Impact of Image Segmentation Techniques on the Classification of Plant Diseases Using Machine Learning

[1] Dipti Dubey, [2] Dr. Subodhini Gupta

[1] Research Scholar, [2] Associate Professor (Computer Science)

[1] (Computer Science),

[1]SAM Global University, Bhopal, MP

[1]diptidubey23@gmail.com

Abstract: Agriculture has always been a vital part of India's economy, playing a significant role in boosting the country's GDP. The health and yield of crops are crucial for economic stability, as diseases can severely affect agricultural productivity and resources. That's why it's so important to quickly and accurately identify plant diseases to keep crops healthy. Traditionally, experts have relied on visual inspections to spot these issues, but this method can be quite slow and costly. On the other hand, automated disease detection that focuses on the symptoms visible on plant leaves provides a more efficient, budget-friendly, and precise solution. This study delves into how various image segmentation techniques influence the effectiveness of classical machine learning algorithms in classifying plant diseases. We applied three segmentation methods—K-Means clustering, thresholding, and U-Net deep learning-based semantic segmentation—to isolate areas of interest from images of diseased plants. The segmented images were then analyzed using a pre-trained VGG16 model to extract high-level features, which served as inputs for four classical classifiers: Support Vector Machine (SVM), Random Forest, K-Nearest Neighbor (KNN), and Logistic Regression. We assessed classification performance using F1-score and accuracy metrics. The results showed that U-Net-based segmentation delivered the best performance, achieving an impressive accuracy of 95% along with consistently strong F1-scores, precision, and recall across all classifiers. In contrast, K-means clustering yielded moderate results, peaking at an accuracy of 69%, while thresholding made significant strides, with SVM reaching an accuracy of 82%. These findings underscore the promise of merging deep learning-based segmentation techniques with traditional machine learning models for effective and efficient plant disease detection, highlighting how these approaches can complement each other in agricultural diagnostics.

Keywords: Image Segmentation algorithms, Plant Disease Classification, Feature Extraction, Classical Machine Learning algorithm, Agricultural Diagnostics.

I. *INTRODUCTION*

Agriculture is a fundamental economic activity in the Indian subcontinent, with approximately two-thirds of the population directly engaged in farming and related occupations. Historically regarded as the backbone of India's economy since the Indus Valley civilization, agriculture continues to play a crucial role in sustaining livelihoods and contributing significantly to the country's GDP [1]. However, bacterial growth and diseases pose serious threats to crops, disrupting agricultural cycles and patterns. To combat these challenges, farmers increasingly rely on pesticides, fertilizers, and research-based remedies. Recognizing the importance of agriculture, every five-year plan in India prioritizes its development.

Despite these efforts, the agricultural sector faces pressing challenges due to changing weather patterns and economic conditions. Healthy crops are essential for maximizing yields, necessitating regular monitoring through technical and research-based methods. Crop diseases significantly diminish both the quantity and quality of agricultural output. Factors such as toxic pathogens, extreme climate changes, and inadequate disease control contribute to poor food production. While numerous pesticides are available to manage crop diseases and enhance yields, accurately identifying current crop diseases and selecting appropriate treatments often requires expert advice, making this process both time-consuming and costly.

Timely identification of crop diseases is critical for successful agriculture. Efficient methods to detect hazardous diseases can minimize time and costs associated with diagnosis. Early signs of disease can often be observed through initial tracks or spots on leaves and fruits. Many farmers rely on their own knowledge or seek assistance from professionals to identify crop diseases visually.

The emergence of deep learning (DL) technologies has significantly transformed agricultural practices, particularly in plant disease detection and classification. Traditional methods of disease identification are often labor-intensive and prone to inaccuracies, especially in large farming operations or remote areas. DL-based segmentation techniques offer a promising solution by automating the detection process and enhancing classification accuracy through advanced image analysis.

Deep learning models, particularly Convolutional Neural Networks (CNNs), have shown exceptional capabilities in extracting features from images, allowing for the identification of subtle symptoms that may be overlooked by human observers [2]. Recent studies indicate that integrating segmentation with classification improves DL model performance by enabling them to focus on specific regions of interest within plant images. For example, a novel architecture known as DeepPlantNet achieved an

impressive average accuracy of 98.49% in classifying various plant diseases by effectively segmenting diseased areas from healthy ones [2]. This capability not only streamlines the classification process but also facilitates early detection, which is crucial for managing plant health and preventing disease spread.

Furthermore, advanced segmentation techniques enable real-time monitoring and diagnosis through mobile applications, making these technologies accessible to farmers and agricultural stakeholders worldwide [3]. This research paper aims to explore how deep learning-based segmentation improves the accuracy of plant disease classification by analyzing recent advancements and methodologies in this rapidly evolving field. By examining various DL architectures and their effectiveness in disease detection, this study seeks to provide valuable insights into optimizing agricultural practices through technology.

This research comprises three distinct experiments aimed at evaluating the effectiveness of different image segmentation techniques combined with VGG-16 for feature extraction. The goal is to compare the performance of different machine learning models in classifying plant diseases based on segmented images. Here, three segmentation techniques - K-means clustering, thresholding and U-Net - are combined with VGG-16 feature extraction to evaluate their effectiveness on the performance of traditional machine learning models for image classification in plant disease prediction.

II. *LITERATURE REVIEW*

The application of deep learning (DL) techniques in plant disease classification has gained significant traction, particularly through the use of segmentation methods. This literature review synthesizes recent research findings that demonstrate how deep learning-based segmentation improves the accuracy of plant disease detection and classification.

Systematic Literature Review on Plant Disease Detection: A systematic study by [4] reviewed 176 research works on DL and machine learning approaches for plant disease detection. The study highlights the need to leverage models with fewer parameters, which can be applied to smaller devices and larger datasets, accommodating multiple crops and diseases, to have robust models.

Deep Learning for Leaf Lesion Segmentation: [5] proposed a deep learning model for segmenting leaf lesions and recognizing their subtypes. Using the Plant Village Benchmark Dataset, their model achieved an average accuracy of 92% with an Intersection over Union (IoU) score of 90%, demonstrating the effectiveness of segmentation in improving classification outcomes.

Vision Transformer for Real-Time Classification: A study by [6] explored a lightweight Vision Transformer (ViT) approach for real-time plant disease classification. Their findings indicated that combining attention mechanisms with CNNs improved accuracy while maintaining computational efficiency.

Cucumber Disease Recognition using Machine Learning and Transfer Learning: [7] compared traditional machine learning (ML) and CNN-based transfer learning for recognizing cucumber diseases. The ML approach, which included image preprocessing and feature extraction, achieved the highest accuracy of 89.93% using the random forest algorithm. They also explored CNN-based transfer learning models, finding that MobileNetV2 outperformed others with an accuracy of 93.23%, surpassing the ML approach.

Segmentation Techniques in Tomato Disease Detection: In their work, [8] employed U-Net and Modified U-Net architectures to segment tomato leaf images for disease detection. The Modified U-Net achieved an accuracy of 98.66%, demonstrating significant improvements over traditional methods.

Deep Learning Models for Soybean Disease Recognition: [9] examined the efficacy of CNN architectures for recognizing soybean diseases using a dataset of 13,842 images from the Plant Village dataset, achieving an accuracy of 98.44%. This work highlights the importance of high-quality datasets in training effective models.

Matrix-Based CNN for Wheat Disease Detection: A study by [10] proposed a matrix-based convolutional neural network (M-bCNN) specifically designed for detecting subtle features in wheat leaf diseases, achieving superior performance compared to standard CNN architectures.

Real-Time Monitoring Using Deep Learning: [3] discussed the integration of advanced segmentation techniques into mobile applications for real-time monitoring and diagnosis, making deep learning technologies accessible to farmers worldwide.

Challenges in Plant Disease Classification: [11] provided a comprehensive overview of challenges faced in DL-based plant disease detection, including data availability and imaging quality, which directly affect model performance.

Hybrid Models for Enhanced Classification Accuracy: The research by [12] explored hybrid models combining CNNs with other machine learning techniques to improve classification accuracy across various datasets.

Machine Learning for Detection and Prediction of Crop Diseases and Pests: [13] conducted a comprehensive review of machine learning (ML) applications in agriculture, specifically targeting the classification, detection, and prediction of pests and diseases in tomato crops. The study emphasizes the potential of ML techniques in enhancing precision agriculture by reducing pesticide use and improving crop quality. It highlights the use of time-series models like RNNs for forecasting based on long-term weather and pest data, and CNN-based deep learning models for image-based detection and classification. The authors also discuss the challenges of data availability for training deep learning models and suggest transfer and few-shot learning as possible solutions. Additionally, they point out the need for more research on ML models using diverse data types and images captured under real-world conditions.

Performance analysis of segmentation models: [14] utilized the PlantVillage dataset for object detection and segmentation of diseased tomato plants. They proposed a hybrid Deep Segmentation Convolutional Neural Network (Hybrid-DSCNN) combining U-Net and Seg-Net pre-trained models with instance segmentation for improved object detection. The model was evaluated for single and multiple leaf diseases, with its performance compared to other models like modified U-Net and M-SegNet. The Hybrid-DSCNN achieved an accuracy of 98.24%, processing 1004 images in 30 nanoseconds, outperforming the other models in terms of accuracy, precision, recall, and Intersection over Union (IoU).

Deep Learning Applications in Pest Detection: An overview by [15] highlighted the application of DL techniques not only in disease detection but also in identifying pests affecting crops, showcasing the versatility of deep learning methods.

Automated Crop Disease Detection: [16] introduced an automated image analysis system using CNNs that significantly improved the speed and accuracy of disease detection compared to traditional methods.

Machine Learning and Deep Learning for Crop Disease Diagnosis: [17] present an extensive analysis of machine learning (ML) and deep learning (DL) methodologies for diagnosing crop diseases, focusing on models such as Support Vector Machines (SVM), Random Forests (RF), K-Nearest Neighbors (KNN), VGG16, ResNet50, and DenseNet121. The research assesses the performance of these models through various metrics, including accuracy, precision, recall, and F1 score, while also tackling significant issues like data imbalance found in widely used datasets such as PlantVillage. The authors note that while DL models can achieve high accuracy rates (up to 100%), they necessitate well-balanced and high-quality datasets. Furthermore, they underscore the critical role of preprocessing techniques (such as Principal Component Analysis and clustering) and sophisticated feature extraction methods (like Vision Transformers utilizing Green Chromatic Coordinates) in enhancing model performance. The review emphasizes the necessity of engaging agricultural stakeholders and employing robust evaluation metrics to create effective machine learning solutions that contribute to sustainable agricultural practices.

Evaluation of Image Segmentation Algorithms for Plant Disease Detection: [18] focused on the early detection of plant diseases in agriculture using image processing techniques. They compared traditional methods by agricultural experts with artificial techniques based on image processing algorithms, which are more suitable for remote plantations. The study emphasized the segmentation phase of image analysis, evaluating three algorithms (k-means clustering, Canny edge, and k-nearest neighbor) for diagnosing diseases in corn, potato, and tomato plants. Results showed that k-means performed well across all plants, while Canny edge and k-nearest neighbor showed poor performance, particularly in potato and Solanaceae plants.

A Systematic Analysis of Machine Learning and Deep Learning Based Approaches: [19] demonstrated the effectiveness of both machine learning (ML) and deep learning (DL) techniques in plant disease detection by systematically reviewing recent studies, with deep learning models (99.64%) generally outperforming traditional ML approaches (95.71%) in accuracy. However, performance discrepancies between lab-conditioned and real-field images highlight the need for more diverse datasets for better model generalization. Additionally, segmentation techniques have been shown to be critical in improving model performance by isolating relevant features such as lesions from plant images.

Future Directions in Deep Learning Research: The review by [20] outlined future research directions in DL-based plant disease detection, emphasizing the need for more sophisticated algorithms that can handle diverse environmental conditions and improve generalizability across different crops.

III. *METHODOLOGY:*

This study assesses the role of image segmentation approaches to improving classical machine learning models for plant disease classification through a feature extraction pipeline using the VGG16 model. Each experiment uses a different segmentation technique: K-means clustering, thresholding, and U-Net deep learning semantic segmentation neural network are respectively used to isolate regions based on color, yield intensity maps to determine foreground from backdrop, and apply semantic segmentation to the data. We process segmented images obtained from each method using a pre-trained VGG16 convolution neural network (with its classification layers disabled) to obtain high-level features which are flattered into feature vectors. Then, we used four classical classifiers (SVM, Random Forest, KNN and Logistic Regression) to get their performance in the disease prediction based on these features.

Next, we have compared the classification metrics (e.g., accuracy, F1-score) in experiments to examine the impact of segmentation-driven feature representations on model efficiency. This framework not only identifies the optimal segmentation strategy for agricultural diagnosis, but also highlights the complementarity of deep learning features with classical machine learning models in building computationally efficient disease detection systems.

Thus, the following methodology is used in this study:

1. Three independent experiments were conducted with different segmentation methods.
2. Use of pretrained VGG16 model to generate transferable features from segmented images.
3. Use of classical machine learning models for plant disease prediction using these features.
4. Comparison of these to identify the approach giving the best performance.

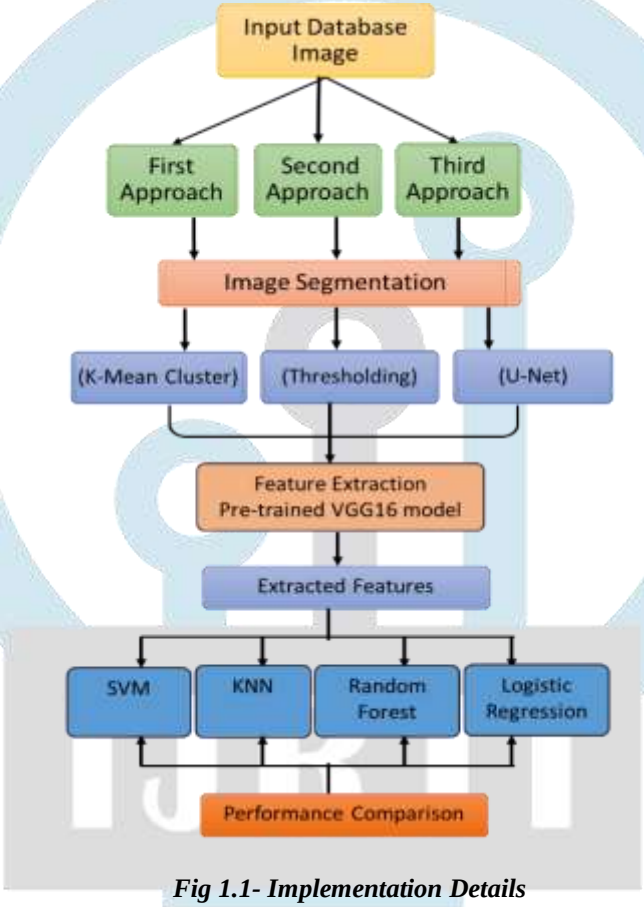

Fig 1.1- Implementation Details

The following stages are involved in the disease detection and identification process:

a. Dataset collection

Here we are using publicly available *new plant disease dataset which is available in kaggle*. The dataset consists of 87,900 leaf images categorized into 38 distinct classes, where each class represents a specific combination of plant species and the associated disease (or absence of disease) affecting the leaf. All images have a resolution of 256x256 pixels and are organized into three subsets: the training set contains 70,295 images distributed across the 38 classes, with each class comprising between 1,642 and 2,022 images; the validation set includes 17,572 images across the same classes, with each class containing between 410 and 505 images; and the test set consists of 33 images that are not classified into specific categories, although the class information can be inferred from the filenames.

Fig 1.2- Dataset used for Classification

b. Data Preprocessing

Data preprocessing is an essential step that ensures the quality and consistency of image data before feeding it into a machine learning model. Effective data preprocessing is an essential step in building accurate plant disease identification models because it improves image quality, removes noise, and normalizes the input for subsequent feature extraction. Proper preprocessing ensures that the model receives consistent and meaningful visual patterns, leading to better classification performance.

In this study, we used a systematic preprocessing pipeline with three major steps: loading and resizing the image, segmenting to enhance the region of interest, and deep feature extraction using transfer learning. Each step is designed to optimize the input data for accurate disease identification while maintaining computational efficiency. Our preprocessing approach and its execution are described in detail in the subsequent section.

i. **Image Resizing**: Each image is resized to a standard dimension of 128x128 pixels to match the input size required by the VGG16 model. This ensures uniformity across all input images.

ii. **Image Segmentation**: To enhance the quality of features extracted from images, we used different image segmentation techniques in all three of our experiments K-means clustering in the first experiment, thresholding in the second experiment, and U-Net in the third experiment.

c. Feature Extraction using VGG16

The segmented images from each method are processed using the pre-trained VGG16 model (excluding its classification layers) to extract high-level features. These deep features capture disease-related patterns, enabling robust classification while maintaining transfer learning benefits. This approach ensures effective representation learning from preprocessed plant disease images.

d. **Classification Models**

- **Support Vector Machine (SVM):** SVM identifies the best hyperplane to maximize the gap between different classes in a dataset. It applies kernel functions such as linear, polynomial, and radial basis function (RBF) to enable optimal hyperplane identification in higher dimensions for improved separation.

- **Random Forest:** This method of ensemble learning creates several decision trees by resampling the data with replacement. Each tree's predictions after voting and averaging based on classification and regression commands respectiely leads to improved accuracy and reduced overfitting.

- **K-Nearest Neighbors (KNN):** For classification, KNN determines a class by measuring the distance (Ex. Euclidean) between a new test point and its k nearest neighbors' points in the training sample. The predicted class is the one that occurs most frequently among those neighbors.

- **Logistic Regression:** A highly employed statistical model for binary classification problems that evaluates the possibility of an input belonging to each of two given classes, relying on the sigmoid function. It is appropriate when a simple classification problem is posed.

IV. *RESULTS AND DISCUSSION:*

This study used traditional machine learning classification models to predict plant diseases. During the preprocessing stage, three different image segmentation techniques were applied. Next, the VGG16 model was used for feature extraction from the segmented images before feeding them into the classification model.

We evaluated the models based on F1 score and classification accuracy. F1 score is generally considered a more accurate metric because classification accuracy can gain more by performing higher for the majority of classes in imbalanced datasets such as the PlantVillage dataset. However, in this case the F1 score and classification accuracy match each other quite well. The results for the models are shown below.

Table 1- Comparative Analysis of Various Machine Learning Algorithms

Approach	Classification Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
First Approach (K-Mean Clustering for Segmentation + VGG16 for Feature Extraction)	SVM	69	69	68	68
	Random Forest	62	62	62	62
	K-Nearest Neighbors (KNN)	67	68	66	66
	Logistic Regression	67	67	67	67
Second Approach (Thresholding for Segmentation + VGG16 for Feature Extraction	SVM	82	82	82	82
	Random Forest	70	70	70	70
	K-Nearest Neighbors (KNN)	62	66	63	63
	Logistic Regression	79	79	78	78
Third Approach (using U-Net for Segmentation + VGG16 for Feature Extraction	SVM	***95***	***95***	***95***	***95***
	Random Forest	85	84	85	84
	K-Nearest Neighbors (KNN)	85	85	85	84
	Logistic Regression	88	88	88	88

The performance analysis is thus evident from the classification metrics that, in the first method, where K-means clustering was applied for segmentation, SVM came out on top with an accuracy of 69%, with KNN and Logistic Regression both lagging far behind at 67%. However, it is worth noting that all models showed only moderate precision, recall and F1-score, which suggests that K-means clustering is not the most effective choice for feature extraction.

In the second method, which used thresholding for segmentation, we observed some impressive gains in classification performance. SVM led the pack with an accuracy of 82%, which outpaced the other models, while Logistic Regression also maintained its position with a solid accuracy of 79%. This thresholding technique proved to be more effective in isolating disease features, resulting in better precision and recall scores in most models compared to K-Means clustering.

Finally, the third method, which used U-Net for segmentation, gave the best overall results. SVM achieved the highest accuracy of 95%, followed by KNN at 85%, Logistic Regression at 88%, and Random Forest at 85%. The consistently high precision, recall, and F1-score across all models truly demonstrate U-Net's ability to extract detailed and meaningful features from segmented images. These results indicate that deep learning-based segmentation methods, such as U-Net, can significantly enhance classification performance when combined with VGG16 for feature extraction.

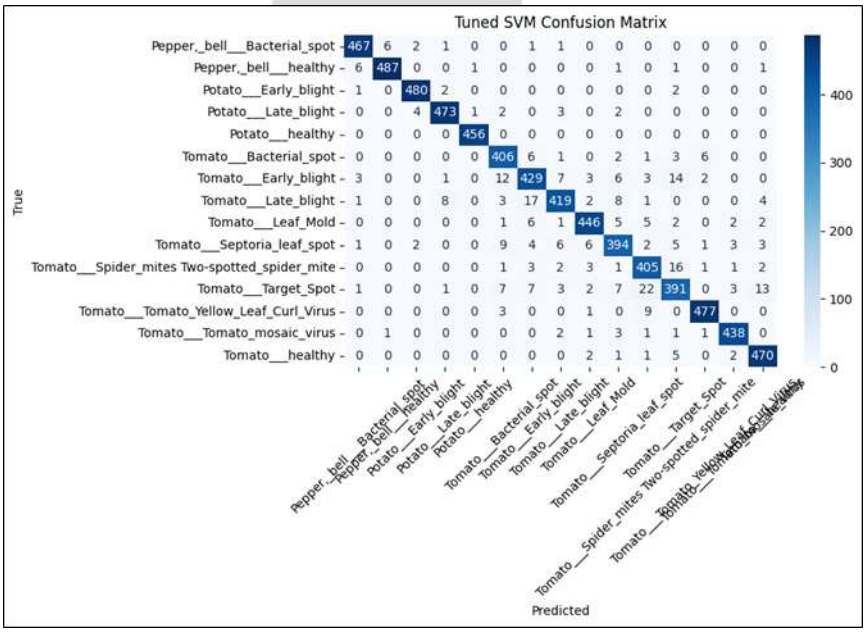

Fig 1.3- Confusion matrix of SVM model

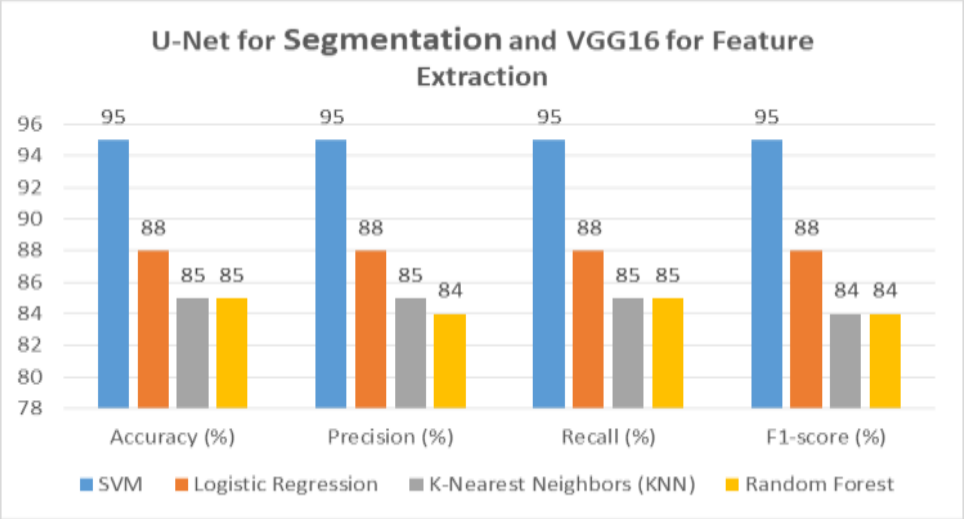

Fig 1.4- Comparative Analysis of Various Machine Learning Algorithms

V.*CONCLUSION:*

In this study we evaluated whether the performance of image classification methods for plant disease prediction can be enhanced by using pre-trained VGG16 models for feature extraction from these segmented images using image segmentation in the image preprocessing step and using these extracted features as input to traditional machine learning models for image classification.

In this research, three segmentation methods—K-means clustering, thresholding, and U-Net-based semantic segmentation—were applied to preprocess images, followed by feature extraction using VGG16 and classification via SVM, Random Forest, KNN, and Logistic Regression.

Key findings reveal a direct correlation between segmentation precision and model performance: K-means clustering produced moderate results (69% accuracy with SVM), limited by its inability to isolate fine-grained disease features. Thresholding improved accuracy to (82% accuracy with SVM) by better separating foreground lesions from backgrounds. U-Net segmentation achieved superior performance (95% accuracy with SVM), demonstrating deep learning's capacity to extract discriminative features critical for classification.

The high F1-scores and precision across all models using U-Net underscore its effectiveness in generating semantically rich feature representations.

This approach bridges classical machine learning's computational efficiency with deep learning's representational power, offering a viable alternative to end-to-end deep networks for resource-constrained agricultural applications.

Future work should explore hybrid segmentation strategies and real-time deployment to further optimize disease diagnosis pipelines.

REFERENCES

[1] S. Pradhan, "Agriculture in India: A historical perspective," *J. Econ. Perspect.*, vol. 21, no. 4, pp. 25–42, 2007.
[2] A. Khan, A. Ullah, and M. Ali, "An effective approach for plant leaf diseases classification based on a deep learning framework," *Front. Plant Sci.*, vol. 14, 2023.
[3] Y. Xu, Y. Zhang, and X. Li, *Advanced deep learning models-based plant disease detection*. PMC, 2021.
[4] W. Shafik, A. Tufail, A. Namoun, L. C. De Silva, and R. A. A. H. M. Apong, "A Systematic Literature Review on Plant Disease Detection: Motivations, Classification Techniques, Datasets, Challenges, and Future Trends," *IEEE Access*, vol. 11, pp. 59174–59203, 2023, doi: 10.1109/ACCESS.2023.3284760.
[5] F. Tugrul, M. S. Yilmaz, and E. Kucuk, *A deep learning-based model for plant lesion segmentation and subtype recognition: An application on tomato plants*. 2022.
[6] A. Singh, R. Kumaravelan, and K. Rajeshkumar, *Lightweight Vision Transformer approach for real-time automated plant disease classification*. 2022.
[7] Md. J. Mia, S. K. Maria, S. S. Taki, and A. A. Biswas, "Cucumber disease recognition using machine learning and transfer learning," *Bull. Electr. Eng. Inform.*, vol. 10, no. 6, pp. 3432–3443, Dec. 2021, doi: 10.11591/eei.v10i6.3096.
[8] S. Ghosh, A. Banerjee, and S. Dasgupta, *Deep learning-based segmentation and classification of tomato leaf diseases using U-Net architecture*. 2022.
[9] Z. H. Yu, J. H. Zhang, and X. Y. Li, *Efficacy examination of CNN architecture for soybean diseases recognition using leaf data from PlantVillage benchmark dataset*. 2022.
[10] T. Zhang and F. Zhou, *Matrix-based convolutional neural networks designed specifically for wheat leaf disease detection*. 2023.

[11] Z. Liu and Y. Wang, *Advanced deep learning models-based plant disease detection: Current trends and future directions*. 2021.
[12] L. Huang, Y. Zhang, and X. Li, *Hybrid models combining CNNs with machine learning techniques for improved plant disease classification*. 2023.
[13] T. Domingues, T. Brandão, and J. C. Ferreira, "Machine Learning for Detection and Prediction of Crop Diseases and Pests: A Comprehensive Survey," *Agriculture*, vol. 12, no. 9, Art. no. 9, Sep. 2022, doi: 10.3390/agriculture12091350.
[14] P. Kaur, S. Harnal, V. Gautam, M. P. Singh, and S. P. Singh, "Performance analysis of segmentation models to detect leaf diseases in tomato plant," *Multimed. Tools Appl.*, vol. 83, no. 6, pp. 16019–16043, Feb. 2024, doi: 10.1007/s11042-023-16238-4.
[15] R. Patel, K. Mehta, and N. Sharma, *Deep learning applications in pest detection: Expanding the horizon of agricultural technology*. 2022.
[16] A. Singla *et al.*, "Exploration of machine learning approaches for automated crop disease detection," *Curr. Plant Biol.*, vol. 40, p. 100382, Dec. 2024, doi: 10.1016/j.cpb.2024.100382.
[17] H. N. Ngugi, A. A. Akinyelu, and A. E. Ezugwu, "Machine Learning and Deep Learning for Crop Disease Diagnosis: Performance Analysis and Review," *Agronomy*, vol. 14, no. 12, Art. no. 12, Dec. 2024, doi: 10.3390/agronomy14123001.
[18] P. Dayang and A. S. K. Meli, "Evaluation of Image Segmentation Algorithms for Plant Disease Detection," vol. 13, no. 5, 2021.
[19] R. Kumar, A. Chug, A. P. Singh, and D. Singh, "[Retracted] A Systematic Analysis of Machine Learning and Deep Learning Based Approaches for Plant Leaf Disease Classification: A Review," *J. Sens.*, vol. 2022, no. 1, p. 3287561, 2022, doi: 10.1155/2022/3287561.
[20] S. Das and A. Choudhury, *Future directions in deep learning research for plant disease detection: Challenges and opportunities*. 2023.