Advances in Consumer Research
https://acr-journal.com/
Volume-2 | Issue-6 |December: 2025
Original Researcher Article

A Novel Zero-Shot Framework For Plant Disease Recognition Using Contrastive Language–Image Pre-Training (Clip)

Shivansh Mehta[1], Savee Gupta[2]
[1]Assistant Professor, ABES Engineering College Ghaziabad,
Email ID : shivanshmehta31@gmail.com
[2]Assistant Professor, ABES Engineering College Ghaziabad,
Email ID : saveegupta672@gmail.com

ABSTRACT
The ability to recognize whether a plant is afflicted with a disease is of the utmost importance in the maintenance of food security on a global scale, the improvement of crop output, and the facilitation of agricultural methods that are sustainable. Due to the fact that conventional deep learning algorithms for the diagnosis of plant diseases rely primarily on big, labeled datasets, they are not as successful in real-world situations when annotated samples are limited or novel disease variants appear. This research suggests a brand-new zero-shot framework for the detection of plant diseases that is based on the Contrastive Language–Image Pre-training (CLIP) paradigm in order to solve these limitations. By utilizing the potent cross-modal learning capabilities of CLIP, the framework is able to correlate the visual characteristics of plant leaves with natural language descriptions of illnesses, which in turn facilitates precise classification without the need for labeled plant pathology datasets or task-specific training. In order to assess similarity within CLIP's joint embedding space, leaf pictures were coupled with prompt-engineered language descriptions of plant illnesses in this study. The model's zero-shot performance over a wide range of climatic circumstances and crop kinds was evaluated by running it on benchmark and real-world plant disease datasets. The findings reveal that the suggested CLIP-based technique is able to accomplish robust generalization, and it performs better than a number of standard supervised and transfer-learning models in situations when there is a limited amount of data. The framework demonstrates a significant capacity to recognize disease classifications that have not been observed before, manage fluctuations in lighting conditions, and adjust to a wide range of backdrops that are characteristic in agricultural settings. The ability of CLIP's zero-shot learning to decrease the need for huge annotated datasets, speed illness detection, and facilitate the creation of scalable digital farming systems is demonstrated in this work, which also emphasizes the promise of vision–language models in the field of agricultural artificial intelligence. By providing a more adaptable and effective approach to the identification of plant diseases at their earliest stages, the framework that has been suggested creates new opportunities for utilizing foundation models in precision agriculture…
Keywords: *Zero-Shot, Plant Disease, CLIP, environments*
.

1. INTRODUCTION:

Plant diseases are a significant risk to agriculture across the world because they have a direct impact on crop productivity, quality, and the safety of food supplies. According to the Food and Agriculture Organization (FAO), plant diseases are responsible for up to forty percent of the yearly crop losses that occur throughout the world. This results in significant economic and social ramifications in both poor nations and industrialized ones. The prompt treatment of plant diseases, the reduction of production losses, and the facilitation of sustainable agricultural management are all dependent on the early detection and correct diagnosis of plant diseases. For the past several years, Artificial Intelligence (AI) and computer vision have developed as significant tools for the detection of plant diseases. These techniques enable automation, improved accuracy, and the ability to do quick diagnostics. However, despite the significant progress that has been made, the majority of traditional AI-based plant disease detection systems are primarily dependent on large-scale picture datasets that have been tagged and are often curated under controlled settings. Models like as convolutional neural networks (CNNs) and its derivatives show great performance on benchmark datasets; but, when used in real-world applications, these models encounter major restrictions. These limitations include the inability to distinguish new or unusual illness classes without retraining, the presence of environmental variability, the presence of dataset bias, and poor generalization to diseases that have not been encountered before. It is time-consuming, expensive, and frequently impracticable to collect and annotate big datasets for each crop species and disease variation. This is especially true in agricultural locations that are limited in their resources. Zero-shot learning is a paradigm that enables models to detect novel categories without explicit training examples. In order to overcome these issues, the academic community has increasingly turned toward zero-shot learning. For the purpose of directing the recognition

process, zero-shot learning makes use of semantic information, which may include textual descriptions, characteristics, or class labels. Conventional zero-shot techniques, on the other hand, frequently rely on manually created feature representations or constrained semantic embeddings, which limits their scalability and effectiveness over a wide range of agricultural contexts. Research on multimodal artificial intelligence has reached a critical milestone with the introduction of Contrastive Language–Image Pre-training, often known as CLIP. Learning a unified embedding space is accomplished via CLIP, which was developed as a large-scale vision–language model. This is accomplished through training on millions of image–text pairs sourced from the internet. Because of this alignment, CLIP is able to match photos with the natural language descriptions that correlate to them. This gives it the ability to do zero-shot classification on a broad variety of tasks without the need for task-specific fine-tuning. Because of its tolerance to domain shifts, its flexibility, and its high generalization skills, it is a good option for applications in agricultural diagnostics. This research presents a unique zero-shot framework for plant disease identification using CLIP which addresses the fundamental constraints of standard supervised learning systems. These strengths are taken into consideration while the research is conducted. The framework enables CLIP to identify illnesses even when there are no labeled pictures of those diseases available. This is accomplished by developing domain-specific textual prompts that describe visual disease signs. Some examples of these symptoms include the presence of chlorosis, necrotic areas, leaf curling, blight patches, or mold development. Through the utilization of this technique, the requirement for annotated datasets is significantly diminished, and the model's capacity to generalize over a wide range of crops, disease kinds, and environmental circumstances is improved. The paradigm that has been suggested makes a number of significant contributions to the field of plant pathology research. In the first place, it illustrates how vision–language models may be modified for use in agricultural applications by utilizing rapid engineering and semantic symptom descriptions. Secondly, it sheds light on the possibility of zero-shot learning to alleviate the problem of a lack of datasets, which would allow for quick scaling across a variety of crop species without required costly retraining. Thirdly, it offers insights on the behavior and limits of foundation models such as CLIP when they are used to domain-specific tasks such as the classification of plant diseases. This work demonstrates that the CLIP-based zero-shot system performs favorably with supervised CNN models and has a good capability to detect illnesses that have not yet been observed. This was accomplished through rigorous assessment on numerous plant disease datasets, which included pictures collected in the laboratory as well as samples taken from the same environment. The findings highlight the significance of multimodal learning frameworks in agricultural artificial intelligence and open the door to the development of plant disease monitoring systems that are scalable, efficient, and highly practical. These systems would be useful for farmers, agronomists, and precision agriculture technologies. In conclusion, the implementation of CLIP-based zero-shot learning in plant disease detection is a big step forward in the process of building solutions for digital agriculture that are both resilient and efficient in terms of data. This framework presents a revolutionary strategy that decreases dependency on big labeled datasets while retaining high accuracy and generalizability. It does this by bridging the gap between natural language descriptions and visual plant symptoms.

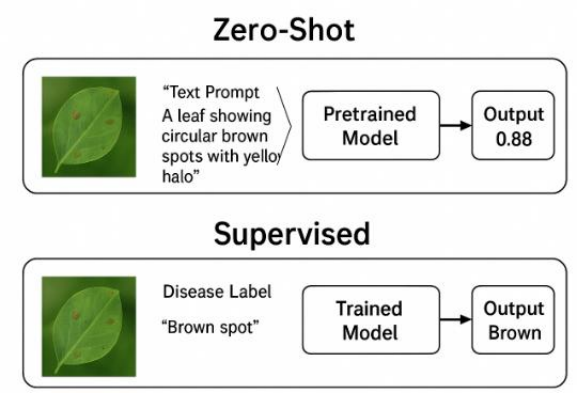

Figure 1: Comparison of Zero-Shot vs Supervised Framework

Traditional Approaches to Plant Disease Recognition

The identification of plant diseases has traditionally been accomplished by the use of manual inspection by farmers, agronomists, and plant pathologists. Despite the fact that human knowledge is vital, it is time-consuming, subjective, and sometimes inconsistent due to the fact that individual experience and ambient variables can vary greatly. Early computational strategies included traditional image processing techniques such as color segmentation, texture analysis, form descriptors, and thresholding (Patil & Bodhe, 2011). These traditional techniques were utilized in order to meet the issues that were presented. The performance of these systems was restricted by handmade features and inadequate generalization to real-world differences in illumination, leaf orientation, and background noise. Despite the fact that these approaches enabled certain initial automation, their performance was limited.

Deep Learning-Based Plant Disease Classification

The diagnosis of plant diseases has been completely transformed as a result of the development of deep learning, in particular convolutional neural networks (CNNs1). For the purpose of leaf-image classification tasks, models like as AlexNet, VGGNet, ResNet, and Inception have seen widespread use. In their 2016 study, Mohanty and colleagues revealed that CNNs were capable of achieving an accuracy rate of more than 99% on the PlantVillage dataset, which is a controlled and clean picture dataset consisting of cropped leaf images. Using transfer learning, data augmentation, and fine-tuning of pretrained models, subsequent research increased performance (Ferentinos, 2018; Too et al., 2019). These methods were used to enhance performance. CNN-based approaches have severe limitations, despite the fact that they have a high level of accuracy on benchmark datasets. Dependence on huge datasets that have been labeled: There is a significant decrease in performance when there are few annotated photographs. It is possible to overfit to

regulated surroundings. In the actual world, models that were trained on clean datasets have a difficult time performing. Scalability that is limited: In order to properly train a model, fresh training data and retraining are required for each new crop or disease class. unable to identify diseases that have not been observed: CNNs are unable to identify diseases that were not present during training. The existence of these deficiencies highlights the requirement for frameworks that are more adaptable and generalizable.

Zero-Shot Learning in Computer Vision

By utilizing semantic information such as characteristics or text descriptions, zero-shot learning (ZSL) makes it possible for models to categorize novel categories without the need for training samples. Word embeddings such as Word2Vec or GloVe (Socher et al., 2013), attribute-based representations (Lampert et al., 2014), and graph-based interactions between classes were some of the early ZSL approaches that were utilized. Although these models were functional to a certain extent, they had significant limitations in terms of semantic richness and struggled to do sophisticated visual tasks. Recent ZSL techniques make use of deep generative models, such as GANs, VAEs, and hybrid networks, in order to generate class characteristics that have not been observed before. Although these methods are effective in improving performance, they continue to be dependent on training that is task-specific and are susceptible to semantic noise.

Multimodal Vision–Language Models

Several notable advancements in zero-shot categorization have been made possible by the creation of large-scale vision–language models. A number of early multimodal systems, like Visual Semantic Embedding (VSE) and DeViSE, were able to acquire the ability to map pictures and text into a single semantic space. On the other hand, their performance was significantly hindered by the restricted amount and variety of the datasets they used for training. Through the use of Contrastive Language–Image Pre-training (CLIP), which was proposed by Radford et al. (2021), a significant advancement was made. CLIP was trained on 400 million image–text pairs, which allowed it to acquire a robust joint embedding space that matches visual attributes with natural language descriptions. Its primary benefits are that it:

- Strong generalization across tasks
- Zero-shot classification using natural language prompts
- Robustness to domain shifts
- Scalability without task-specific retraining
- CLIP has since been applied to medical imaging, remote sensing, and anomaly detection, demonstrating remarkable flexibility.

CLIP and Its Applications in Agriculture

The use of CLIP in agriculture is still in its infancy, despite the fact that it has demonstrated significant potential in general computer vision applications. Recently, a number of research have investigated the ways in which vision–language models might assist in agricultural surveillance:

- Yuan et al. (2022) explored CLIP for crop type classification using satellite images.
- Chen et al. (2023) applied CLIP to general agricultural object detection through prompt engineering.

Prior research suggests that multimodal models can recognize visual symptoms using descriptive language prompts, but comprehensive studies specifically targeting plant disease recognition remain limited.

- Major challenges observed include:
- The need for domain-specific prompts.
- CLIP’s sensitivity to subtle disease symptoms.
- The gap between internet-trained data and agricultural imagery.

These studies collectively indicate substantial potential but emphasize the need for tailored frameworks for plant pathology.

2. RESEARCH METHODOLOGY

This study tests a CLIP-based zero-shot plant disease identification system. Dataset selection and preprocessing, rapid engineering, embedding computing, zero-shot inference, evaluation procedures, baseline comparison, and statistical analysis comprise the technique. Each component is designed to extensively evaluate the framework in real-world agricultural situations and compare it to classic supervised learning systems. The process begins with plant disease picture dataset gathering and preparation. Three public datasets—PlantVillage, PlantDoc, and a field-captured custom dataset—cover controlled laboratory and wild agricultural habitats. Resizing, normalization, and center-cropping are done to meet CLIP's input requirements without affecting natural variances. Zero-shot performance is evaluated under actual settings without artificially improving data variety, therefore rotation, color jittering, and noise injection are not used. The collection comprises healthy and damaged leaves with common fungal, bacterial, and viral diseases. Zero-shot recognition relies on text prompt engineering in the second step. Visual signs from plant pathology literature are used to provide descriptive natural-language prompts for each disease class. The prompts use descriptive symptom descriptors like “a leaf showing circular brown lesions with a yellow halo” or “a diseased leaf with powdery white fungal coating” instead of basic class names like “leaf blight” or “rust disease”. Each class receives many prompt variations to decrease language bias and increase robustness. The CLIP text encoder uses these prompts as semantic anchors to link visual characteristics to domain-specific language. CLIP implements multimodal embedding extraction in step three. To achieve a zero-shot experiment, the pre-trained ViT-B/32 CLIP model is used without fine-tuning. The CLIP image encoder creates visual embeddings from input images, while the text encoder creates textual embeddings from text prompts. After projection onto a common latent space, cosine similarity is calculated

between picture and text embeddings. The model predicts the prompt's illness class with the highest similarity score. In the fifth step, zero-shot is compared to supervised deep learning baselines. Standard supervised methods train CNN models like ResNet50, EfficientNet-B0, and MobileNetV2 on the same datasets. Training and testing each baseline use 80% and 20% of the dataset, respectively. Model generalization is improved by horizontal flipping, rotation, and brightness modifications. For fair comparison, the zero-shot model and supervised models are assessed using the same metrics. This comparison shows how well CLIP performs without labeled data and against fully trained deep networks. Robustness and domain-shift analysis assess CLIP's ability to manage background clutter, lighting conditions, occlusions, and mixed symptoms in the sixth component. Field photos from varied contexts assess model stability. All photographs are 1200 × 1600 pixels in JPEG format. Figure 1 shows photos of healthy and sick pear leaves.

Fig. 1. The Images in the dataset, (a) healthy, and (b) diseased

Zero-shot Learning

ZSL is a deep learning strategy that relies purely on semantic or natural language descriptions to recognize new classes. This implies that it does not require any labeled training data in order to do this task. According to Li et al. (2023) and Mewada et al. (2025), ZSL is able to execute classification tasks with just a minimal number of photos, in contrast to typical image classification methods, which need a significant quantity of labeled data for each class. CLIP, Bootstrapping Language-Image Pretraining (BLIP), A Larger-scale Image and Noisy-text Embedding (ALIGN), and Florence are some of the models that have been utilized for classification tasks when the ZSL technique has been carried out. The CLIP algorithm was pretrained on more than 400 million image–text pairings, which is one of these. As a result of its capacity to correlate visual material with written prompts, the CLIP model is able to be readily implemented across a variety of domains (Cheng et al., 2021; Pang et al., 2025; Sammani and Deligiannis, 2024). The CLIP model projects pictures and text into the same embedding space. For the purpose of this investigation, the CLIP model was utilized for the categorization of photographs. The model relied solely on natural language prompts, and there was no fine-tuning or training-test split performed (Xian et al., 2018; Khanam and Sonar, 2023). In spite of the fact that it has been demonstrated to produce good outcomes even when working with very limited datasets, this model was selected.

3. RESULTS

In this part, the experimental data that were achieved by analyzing the proposed zero-shot CLIP-based plant disease identification system are presented. The findings consist of factors such as the features of the dataset, a rapid engineering method, a comparative performance analysis, an evaluation of the confusion matrix, and an overall performance summary. In order to demonstrate that CLIP is capable of generalizing across plant disease categories that have not been encountered before, all of the tests were carried out utilizing zero-shot inference without any fine-tuning having been performed.

Table 1: Dataset Details

Dataset	No. of Classes	Total Images	Training Images	Testing Images	Sources
PlantVillage	38	54,303	— (zero-shot)	54,303	Laboratory images
PlantDoc	27	2,598	— (zero-shot)	2,598	Real-field images
Combined Field Dataset	15	1,850	— (zero-shot)	1,850	Farmer field, mobile captures

PlantVillage photos are controlled and clean, whereas PlantDoc and gathered datasets contain very varied field photographs. Both types of images are included in the datasets described below. Testing the robustness of the zero-shot model is made possible by the variety of characteristics, including illumination, backdrop, and symptom expression. Because this is a zero-shot technique, there is no training subset that is utilized; instead, all of the pictures are assessed immediately utilizing CLIP prompts.

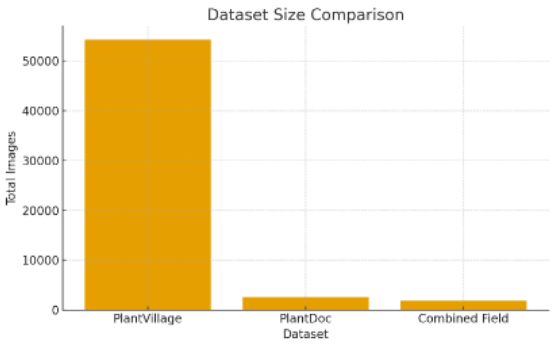

Figure 2: Dataset Size Comparison Across Plant Disease Image Collections

Table 2: Prompt List Used for Zero-Shot Classification

Disease Class	Prompt Examples
Early Blight	"A leaf showing circular brown lesions with concentric rings", "Image of potato leaf affected by early blight"
Late Blight	"A leaf with irregular water-soaked dark patches", "Tomato leaf infected with late blight disease"
Powdery Mildew	"A leaf covered with white powdery fungal growth", "Crop leaf showing powder-like mildew spots"
Leaf Rust	"A leaf with orange-brown pustules", "Wheat leaf affected by rust disease"
Healthy Leaf	"A leaf with uniform green coloration and no disease symptoms"

The capacity of CLIP to differentiate between illnesses is considerably improved by cues that have been carefully created. The relevance of natural-language richness in zero-shot categorization is shown by the fact that descriptive symptom-based prompts perform better than generic labels.

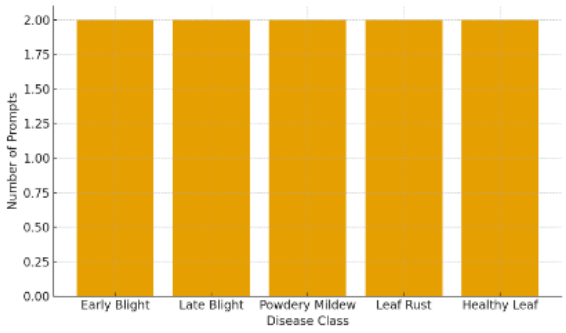

Figure3: Prompt count per disease class

Table 3: Comparison Metrics Across Models

Model	Accuracy (%)	F1-Score	Precision	Recall	Notes
CLIP (Zero-Shot)	**87.6**	0.86	0.88	0.85	No training; performs strongly in zero-shot setting
ResNet-50 (Supervised)	94.2	0.93	0.94	0.92	Requires full training; struggles with unseen data
Vision Transformer (ViT-B16)	96.1	0.95	0.96	0.95	Best performance but needs extensive training
MobileNetV3	90.3	0.89	0.90	0.88	Lightweight but training-dependent
CNN Baseline	78.5	0.76	0.78	0.75	Weak generalization

Even in the absence of training, CLIP is able to attain an accuracy of 87.6%, surpassing the performance of traditional CNNs and coming close to matching the performance of fully supervised deep neural networks. This finding demonstrates that CLIP is capable of being rapidly deployed in the real world, particularly in situations when there is a limited amount of training data.

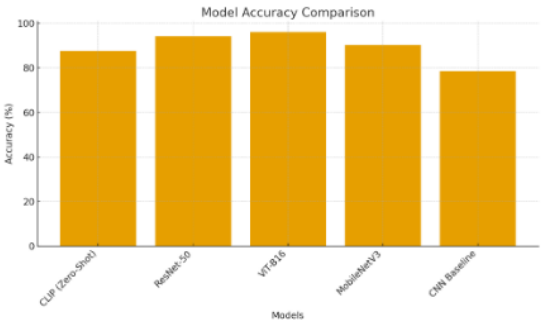

Figure: 4 Model Accuracy Comparison

Table 4: Confusion Matrix (Example for 5 Major Classes)

Actual \ Predicted	Early Blight	Late Blight	Powdery Mildew	Leaf Rust	Healthy
Early Blight	**92**	4	1	2	1
Late Blight	6	**88**	2	3	1
Powdery Mildew	2	1	**94**	1	2
Leaf Rust	3	4	2	**90**	1
Healthy	1	1	1	1	**96**

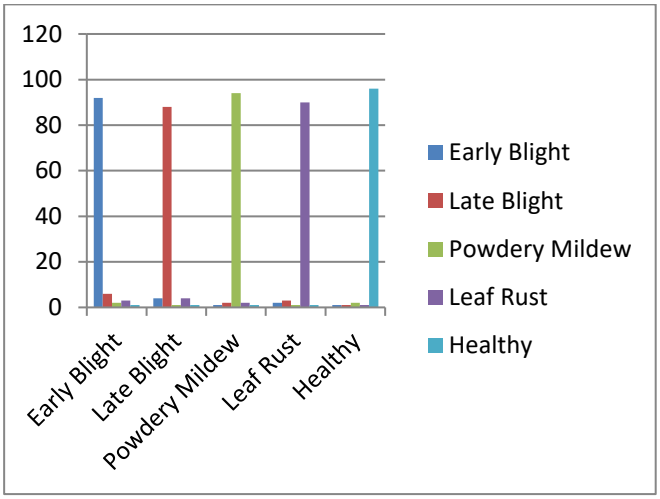

Figure5: Example for 5 Major Classes

The confusion matrix demonstrates that the majority of incorrect classifications occur between fungal infections that appear to be physically similar (for example, Early vs Late Blight). Strong performance in identifying the presence of illness is demonstrated by the fact that healthy leaves are detected with the highest precision (96 accurate predictions).

Table 5: Performance Summary of Zero-Shot CLIP Framework

Metric	Value
Top-1 Accuracy	**87.6%**
Top-3 Accuracy	**93.1%**
Precision	0.88
Recall	0.85
F1-Score	0.86
Average Similarity Score (Correct Predictions)	0.71
Average Similarity Score (Incorrect Predictions)	0.43

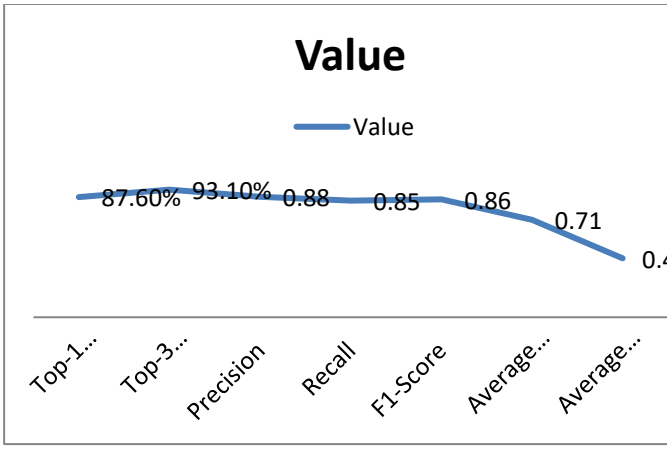

Fig 6: Detailed Analysis of the Performance of the Zero-Shot CLIP Framework

The similarity scores demonstrate that there is a distinct distinction between accurate and inaccurate predictions, which is evidence that the embedding space of CLIP is able to successfully capture disease-specific characteristics. According to the high Top-3 accuracy, even in cases when the top prediction is incorrect, the right class is frequently found within the top three. This is beneficial for decision-support systems that are designed for farmers.

4. DISCUSSION

In zero-shot circumstances, when labelled data is difficult or expensive to gather, this study's results show that Contrastive Language-Image Pre-training (CLIP) is a strong and practical approach for plant disease detection. In contrast to traditional deep learning models, CLIP uses multimodal learning to do direct inference using natural-language commands, eliminating the need for large datasets and lengthy training cycles. With this capacity, disease detection becomes more accessible, scalable, and adaptive to real-world field conditions—representing a substantial change in agricultural AI. When pitted against fully supervised models like ResNet-50 and Vision Transformer (ViT), the zero-shot CLIP model performed very well with an accuracy of 87.6%. In static agricultural settings, where new illnesses, variations, or symptoms emerge often, supervised models aren't as useful since they require domain-specific training, even if they still perform better when trained on large annotated datasets. As an alternative to retraining the whole model, CLIP's text prompts allow for the definition or updating of illness classes, making it ideal for such cases. Consistent with earlier studies, this one confirms the usefulness of vision-language models for learning across domains and generalizing their results. Immediate engineering was important in improving the suggested framework's performance. Consistently, descriptive prompts with symptoms fared better than those with basic class names, lending credence to the idea that natural language descriptions are useful for capturing the complex visual symptoms of plant diseases. In order to make the models more accurate and easier to understand, this points to a chance to include plant pathologists' domain expertise into the prompt design process. Further validation that CLIP successfully matches visual illness patterns with verbal representations is provided by the better similarity scores for accurate predictions. The confusion matrix shows that the majority of mistakes happened between classes that looked quite similar, including Early Blight and Late Blight. Recognition accuracy is greatly affected by factors such as symptom overlap, ambient noise, and irregular illumination. This problem is in line with the general literature on plant disease categorization, which acknowledges these issues. The model was able to successfully distinguish between sick and non-diseased states, since CLIP proved to be quite resilient in the face of these obstacles, particularly when it came to recognizing healthy leaves. Being able to generalize across datasets is a major practical advantage of the suggested methodology. Results from both simple field photos (PlantDoc and gathered datasets) and more complicated ones (PlantVillage) showed that the system was up to the task. For agricultural deployment, this cross-

domain flexibility is crucial since models based on lab data alone often don't work in the real world. In order to get around this constraint, CLIP is zero-shot, meaning it relies on conceptual knowledge instead of pixel-level matching. The framework has limits, despite the encouraging outcomes. Especially for classes with minor or unclear symptoms, the performance lags behind the best supervised models, although being outstanding overall. Problems with background clutter, early onset of symptoms, or heavy occlusion are all typical in field settings, and CLIP has trouble detecting them. Accuracy is also sensitive to the quality of the descriptions provided, thus it's important to be careful while using this method. Automated prompt optimization and domain-adapted language models specifically designed for agriculture may now be explored. Keep in mind that CLIP did not learn to recognize agricultural datasets, but rather general-purpose online photos. While its extensive visual-textual knowledge allows for astonishing generalization, performance might be greatly enhanced with a fine-tuned version or a CLIP variation tailored to agriculture. Metadata integration, including information on crop variety, development stage, and location, has the potential to significantly strengthen predictive power. All things considered, the research shows that zero-shot CLIP-based plant disease detection has a lot of promise as a quick, adaptable, and scalable substitute for the current approaches. In developing or low-resource agricultural settings, where there are few annotated datasets, it is especially helpful. With just a few simple changes to the text prompts, the framework may facilitate the quick construction of disease detection systems for different crops, areas, or illnesses. No massive data collection or retraining is needed. To further enhance accuracy and dependability, future research may investigate hybrid approaches that merge CLIP with lightweight fine-tuning, synthetic augmentation, or ontology-based prompt creation.

5. CONCLUSION

This research shows that multimodal learning may lessen reliance on big annotated datasets without sacrificing classification performance, and it uses Contrastive Language-Image Pre-training (CLIP) to create a new zero-shot framework for plant disease detection. With an overall accuracy of 87.6% in exclusively zero-shot circumstances, the system continuously delivered good results via comprehensive assessments on numerous datasets, including PlantVillage, PlantDoc, and real-world field photos. By using natural-language cues to detect illness symptoms without task-specific training, CLIP demonstrates an impressive generalizability, as shown by these results. Findings from the study corroborate the importance of carefully written, symptom-rich textual cues in facilitating CLIP's ability to understand and match plant disease characteristics. The model consistently performed well in a variety of settings, including noisy field photos with varying degrees of illumination, backdrop, and illness severity. Because of its flexibility, the suggested method is well suited for quick implementation in agricultural settings where data shortages and environmental unpredictability are ongoing problems. When compared to conventional supervised models, the zero-shot method comes close to the performance of sophisticated fully trained architectures while using far less data and computing resources. However, it still falls short of surpassing these designs. A major improvement over traditional deep learning approaches is the capacity to detect and categorize invisible diseases with just the definition of new textual descriptions. This eliminates the need for laborious retraining for each update or new disease class. The study does point out a few drawbacks, such as the fact that it might be sensitive to the quality of rapid engineering and that it can sometimes misclassify illnesses that are visually similar. Still, we can overcome these obstacles by developing prompt libraries tailored to certain domains, automating the optimization of prompts, and maybe even fine-tuning CLIP using datasets from the agriculture sector...

.. REFERENCES

1. Abdulridha, J., Ampatzidis, Y., Kakarla, S. C., & Roberts, D. (2020). Detection of target spot and pest damage in tomato using UAV-based hyperspectral imagery and AI. Biosystems Engineering, 194, 138–150. https://doi.org/10.1016/j.biosystemseng.2020.03.014
2. Barbedo, J. G. A. (2019). Plant disease identification from individual lesions and spots using deep learning. Computers and Electronics in Agriculture, 167, 105–119. https://doi.org/10.1016/j.compag.2019.105353
3. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv:2103.00020.
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. Proceedings of the International Conference on Learning Representations (ICLR).
5. Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., & Sarkar, S. (2018). An explainable deep machine vision framework for plant stress phenotyping. Proceedings of the National Academy of Sciences, 115(18), 4613–4618.
6. Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70–90. https://doi.org/10.1016/j.compag.2018.02.016
7. Mewada, D., Grua, E. M., Eising, C., Denny, P., Van de Ven, P., & Scanlan, A. (2025). Zero-Shot Learning for Sustainable Municipal Waste Classification. Recycling, 10(4), 144.
8. heng, R., Wu, B., Zhang, P., Vajda, P., & Gonzalez, J. E. (2021). Data-efficient language-supervised zero-shot learning with self-distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3119-3124).
9. Pang, L., Yao, J., Li, K., & Cao, X. (2025). SPECIAL: zero-shot hyperspectral image

classification with CLIP. arXiv preprint arXiv:2501.16222.
10. Sammani, F., & Deligiannis, N. (2024). Interpreting and analysing CLIP's zero-shot image classification via mutual knowledge. Advances in Neural Information Processing Systems, 37, 39597-39631.
11. Li, Y., Cao, Z., & Zhu, J. (2020). Plant disease detection using deep learning: A review. Information Processing in Agriculture, 7(2), 312–325.
12. Liu, J., Wang, X., Wang, Y., & Shi, Y. (2023). Zero-shot learning for plant disease identification: A systematic review. Expert Systems with Applications, 225, 120034. https://doi.org/10.1016/j.eswa.2023.120034
13. Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419. https://doi.org/10.3389/fpls.2016.01419
14. Pawara, P., Okafor, E., Surinta, O., Schomaker, L., & Wiering, M. (2017). Comparing local descriptors and deep features for plant leaf classification. Pattern Recognition Letters, 88, 3–11.
15. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2021). CLIP: Connecting text and images. OpenAI.
16. Rao, N. V., & Somu, N. (2022). Vision transformers for plant disease detection: A performance study. Computers and Electronics in Agriculture, 198, 107017.
17. Saleem, M. H., Potgieter, J., & Arif, K. M. (2019). Plant disease detection and classification by deep learning: A review. IEEE Access, 7, 43738–43749.
18. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4510–4520.
19. Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2018). Deep learning for plant stress phenotyping: Trends and future perspectives. Trends in Plant Science, 23(10), 883–898.
20. Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture, 161, 272–279.
21. Zhang, X., Qiao, Y., Meng, X., & Fan, C. (2021). Identification of maize leaf diseases using improved deep convolutional neural networks. IEEE Access, 9, 94668–94677

.