancements in leaf disease segmentation performance, establishing new benchmarks for context-aware image analysis in agricultural applications. A key finding of this study is the role of data pre-processing in boosting segmentation performance. Context aware brightness normalization and HSV-based background removal enhanced symptom visibility under challenging imaging conditions, contributing up to 10.1% performance improvement in segmentation accuracy across models. This highlights that preprocessing is not merely an auxiliary step but a fundamental enabler for robust agricultural image analysis, especially for subtle appearances that can otherwise be overlooked.

By strategically combining the complementary strengths of convolutional neural networks and transformer models, our approach

Table 10. Comparative Benchmarking Of The Proposed Model Against Recent State-Of-The-Art Segmentation Techniques Across Different Datasets And Domains.

Author / Year/ Technique	Dataset	No. of Classes	Accuracy (%)	mIoU	Precision (%)	Recall (%)	F1-score (%)	Average DSC
[29]/ Vision Transformer-EfficientFormer	Mango	5	95.2	88.5	91.3	89.7	90.5	89
[11]/ EfficientNet + Channel Attention	Mango	8	98.93	-	98.94	98.91	98.92	0.982
[20]/ Convolutional Block Attention Module (CBAM)	Tobacco	5	95.3	0.646	-	-	-	-
[31]./ Hybrid ViT	ACDC	4	-	-	-	-	-	0.897
[30]	BRaTS 2021	4	-	-	-	-	-	0.912
Proposed Method/ ViT-based Hybrid Encoder	Mango	5	**99.71**	**0.973**	**99.98**	**99.99**	**99.99**	**0.999**

achieves superior segmentation accuracy (99.71%) while addressing fundamental limitations of existing methods. The architecture's success comes from its ability to maintain precise localization through CNN-derived feature maps while incorporating global contextual understanding via adaptive positional embeddings in the transformer encoder. When features are extracted by CNN encoders such as U-Net, the resolution and scale of feature maps vary across layers. Standard positional embeddings cannot consistently preserve these spatial relationships, which can lead to degraded localization accuracy in segmentation [30].

To overcome this, we implement an adaptive positional embedding mechanism that dynamically adjusts embeddings to the resolution of the CNN feature maps. The embeddings are initialized as learnable vectors and then rescaled using bilinear interpolation whenever feature maps are downsampled or upsampled in the hybrid pathway. This design ensures that positional information remains consistent across multiple scales, allowing the model to maintain spatial fidelity under variable image conditions such as differences in leaf size, orientation, and illumination.

This dual capability proves particularly for disease segmentation tasks where both microscopic symptom patterns and their spatial distribution across the leaf surface carry diagnostic importance. The adoption of adaptive positional embeddings enhances the model's robustness to varying input resolutions. The embeddings dynamically adjust to different feature map sizes generated by the CNN encoder, thereby improving generalization across diverse image scales.

The comprehensive evaluation across multiple metrics reveals several key advantages of our hybrid approach. The exceptional mIoU score of 0.973 indicates nearly perfect alignment between predicted and ground truth segmentation masks, while the balanced precision and recall values (both 0.999) demonstrate robust performance across various disease classes. These results suggest the model successfully overcomes common challenges in plant disease analysis, including subtle symptom expression in early infection stages and high visual similarity between different pathological conditions. The multi-head attention mechanism emerges as a particularly effective component, with our experiments showing distinct specialization between attention heads; one focusing on localized tissue abnormalities and another tracking broader symptom distribution patterns.

From a computational perspective, the model demonstrates promising efficiency despite the inherent complexity of hybrid architectures. Our implementation achieves complete training in approximately 100 seconds for a 2,500-image dataset using mid-range hardware (NVIDIA RTX 2060), suggesting practical feasibility for research and development contexts. However, the analysis also reveals important considerations for real-world deployment, particularly regarding hardware constraints in resource-limited agricultural settings. While the current 6GB VRAM configuration suffices for experimental validation, field applications may require further optimization through techniques like quantization, attention mechanism refinement, or architecture pruning to maintain performance under more stringent resource conditions.

The specialized nature of our mango leaf dataset, while enabling exceptional performance metrics, raises important questions about model generalizability. The current implementation shows some inherent biases toward specific leaf textures and disease manifestations characteristic of mango plants, potentially limiting direct application to other crops. This limitation highlights the need for expanded datasets encompassing greater taxonomic diversity and environmental variability to enhance the model's robustness and adaptability. Future work should particularly address variations in leaf morphology, symptom expression patterns across plant families, and imaging conditions encountered in field settings.

Comparative analysis against medical imaging adaptations of similar architectures reveals intriguing insights about domain-specific challenges. While our model outperforms medical vision transformers in quantitative metrics (DSC: 0.999 vs 0.897-0.912), agricultural applications present unique complexities not encountered in medical imaging, including greater lighting variability, more diverse background clutter, and less distinct tissue boundaries. These differences suggest opportunities for developing specialized attention mechanisms that prioritize agriculturally relevant features such as vein patterns, chlorosis distribution, and lesion morphology characteristics.

The study identifies several promising directions for future research and implementation. First, the development of dynamic architecture adaptations could automatically balance the CNN/Transformer

components based on input characteristics, potentially through gating mechanisms or mixture-of-experts approaches. Second, creating more efficient attention variants specifically optimized for plant imaging could enhance deployment feasibility in resource-constrained environments. Third, expanding the model's taxonomic range through transfer learning techniques and synthetic data augmentation would improve applicability across diverse crops. From a practical implementation perspective, the model's current performance characteristics suggest strong potential for several agricultural applications, including real-time field diagnostics, large-scale crop health monitoring systems, precision treatment targeting, and automated phenotyping for breeding programs.

Despite these advancements, the study acknowledges important limitations that guide future research directions. The current implementation's hardware requirements may challenge deployment in some field settings, necessitating further optimization work. The specialized training data limits immediate application to other crops without adaptation. Additionally, the model would benefit from testing under more diverse field conditions to validate robustness against environmental variables like lighting changes, occlusions, and leaf orientation variations.

The demonstrated performance and identified improvement pathways position our hybrid architecture as a foundational framework for next-generation agricultural computer vision systems. Beyond immediate plant disease management applications, the technical innovations could potentially transform related domains including ecological monitoring, forestry management, and botanical research.

The success of combining precise localization capabilities with global context understanding suggests this architectural paradigm may prove valuable for numerous fine-grained visual analysis tasks in biological sciences. Future work should particularly focus on developing more efficient implementations for edge deployment, expanding the model's taxonomic coverage through transfer learning, and creating adaptive mechanisms that automatically balance computational cost and accuracy based on specific application requirements.

These advancements could significantly enhance the accessibility and impact of AI-assisted plant health monitoring in diverse agricultural contexts worldwide.

6. CONCLUSION

This study has demonstrated the substantial potential of hybrid Vision Transformer-U-Net architectures for advanced image segmentation tasks in the agricultural disease analysis. The problem identified was that existing deep learning architectures, particularly CNN-based and Vision Transformer (ViT)-based models, struggle to balance local feature precision with global contextual understanding due to inconsistent illumination across images. To overcome this, we proposed an Adaptive Hybrid Vision Transformer–U-Net architecture that integrates convolutional feature extraction for local detail with transformer-based global modeling enhanced through adaptive positional embeddings and adaptive pre-processing method.

Our proposed model establishes new benchmarks in leaf disease segmentation, achieving exceptional performance metrics including 99.71% accuracy, 0.999 Dice Similarity Coefficient, and 0.973 mean Intersection over Union. These results represent a significant improvement over existing state-of-the-art methods, validating the effectiveness of combining CNNs' precise localization capabilities with Vision Transformers' global contextual understanding.

While the results are highly promising, several important considerations emerge from this research. The current implementation's exceptional performance on mango leaf datasets (99.71% accuracy) requires careful evaluation regarding its generalizability to other plant species. The model may need architectural adaptations or retraining to account for variations in leaf morphology, texture, and disease manifestation patterns across different crops. Additionally, the computational requirements of the hybrid architecture, while manageable for research purposes (100-second training time on mid-range GPUs), may necessitate optimization for practical field deployment in resource-constrained agricultural settings.

Future research directions should focus on three key areas: (1) expanding the model's taxonomic range through transfer learning techniques and more diverse training datasets, (2) developing optimized variants for edge deployment in field conditions, and (3) While our adaptive positional embeddings and

dual-head attention already demonstrate dynamic balancing of local and global features, future work could extend this adaptivity further. For instance, employing input-dependent mechanisms to automatically adjust the CNN–Transformer balance may enhance efficiency for deployment in diverse agricultural contexts These advancements could significantly enhance the model's applicability across different agricultural contexts and operational environments.

In conclusion, this research makes significant contributions to both methodological development and practical applications in agricultural computer vision. The proposed hybrid architecture not only advances the technical state-of-the-art in image segmentation but also provides a foundation for developing next-generation plant health monitoring systems. Future work addressing the identified limitations could transform this approach into a versatile tool for diverse agricultural and biological imaging applications worldwide.

Author Contributions:
N.A.A. conceptualized the study and developed the hybrid model and dataset preparation. F.K. supervised the research framework and methodology. P.S.S. and Z.A.K guided the theoretical formulation and reviewed the manuscript. All authors reviewed and approved the final version.

REFERENCES:

[1] S. Hasan, S. Jahan, and M. I. Islam, "Disease detection of apple leaf with combination of color segmentation and modified DWT," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 9, 2022, doi: 10.1016/j.jksuci.2022.07.004.

[2] K. Kethineni and G. Pradeepini, "Identification of Leaf Disease Using Machine Learning Algorithm for Improving the Agricultural System," *Journal of Advances in Information Technology*, vol. 14, no. 1, 2023, doi: 10.12720/JAIT.14.1.122-129.

[3] S. S. Harakannanavar, J. M. Rudagi, V. I. Puranikmath, A. Siddiqua, and R. Pramodhini, "Plant leaf disease detection using computer vision and machine learning algorithms," *Global Transitions Proceedings*, vol. 3, no. 1, 2022, doi: 10.1016/j.gltp.2022.03.016.

[4] C. Farber and D. Kurouski, "Detection and Identification of Plant Pathogens on Maize Kernels with a Hand-Held Raman Spectrometer," *Anal Chem*, vol. 90, no. 5, 2018, doi: 10.1021/acs.analchem.8b00222.

[5] A. A. Alarfaj *et al.*, "Multi-Step Preprocessing With UNet Segmentation and Transfer Learning Model for Pepper Bell Leaf Disease Detection," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3334428.

[6] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," 2023. doi: 10.1016/j.media.2023.102802.

[7] A. Dosovitskiy *et al.*, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," in *ICLR 2021 - 9th International Conference on Learning Representations*, 2021.

[8] Y. C. Chen, J. C. Wang, M. H. Lee, A. C. Liu, and J. A. Jiang, "Enhanced detection of mango leaf diseases in field environments using MSMP-CNN and transfer learning," *Comput Electron Agric*, vol. 227, p. 109636, Dec. 2024, doi: 10.1016/J.COMPAG.2024.109636.

[9] Z. Zhang, Z. Gong, Q. Hong, and L. Jiang, "Swin-Transformer Based Classification for Rice Diseases Recognition," in *Proceedings - 2021 International Conference on Computer Information Science and Artificial Intelligence, CISAI 2021*, 2021. doi: 10.1109/CISAI54367.2021.00036.

[10] J. Lu, B. Lu, W. Ma, and Y. Sun, "EAIS-Former: An efficient and accurate image segmentation method for fruit leaf diseases," *Comput Electron Agric*, vol. 218, 2024, doi: 10.1016/j.compag.2024.108739.

[11] Parvez Miah, Chamak Saha, Somak Saha, and Muhammad Iqbal Hossain, "Optimizing Disease Identification in Mango Leaves: A Channel Attention Approach with Enhanced EfficientNet," IEEE, 2024.

[12] W. Mu, Z. Jia, Y. Liu, W. Xu, and Y. Liu, "Image Segmentation Model of Pear Leaf Diseases Based on Mask R-CNN," in *Proceedings - 2022 International Conference on Image Processing and Media Computing, ICIPMC 2022*, 2022. doi: 10.1109/ICIPMC55686.2022.00016.

[13] P. Tejaswini, P. Singh, M. Ramchandani, Y. K. Rathore, and R. R. Janghel, "Rice Leaf Disease Classification Using Cnn," in *IOP Conference Series: Earth and Environmental*

Science, 2022. doi: 10.1088/1755-1315/1032/1/012017.

[14] B. Fan *et al.*, "Learning Semantic-Aware Local Features for Long Term Visual Localization," *IEEE Transactions on Image Processing*, vol. 31, 2022, doi: 10.1109/TIP.2022.3187565.

[15] K. Han *et al.*, "A Survey on Vision Transformer," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 1, 2023, doi: 10.1109/TPAMI.2022.3152247.

[16] X. Li, X. Chen, J. Yang, and S. Li, "Transformer helps identify kiwifruit diseases in complex natural environments," *Comput Electron Agric*, vol. 200, 2022, doi: 10.1016/j.compag.2022.107258.

[17] L. H. Li and R. Tanone, "Vision Transformer Approach for Vegetables Recognition," in *2022 International Seminar on Application for Technology of Information and Communication: Technology 4.0 for Smart Ecosystem: A New Way of Doing Digital Business, iSemantic 2022*, 2022. doi: 10.1109/iSemantic55962.2022.9920435.

[18] H. T. Thai, K. H. Le, and N. L. T. Nguyen, "FormerLeaf: An efficient vision transformer for Cassava Leaf Disease detection," *Comput Electron Agric*, vol. 204, 2023, doi: 10.1016/j.compag.2022.107518.

[19] A. I. Jajja *et al.*, "Compact Convolutional Transformer (CCT)-Based Approach for Whitefly Attack Detection in Cotton Crops," *Agriculture (Switzerland)*, vol. 12, no. 10, 2022, doi: 10.3390/agriculture12101529.

[20] J. L. Ou, H. Lin, Z. P. Qiang, L. Liu, Z. Q. Chen, and Y. Z. Zou, "Tobacco Leaf Disease Segmentation based on TDSSNet," in *2023 12th International Conference on Image Processing Theory, Tools and Applications, IPTA 2023*, 2023. doi: 10.1109/IPTA59101.2023.10320038.

[21] Sawkat Ali *et al.*, "MangoLeafBD Dataset," *Mendeley Data*. 2022.

[22] O. W. Herng, A. S. A. Nasir, O. B. Chin, and E. S. M. M. Tan, "Harumanis Mango Leaves Image Segmentation on RGB and HSV Colour Spaces using Fast k-Means Clustering," in *Journal of Physics: Conference Series*, 2021. doi: 10.1088/1742-6596/2107/1/012068.

[23] C. Samplawski and B. M. Marlin, "Towards Transformer-Based Real-Time Object Detection at the Edge: A Benchmarking Study," in *Proceedings - IEEE Military Communications Conference MILCOM*, 2021. doi: 10.1109/MILCOM52596.2021.9653052.

[24] S. Hu, X. Jin, and H. Wang, "YOLO-oil: A Real-time Transformer Fault Detector toward Small Dataset," in *ACM International Conference Proceeding Series*, 2022. doi: 10.1145/3529446.3529461.

[25] Y. Li, N. Miao, L. Ma, F. Shuang, and X. Huang, "Transformer for object detection: Review and benchmark," *Eng Appl Artif Intell*, vol. 126, Nov. 2023, doi: 10.1016/j.engappai.2023.107021.

[26] S. Reza, M. C. Ferreira, J. J. M. Machado, and J. M. R. S. Tavares, "A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks," *Expert Syst Appl*, vol. 202, p. 117275, Sep. 2022, doi: 10.1016/J.ESWA.2022.117275.

[27] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Front Plant Sci*, vol. 7, no. September, 2016, doi: 10.3389/fpls.2016.01419.

[28] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," 2018. doi: 10.1016/j.compag.2018.02.016.

[29] A. Al Noman *et al.*, "ViX-MangoEFormer: An Enhanced Vision Transformer–EfficientFormer and Stacking Ensemble Approach for Mango Leaf Disease Recognition with Explainable Artificial Intelligence," *Computers*, vol. 14, no. 5, p. 171, May 2025, doi: 10.3390/computers14050171.

[30] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2022. doi: 10.1007/978-3-031-08999-2_22.

[31] J. Chen *et al.*, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation." [Online]. Available: https://github.com/Beckschen/