**Agentic AI based Multimodal Framework for Plant Disease Diagnosis Using CLIP and Vision-Enabled LLMs**

Recent advancements in computer vision and deep learning have significantly improved plant disease detection, with Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) achieving state-of-the-art performance. This survey synthesizes findings from key studies, emphasizing major datasets, top-performing models, and accuracy benchmarks. ResNet and DenseNet architectures dominate laboratory-condition datasets, while hybrid CNN-ViT models excel in field conditions.

**Major Datasets in Plant Disease Detection**

1. **PlantVillage Dataset**  
   The PlantVillage dataset remains a widely used benchmark, comprising over 54,000 laboratory-captured images of 38 disease classes across 14 crop species. Its homogeneous background facilitates high accuracy in controlled environments. Top-performing models include ResNet34 (95.3% accuracy), DenseNet201 with a Spotted Hyena Optimizer (98.21% accuracy), and ViT-B/16 (96.8% accuracy). Despite these results, PlantVillage models show limited generalization to real-world conditions.
2. **PlantDoc Dataset**  
   PlantDoc introduces 2,569 field images with complex backgrounds and multiple leaves per image. With 27 disease classes across 13 plant species, models such as Mask R-CNN achieved 83.6% mean Average Precision (mAP), while EfficientNet-B7 attained 81.2% accuracy. Challenges include annotation inconsistencies from non-expert labeling.
3. **FieldPlant Dataset**  
   FieldPlant addresses the limitations of PlantDoc by incorporating 5,170 rigorously annotated field images with 8,629 individual leaves across 27 diseases. YOLOv7 achieved 77.4% mAP, while the Swin Transformer attained 79.1% accuracy by leveraging shifted window-based self-attention. This dataset enhances real-world applicability with occlusion and lighting variations.
4. **AI Challenger 2018**  
   This dataset contains over 10,000 images of 10 crop species spanning 27 disease categories, emphasizing high intra-class variability. Inception-ResNet-v2 achieved 92.4% accuracy, while Xception attained 91.8% accuracy through efficient depthwise separable convolutions.
5. **Rice Disease Dataset**  
   Specialized for detecting rice blast, bacterial blight, and brown spot, this dataset includes 4,200 images with spectral and spatial annotations. U-Net achieved 89.7% segmentation accuracy, while MobileNetV3 optimized for edge devices achieved 87.3% accuracy

**Performance Analysis of Model Architectures**

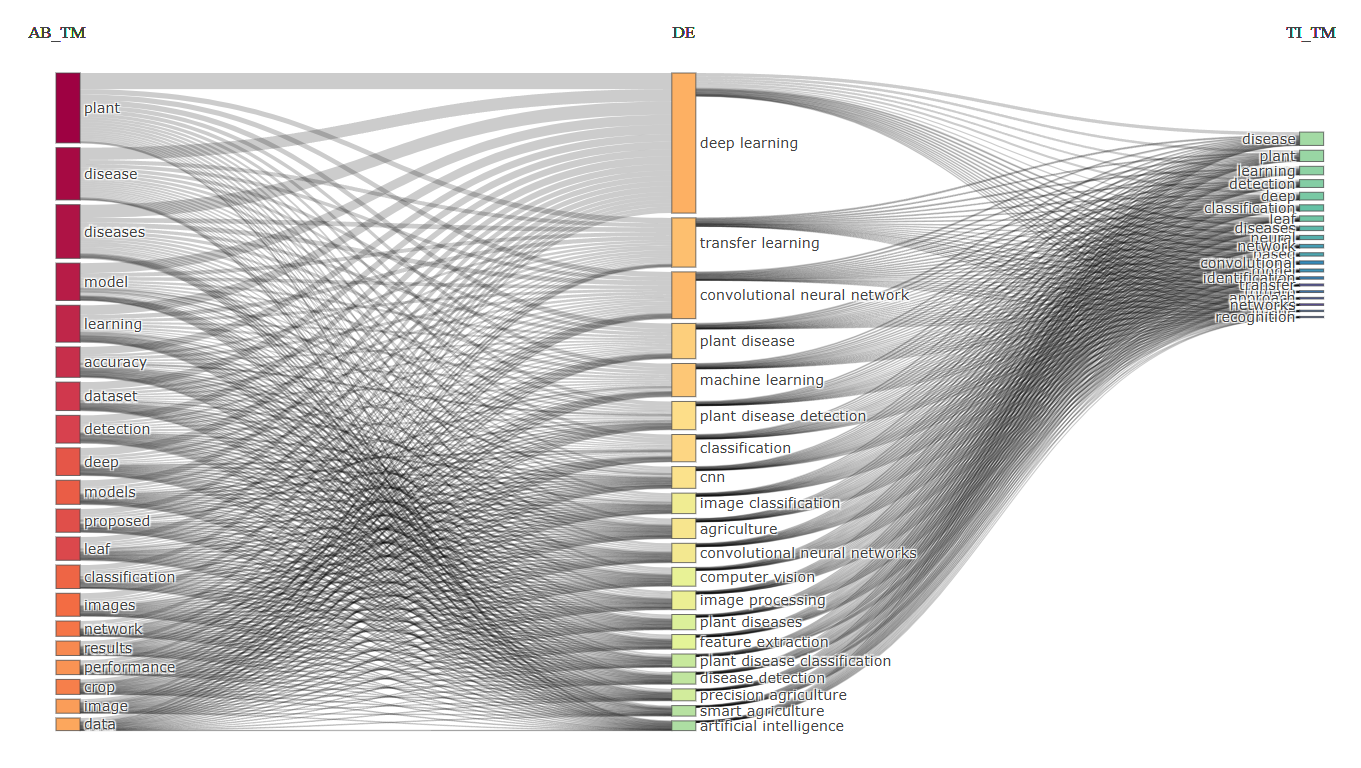
**CNN-Based Models**

* **ResNet Variants:** ResNet50 achieved **93.1%** accuracy on PlantVillage, while ResNet152 improved robustness to image noise with deeper layers. Hybrid ResNet-GAN frameworks improved accuracy by 6.2% on limited-data classes.
* **DenseNet Architectures**: DenseNet201 consistently outperformed ResNet models by 3–5% on multi-class datasets, attributed to enhanced feature reuse through dense connectivity.
* **EfficientNet**: EfficientNet-B4 balanced strong accuracy (90.2%) with parameter efficiency (14M parameters), making it ideal for mobile deployment.

**Vision Transformers**

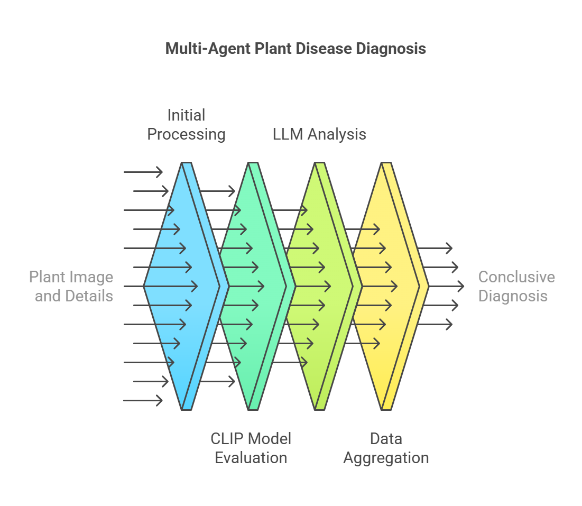
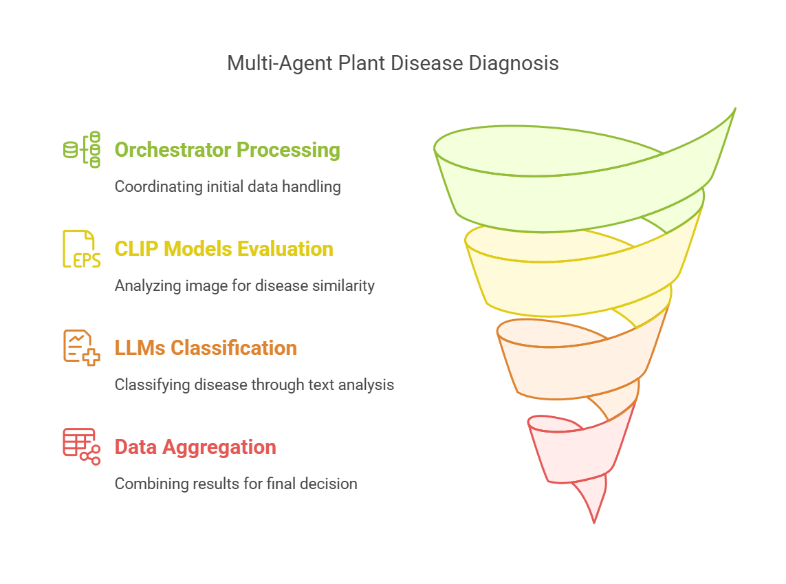
* **ViT-L/16:** Pre-trained on ImageNet-21k and fine-tuned on PlantVillage, ViT-L/16 achieved **96.8%** accuracy, demonstrating superior cross-dataset generalization.
* **Swin Transformer**: With hierarchical feature extraction, Swin-Tiny attained **84.3%** accuracy on FieldPlant, showcasing improved scalability and performance in real-world conditions.

This survey highlights the evolving landscape of plant disease detection, where CNNs excel in controlled settings, while hybrid CNN-ViT models demonstrate superior robustness in field conditions. The integration of transformer architectures continues to improve model performance in complex environments, underscoring the importance of dataset selection and model architecture for effective plant disease detection.



**Methodology**

In our proposed framework, an image of a plant part (e.g., leaf or stem), along with the corresponding part and plant name, is initially processed by an orchestrator agent. This orchestrator subsequently invokes two distinct sets of agents: one comprising multiple CLIP variants and another consisting of several vision-enabled LLM agents. A structured knowledge base, containing a comprehensive list of plants, their associated diseases, and detailed visual descriptions of each disease, is utilized to extract the relevant visual descriptors for the selected plant. The extracted descriptors are then concurrently input to two or three CLIP models alongside the source image, enabling a parallel evaluation that identifies the most strongly matching disease class based on similarity metrics. Simultaneously, the same visual descriptions, embedded within a rigorously structured prompt template, are provided to three or four LLMs to perform a zero-shot classification of the disease depicted in the image. The outputs from both the CLIP models and the LLM agents, accompanied by their respective confidence scores, are subsequently aggregated and forwarded to a final orchestrator LLM. This orchestrator synthesizes the collected data to generate a conclusive diagnosis of the plant disease, thereby ensuring a robust and multi-faceted classification approach.



**Research Problem:**

The absence of suitable training datasets for a large variety of indigenous crops limits the development of effective automated disease detection systems, necessitating novel approaches leveraging zero-shot and few-shot learning paradigms.

**Research Questions:**

1. **Dataset Availability Challenge:** How can the inherent limitation of insufficient training datasets for diverse indigenous crops be addressed through zero-shot classification frameworks?
2. **Knowledge Transfer Efficacy:** To what extent can visual descriptors obtained from domain literature enable effective zero-shot, pure zero-shot, and few-shot classification of plant diseases across taxonomically diverse crops?
3. **Vision-Language Model Application:** What is the optimal methodology for incorporating vision-language models (VLMs) as guidance mechanisms in plant disease detection systems where traditional supervised learning approaches are infeasible?
4. **Transfer Learning Boundaries:** How can we quantitatively assess the generalization capability of zero-shot learning frameworks across different plant families and disease manifestations with minimal or no prior examples?

**References**

1. Exploration of machine learning approaches for automated crop disease detection – Elsevier
2. Leaf-Based Plant Disease Detection and Explainable AI – IEEE
3. Significant plant virus diseases in India and a glimpse of modern disease management technology – Springer

