Journal of Soft Computing Exploration

Homepage: https://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

A comparative benchmark of vision transformer architectures for chili leaf disease classification

Acihmah Sidauruk[1], Danang Wijayanto[2], I Made Artha Agastya[3], Jumanto Unjung[4], Mulia Sulistiono[5]
[1,2,3,5]Department of Computer Science, Universitas AMIKOM Yogyakarta, Indonesia
[4]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received April 22, 2026
Revised May 31, 2026
Accepted July 4, 2026

Keywords:

Image classification
Chili plant disease
Self-supervised learning
DINOv2
Vision transformer

ABSTRACT

Chili plant disease detection represents a critical component for enhancing agricultural productivity. Although Convolutional Neural Networks (CNN) have demonstrated promising results, they encounter limitations in capturing global contextual relationships within images. However, existing Vision Transformer studies on plant disease commonly assess only a single architecture, leaving the relative performance of different Vision Transformer families on chili disease data largely unexamined. This research aims to conduct a comparative benchmark analysis of five Vision Transformer-based architectures ViT, Swin Transformer, MaxViT, DINOv2, and EVA-02 to identify the most optimal model for chili plant disease classification. The methodology begins with data preprocessing and augmentation on a chili leaf dataset comprising five classes: healthy, leaf curl, leaf spot, whitefly, and yellowish. Each model is then fine-tuned under consistent training configurations with early stopping to prevent overfitting, and evaluated using accuracy, precision, recall, F1-score, and AUC. The results indicate that DINOv2 achieves superior performance with 96% accuracy, 96% precision, 96% recall, 96% F1-score, and 99% AUC, along with the highest training efficiency through convergence at epoch 12, outperforming ViT (92%), Swin (88%), MaxViT (88%), EVA-02 (86%), and previous CNN-based approaches. These findings confirm the superior potential of Vision Transformers, particularly self-supervised models, as a promising alternative for agricultural disease detection applications. The main contribution of this study is the first unified, head-to-head benchmark of five distinct Vision Transformer families for chili leaf disease classification, providing practical guidance on model selection in terms of both accuracy and training efficiency.

Corresponding Author:

Danang Wijayanto,
Department of Computer Science,
Universitas AMIKOM Yogyakarta,
Ring Road Utara Street, Ngringin, Condongcatur, Depok District, Sleman Regency, Special Region of Yogyakarta 55281, Indonesia
Email: danangwijayanto@students.amikom.ac.id
https://doi.org/10.52465/joscex.v7i3.62

1. INTRODUCTION

Chili (Capsicum annuum L.) represents one of the most economically significant and highly valuable potential commodities [1]–[4]. Global chili pepper production reaches 4.1 million tons distributed across various countries, with the majority concentrated in Asian regions [5]. Nevertheless, optimal productivity levels have not yet been achieved by farmers, who face substantial challenges [6]. These challenges are primarily attributed to pest and disease disturbances that receive inadequate treatment, resulting in significant crop damage and ultimately leading to harvest failure [7]. Traditional methods in plant disease diagnosis encounter various significant limitations [8], [9]. Additional challenges arise because identical symptoms can be caused by diverse factors (fungi, bacteria, or non-infectious diseases), which complicates accurate diagnosis and pathogen identification. The need to address existing problems in traditional diagnosis has driven the development of novel methods for detecting and identifying pathogen presence in plant pathology [10].

The advancement of deep learning technology offers promising solutions to address problems in plant disease diagnosis by understanding patterns [11]. Research [12], [13] explains the importance of growing demands for Precision Agriculture (PA), which has driven researchers to evaluate machine learning utilization in predicting harvest yields and determining variables that influence them. Similar approaches can be applied to plant disease detection using deep learning methods such as CNN. CNN represents a deep learning method commonly utilized for solving complex problems regarding image recognition and has been recognized since 2012 [14].

Although CNN has demonstrated good performance [15]–[18], Vision Transformer (ViT) emerges as a promising new approach in image classification. Various studies have shown favorable results from ViT utilization developed by Google Research, with diverse applications in image recognition tasks [19]. Transformer architectures (ViT) are now increasingly recognized as one of the prominent neural network types, with the capability to process outputs using highly expressive and effective attention mechanisms [20]. ViT divides each image into a sequence of fixed-length tokens, then applies multiple Transformer layers to model global relationships among tokens for classification [21]. CNN is widely used in machine learning fields and excels in feature extraction, yet it has limitations in capturing global contextual relationships among image features. Conversely, ViT applies attention mechanisms to understand global relationships among features [22].

Previous research has extensively conducted experiments using vision transformer models in chili disease classification, yet many studies are still limited to a single Vision Transformer architecture [23]–[25]. In particular, Kursun and Koklu [23] combined AlexNet with a Vision Transformer, improving accuracy from 86.43% with AlexNet alone to 94.80% with the hybrid model on an Indian chili-leaf dataset, but evaluated only this single hybrid configuration. Cheng et al. [24] benchmarked three ViT variants (ViT-B/16, ViT-B/32, ViT-L/16) against three YOLOv8 variants (YOLOv8n-cls, YOLOv8s-cls, YOLOv8m-cls) on 70 general plant-leaf health classes, reporting a best test accuracy of 97.7%, yet their focus was the ViT-versus-YOLO trade-off rather than a comparison among diverse ViT families. Chen et al. [25] applied a Swin Transformer enhanced with an image-enhancement pipeline to a 31-class general plant dataset in which chili is only one class, raising accuracy from 88.32% to 89.03% while reporting accuracy as the sole metric. However, to the best of our knowledge, no study has benchmarked multiple architecturally diverse ViT models specifically for chili disease detection under a unified training and evaluation protocol, which is the gap this study addresses. To obtain the most optimal model, experiments involving several ViT-based models are required to evaluate the most superior model for use in chili disease classification. This study conducts experiments to implement several vision transformer-based models such as ViT, Swin, MaxVit, DINOv2, and EVA-02 in chili disease classification. Each model to be tested possesses different characteristics. ViT serves as a pure transformer model for vision [26]. Swin Transformer employs hierarchical approaches and shifted windows [27]. MaxViT combines the strengths of CNN and transformer [28]. DINOv2 utilizes self-supervised learning approaches [29]. EVA-02 employs masked image modeling [30].

This study aims to conduct comparative analysis of each model's performance and identify the most ideal model for use in chili disease detection application domains. Performance evaluation is conducted by analyzing accuracy, precision, recall, F1-score, and AUC for each model. The training process is monitored through training loss, validation loss, and early stopping implementation to prevent overfitting. The main contributions of this study are summarized as follows: (1) we present, to our knowledge, the first unified head-to-head benchmark of five architecturally diverse Vision Transformer families, namely ViT, Swin Transformer, MaxViT, DINOv2, and EVA-02, for chili leaf disease classification under identical training and evaluation protocols; (2) we provide a comprehensive evaluation using accuracy, precision, recall, F1-score, and AUC, together with training-efficiency analysis based on convergence epoch, rather than accuracy alone; we show that the self-supervised DINOv2 model achieves the best accuracy and the fastest convergence, and we derive practical guidance for selecting Vision Transformer models in agricultural disease-detection applications.

2. METHOD

This chili disease classification research conducts systematic experiments consisting of several stages including dataset collection, preprocessing, transfer learning for each model, and evaluation. The experimental workflow can be observed in Figure 1.

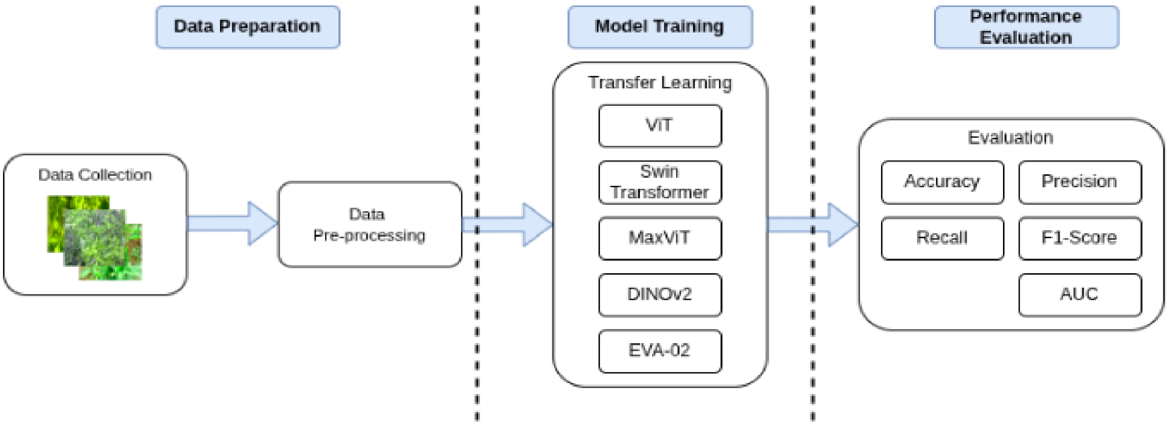

Figure 1. Methodology workflow

Dataset

The dataset is a publicly available chili leaf collection from Kaggle[[31] with a total of 500 images distributed as 80% training, 10% validation, and 10% testing. This dataset consists of several classes including healthy, yellowish, leaf spot, leaf curl, and whitefly. Visual examples of each class can be observed in Figure 2. The source does not document the image acquisition details or any prior use, which is noted as a limitation in the Conclusion.

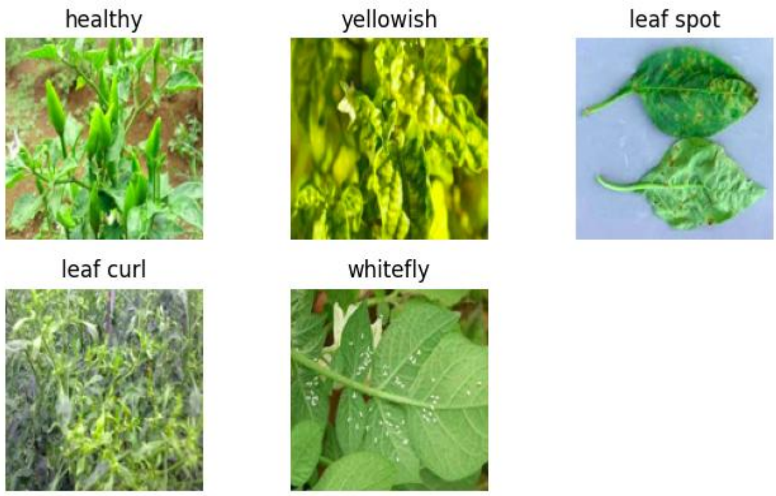

Figure 2. Sample images for each class

Yellowish is characterized by leaves that begin to wilt starting from the lower leaves turning yellow and progressively spreading [18]. Leaf curl presents symptoms of leaf rolling or curling, leaf vein thickening, and leaf edges that curl upward, causing new leaves to become small-sized, flower drop, and plants fail to produce fruit [32]. Leaf spot is identified by the appearance of round brown spots on leaves caused by the fungus Cercospora capsici, which can spread through wind, rainwater, vector pests, and agricultural tools [18]. Whitefly, particularly the species Bemisia tabaci Gennadius, serves as a vector in plant virus disease transmission especially yellow leaf curl disease in chili plants, and can cause damage to various types of agricultural crops [33].

Data Preprocessing

In the preprocessing stage, several augmentation and normalization techniques are applied to process the dataset before entering the transfer learning stage. Several preprocessing techniques follow the methodology from [34] to optimize data representation before transfer learning. This study utilizes various input sizes such as 224 × 224 for ViT, Swin-transformer, and MaxViT models, 336 × 336 for EVA-02, and 518 × 518 for DINOv2. This input size variation is attributed to different model selection strategies: DINOv2 employs 518 × 518 resolution as its default architecture, while EVA-02 is configured at 336 × 336 to adjust total params with other models within the (22M - 30M) params range. This study focuses on efficient and equivalent parameter counts despite input size differences that may affect computational requirements.

Data augmentation is performed on training data, including random resized crop, random horizontal flip, RandAugment, and ColorJitter before being converted to tensors and normalized using ImageNet parameters. Conversely, validation and test data undergo only resize and normalization stages without additional augmentation to ensure consistent evaluation. Detailed configurations and parameters for each preprocessing technique can be observed in Table 1.

Table 1. Data preprocessing pipeline

Name	Value
Resize	(256, 368, 550)
RandomResizedCrop	size = (224, 336, 518), scale = (0.75, 1.0), ratio =(0.9, 1.1)
RandomHorizontalFlip	P = 0.5
RandAugment	num_ops=2,magnitude=12
ColorJitter	brightness=0.1,contrast=0.1,saturation=0.1
Normalization	mean=[0.485, 0.456, 0.406], std=[0.229,0.224, 0.225]

Transfer Learning

Transfer learning is applied using five Vision Transformer-based pre-trained models with distinct architectures, each selected in its smaller version to balance performance and parameter efficiency. ViT [26] serves as the foundational architecture, processing images by dividing them into patches projected into lower-dimensional space, with self-attention enabling global information integration across the entire image from the earliest layers. Swin Transformer [27] extends this with a hierarchical architecture and shifted windowing scheme that limits self-attention to local windows while maintaining cross-window connectivity, achieving linear complexity relative to image size. MaxViT [28] addresses scalability limitations through multi-axis attention combining blocked local and dilated global attention, enabling global-local spatial interactions at arbitrary resolutions. DINOv2 [29] is a foundation model trained via self-supervised learning on diverse curated data, producing general-purpose visual features transferable across various image distributions. EVA-02 [30] further advances plain ViT through architectural improvements and optimized Masked Image Modeling (MIM) pre-training, achieving strong visual representations with moderate model size. All models are fine-tuned on the preprocessed dataset and evaluated to measure classification performance.

Each model in the transfer learning process will utilize a batch size of 16, learning rate of 0.0001, Adam Optimizer, ReduceLROnPlateau scheduler, and unfreezing of the final block in the model architecture. Additionally, early stopping techniques will be implemented following [35]. The patience value employed is 10 to prevent the model from experiencing overfitting, with a maximum of 100 epochs.

Evaluation

To evaluate the performance of the fine-tuned models, this study employs metrics, namely accuracy, precision, recall, and F1-score, all of which are derived from the confusion matrix. The confusion matrix is a table that compares the labels predicted by the model against the true labels, from which four basic counts are obtained for each class: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [36], [37]. Based on these counts, accuracy denotes the overall proportion of correctly classified samples, computed as the number of correct predictions divided by the total number of predictions [38]. Precision measures, among all samples the model assigns to a class, the fraction that genuinely belongs to that class, and therefore reflects the model's ability to avoid false positives [38]. Recall, also called sensitivity, measures the fraction of samples that truly belong to a class and are correctly identified, reflecting the model's ability to avoid false negatives [38]. The F1-score combines precision and recall into a single value through their harmonic mean, providing a balanced measure that is particularly useful when the number of samples per class is uneven [38]. Additionally, AUC-ROC metric is also employed as a popular indicator to analyze model performance more comprehensively [39]. The formal definitions of these metrics are given in Equations (1) to (5).

.

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

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

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

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

$$AUC = \int_0^1 ROC\ (t)dt \quad (5)$$

3. RESULTS AND DISCUSSIONS

The results of experiments on each model demonstrate diverse outcomes in early stopping performance. Each model reaches early stopping at different epochs, which can be observed in Table 2.

Table 2. Early-stopping (convergence) epoch of each model

Model	Stopping Epoch
ViT	30
Swin-Transformer	15
MaxViT	23
DINOv2	12
EVA-02	15

Experimental results demonstrate diverse convergence variations across Vision Transformer-based models. DINOv2 exhibits the fastest convergence with stopping epoch at epoch 12, followed by Swin and EVA-02, both of which terminate at epoch 15. MaxViT achieves convergence at epoch 23, while ViT requires the longest training duration with stopping epoch at epoch 30. These differences in stopping epochs indicate varying training efficiency, where models with lower stopping epochs demonstrate superior adaptation capabilities to the chili plant disease dataset.

Training and Validation Results

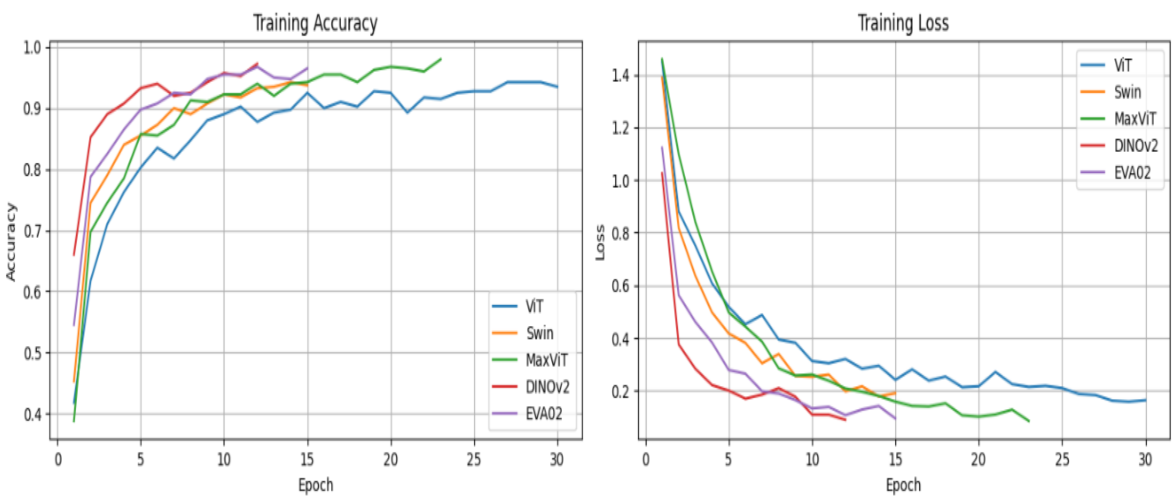

Figure 3. Training accuracy and loss of each model

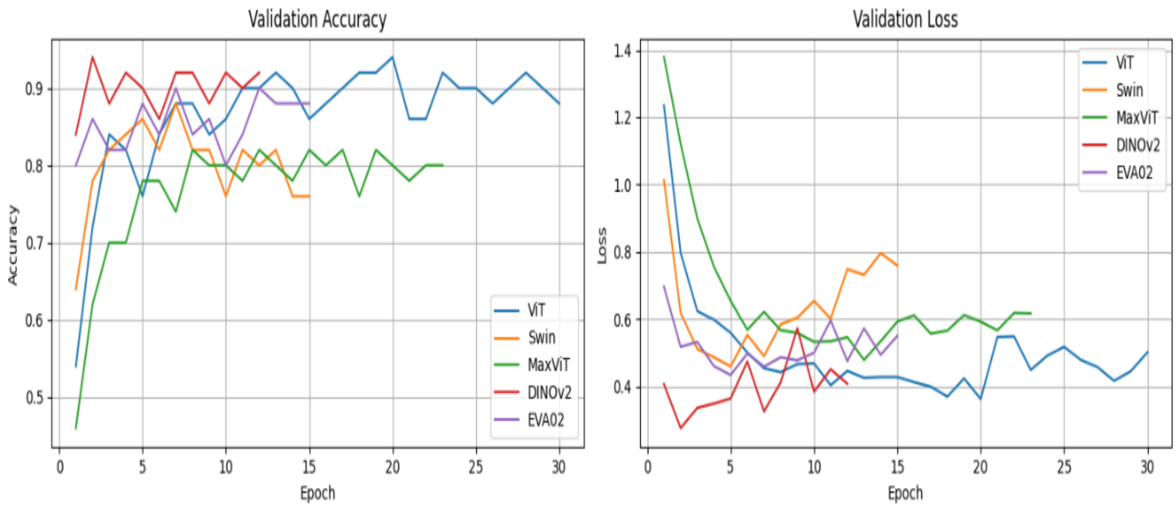

Figure 4. Validation accuracy and loss of each model

Based on Figure 3, the training accuracy results are presented. Each model experiences accuracy improvement and loss reduction as epochs progress. DINOv2 and MaxViT demonstrate the highest accuracy performance compared to other models. Additionally, all models exhibit consistent loss reduction throughout increasing epochs, indicating that the initial model learning process proceeds effectively in adjusting their respective weights.

Figure 4 presents the validation accuracy and loss results for each model. DINOv2 demonstrates the most stable and highest validation accuracy from the initial stages. This indicates that DINOv2 possesses exceptional generalization capabilities toward previously untrained data. EVA-02 also exhibits satisfactory validation performance with reasonably stable accuracy, albeit slightly below DINOv2 levels. Conversely, ViT demonstrates gradual and stable accuracy improvement throughout the training duration, successfully achieving accuracy levels approaching DINOv2 after epoch 20. In contrast, Swin and MaxViT models display fluctuating accuracy trends that remain lower than other models, which may indicate an imbalance between learning capabilities during training and generalization to validation data. In the validation loss graph, DINOv2 again demonstrates the most optimal results, achieving the lowest loss compared to other models. EVA-02 and ViT models also exhibit satisfactory loss reduction, although ViT requires additional epochs to reach its stability point. Meanwhile, Swin and MaxViT models show loss values that tend to fluctuate and fail to achieve significant reduction after several initial epochs. This condition indicates that both models tend to be less stable in the generalization process and potentially experience overfitting or encounter difficulties in maintaining performance on validation data.

Overall, the training and validation loss results demonstrate that the DINOv2 model with the fastest early stopping not only achieves the most rapid convergence but also exhibits superior generalization capabilities with the most stable and lowest validation loss. This indicates that models capable of achieving optimal convergence within shorter timeframes tend to possess superior generalization abilities compared to other models that require more extensive training epochs.

Performance Results

Performance evaluation of the transformer models reveals diverse classification capabilities across the five disease classes. While certain models demonstrate exceptional accuracy for specific disease types, others exhibit challenges in distinguishing between visually similar symptoms. The detailed confusion matrix analysis provides comprehensive insights into each model's error patterns. The variation in classification performance across models can be attributed to their distinct architectural designs and pre-training strategies. Models with self-supervised pre-training such as DINOv2 tend to develop more robust feature representations, whereas models relying solely on supervised pre-training may struggle with subtle inter-class visual similarities. In each confusion matrix shown in Figures 5 to 9, the rows correspond to the true (actual) disease class and the columns to the class predicted by the model. A diagonal cell therefore reports the number of test images of a given class that were classified correctly, whereas an off-diagonal cell reports the images of the row class that were

.

misassigned to the column class. Each value is a count of test images rather than a percentage, and the color shading serves only as a visual cue, with darker cells corresponding to higher counts and lighter cells to lower ones. Since correct predictions accumulate along the diagonal, a pronounced dark diagonal accompanied by near-empty off-diagonal cells reflects strong classification performance. These per-cell counts constitute the true positives and false negatives from which the accuracy, precision, recall, and F1-score reported in this section are derived; consequently, the matrices and the corresponding metric tables remain mutually consistent.

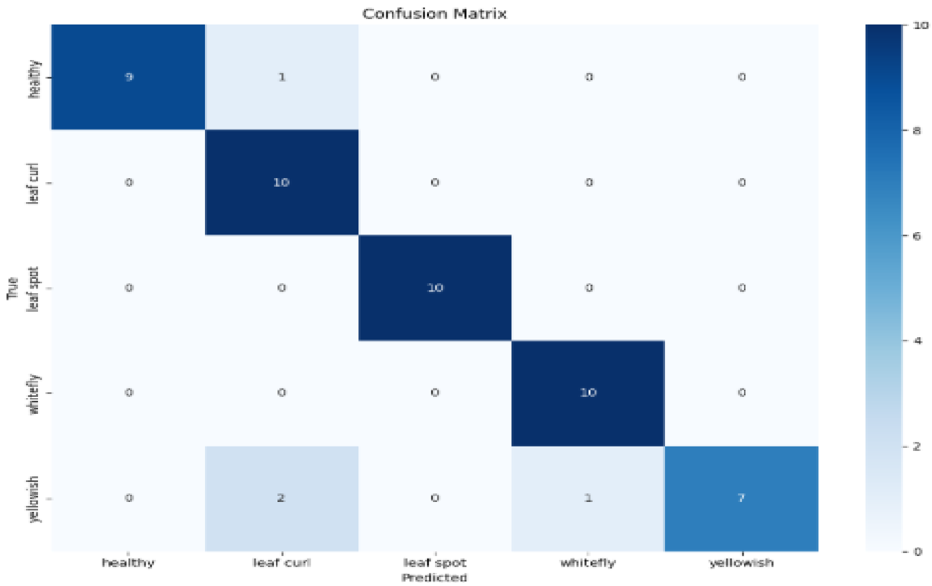

Figure 5. Confusion matrix of the ViT model (rows = true class, columns = predicted class)

Figure 5 presents the confusion matrix of the ViT model, which achieves perfect classification for leaf curl, leaf spot, and whitefly classes without prediction errors. The healthy class experiences 1 False Negative (misclassified as leaf curl). The yellowish class demonstrates 3 False Negatives (2 predicted as leaf curl, 1 as whitefly).

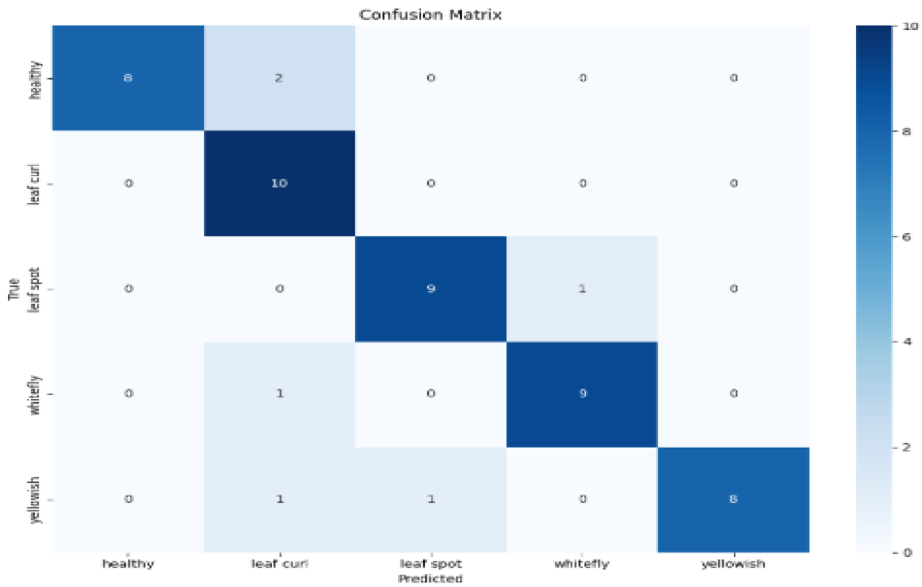

Figure 6. Confusion matrix of the Swin-Transformer model (rows = true class, columns = predicted class)

Figure 6 presents the confusion matrix of the Swin Transformer model, which achieves perfect classification only for the leaf curl class. The healthy class experiences 2 False Negatives (misclassified as leaf curl). The leaf spot class experiences 1 False Negative (misclassified as whitefly). The whitefly class experiences 1 False Negative (misclassified as leaf curl). The yellowish class experiences 2 False Negatives (1 predicted as leaf curl, 1 as leaf spot).

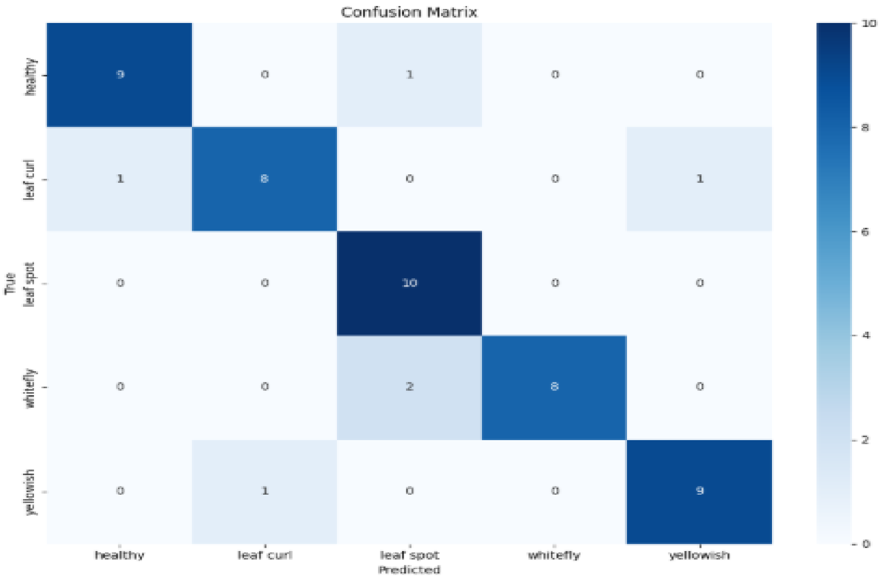

Figure 7. Confusion matrix of the MaxViT model (rows = true class, columns = predicted class)

Figure 7 presents the confusion matrix of the MaxViT model, which achieves perfect classification only for the leaf spot class. The healthy class experiences 1 False Negative (misclassified as leaf spot). The leaf curl class experiences 2 False Negatives (1 predicted as healthy, 1 as yellowish). The whitefly class experiences 2 False Negatives (misclassified as leaf spot). The yellowish class experiences 1 False Negative (misclassified as leaf curl).

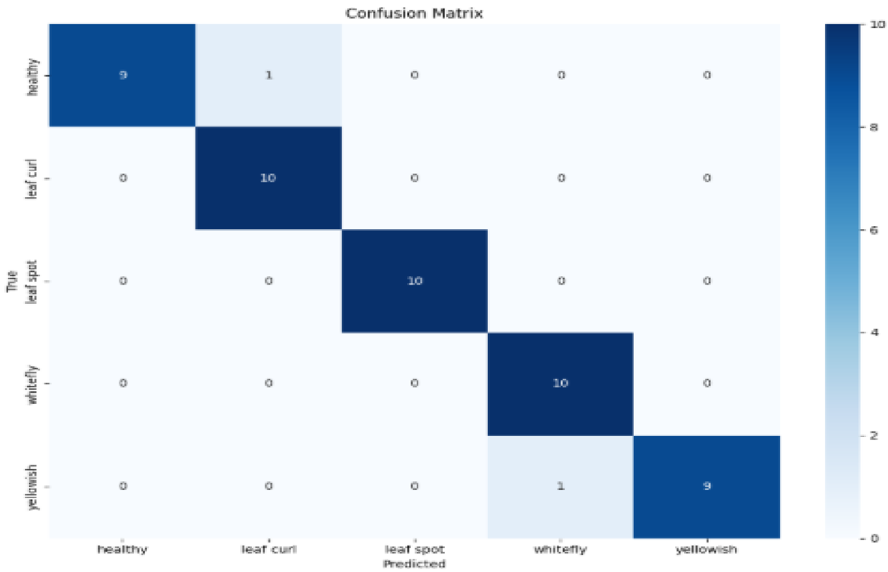

Figure 8. Confusion matrix of the DINOv2 model (rows = true class, columns = predicted class)

Figure 8 presents the confusion matrix of the DINOv2 model, which achieves perfect classification for leaf curl, leaf spot, and whitefly classes without prediction errors. The healthy class experiences 1 False Negative (misclassified as leaf curl). The yellowish class experiences 1 False Negative (misclassified as whitefly).

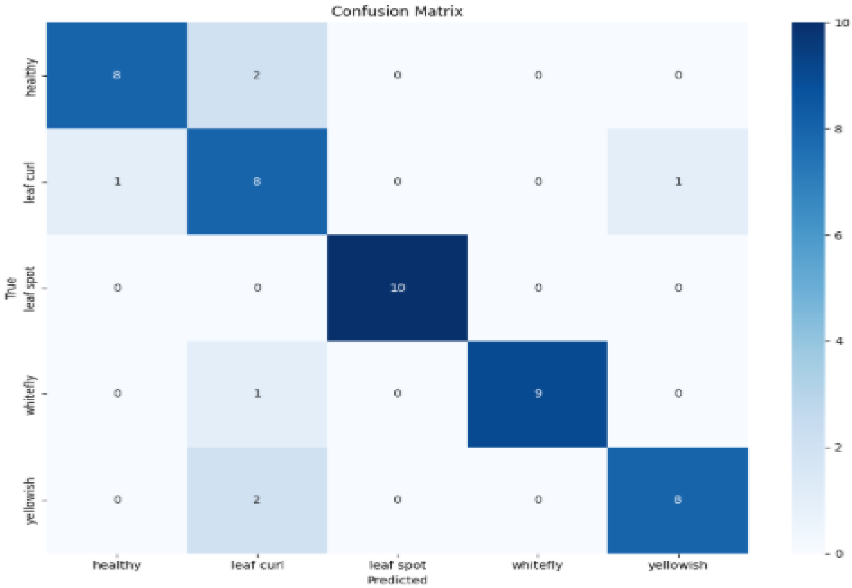

Figure 9. Confusion matrix of the EVA-02 model (rows = true class, columns = predicted class)

Figure 9 presents the confusion matrix of the EVA-02 model, which achieves perfect classification only for the leaf spot class. The healthy class experiences 2 False Negatives (misclassified as leaf curl). The leaf curl class experiences 2 False Negatives (1 predicted as healthy, 1 as yellowish). The whitefly class experiences 1 False Negative (misclassified as leaf curl). The yellowish class experiences 2 False Negatives (misclassified as leaf curl).

Table 3. Performance results of each model on the chili leaf test set

Model	Accuracy	Precision	Recall	F1-score	AUC
ViT	92%	94%	92%	92%	0.98
Swin	88%	90%	88%	88%	0.99
MaxViT	88%	89%	88%	88%	0.98
DINOv2	96%	96%	96%	96%	0.99
EVA-02	86%	88%	86%	87%	0.98

Table 3 presents the performance metrics of each model in classification tasks. DINOv2 achieves the highest performance with 96% accuracy and demonstrates consistently high values across other metrics. ViT, which requires the longest training duration, exhibits the second-best performance with 92% accuracy, indicating that vanilla Transformer architectures necessitate more iterations to achieve optimal representation but produce superior prediction quality EVA-02, despite achieving rapid convergence 15 epochs, demonstrates the lowest performance with 86% accuracy, indicating potential underfitting or limitations in extracting complex features from the chili disease dataset. Conversely, models with hybrid architectures such as Swin and MaxViT exhibit moderate performance at 88%. All models demonstrate high AUC values 0.98-0.99, indicating excellent discriminative capabilities in distinguishing disease classes. However, significant variations in accuracy reveal differences in model capabilities for handling class imbalance and classification threshold optimization on the chili plant disease dataset. A plausible explanation for these differences lies in the pre-training strategy adopted by each architecture. DINOv2 is pre-trained through self-supervised learning on large and diverse image collections, producing general-purpose visual features that transfer effectively to the small chili-leaf dataset. These features also remain robust to the subtle and visually similar symptoms shared by classes such as yellowish, leaf curl, and whitefly, as reflected by the minimal off-diagonal errors in Figure 8. In contrast, the supervised pre-training of EVA-02, together with the window-based hybrid designs of Swin and MaxViT, appears less capable of adapting its features under the limited data available. This limitation gives rise to the higher number of false negatives observed in their confusion matrices and, consequently, to their lower accuracy.

Table 4. Comparison with previous research

Scheme	Model	Accuracy	Precision	Recall	F1-Score	AUC
[16]	CNN	94%	95%	94%	94%	-
[18]	VGG16	94%	82%	82%	82%	-
[40]	Xception	91%	91%	91%	91%	-
[41]	EfficientLeafNet-B4	82%	78%	77%	77%	0.91
[42]	MobileNetV2	90%	92%	92%	92%	-
Our Experiment	ViT	92%	94%	92%	92%	0.98
	Swin-Transformer	88%	90%	88%	88%	0.99
	MaxViT	88%	89%	88%	88%	0.98
	DINOv2	**96%**	**96%**	**96%**	**96%**	**0.99**
	EVA-02	86%	88%	86%	87%	0.98

Table 4 presents a performance comparison between the proposed Vision Transformer models and previous CNN-based research [16],[18],[38]–[40]. DINOv2 achieves the highest accuracy 96%, surpassing all other models including the best-performing CNN 94%, although results demonstrate diverse performance variations for several other ViT models such as ViT 92%, Swin 88%, MaxViT 88%, and EVA-02 86%.

Experimental results demonstrate that DINOv2 achieves superior performance among the five evaluated Vision Transformer models. However, the performance variations exhibited by other models indicate substantial untapped potential within Vision Transformer architectures for chili plant disease classification. Models such as ViT, Swin Transformer, MaxViT, and EVA-02 retain considerable improvement opportunities through the implementation of advanced data augmentation techniques, optimal hyperparameter fine-tuning, and training strategies tailored to the specific characteristics of each architecture. The limitations of static parameter configurations in this study suggest that with appropriate methodological optimization, Vision Transformer-based models potentially achieve enhanced performance in future applications.

4. CONCLUSION

This study successfully evaluates the performance of five Vision Transformer-based models (ViT, Swin Transformer, MaxViT, DINOv2, and EVA-02) in chili plant disease detection with significant results. DINOv2 demonstrates superior performance with 96% accuracy, 96% precision, 96% recall, 96% F1-score, and 99% AUC, while simultaneously achieving the highest training efficiency through convergence at epoch 12. These results confirm that DINOv2 with its self-supervised learning approach exhibits effectiveness in the plant disease detection domain. More broadly, the ability of Vision Transformers to capture long-range dependencies and global contextual relationships through self-attention appears well suited to identifying the complex and spatially distributed symptoms present on chili leaves. Although performance varies across the models, the strong results obtained by DINOv2 highlight the promise of Vision Transformer architectures for plant disease detection, with further gains likely achievable through hyperparameter optimisation and more adaptive training strategies.

Beyond identifying the best-performing model, these findings carry practical implications. For chili disease detection under limited labelled data, a self-supervised foundation model such as DINOv2 offers an attractive balance between high accuracy and low training cost, and it can serve as a strong default backbone for agricultural screening tools. Nevertheless, the present study has several limitations. First, the evaluation relies on a single, relatively small public dataset of only 500 images, which constrains the generalisability of the results. Second, each model was fine-tuned using a single fixed set of hyperparameters and input resolutions rather than being optimised individually; the reported gaps between models may therefore narrow under architecture-specific tuning.

These limitations point to several directions for future research. First, evaluating the models on larger and more diverse chili-disease datasets would provide a more reliable assessment of their generalisability. Second, architecture-specific hyperparameter optimisation, combined with advanced augmentation tailored to each model, could reveal the full potential of the architectures that underperformed in this study. Finally, quantifying the trade-off between accuracy and computational cost, and deploying the best model on mobile or edge devices, would move the approach closer to practical, on-site disease detection for farmers.

REFERENCES

[1] F. Ahmad, K. Kusumiyati, M. A. Soleh, M. R. Khan, dan R. S. Sundari, “Chili crop innovation : Exploring enclosed growing designs for varied varieties — A review,” *Agrosyst Geosci Env.*, no. March, hal. 1–11, 2024, doi: 10.1002/agg2.20491.

[2] M. Hossain *et al.*, “Enhancing growth , yield , and nutritional value of Capsicum Annuum : Evaluating micronutrient efficiency and varietal performance,” *J. Agric. Food Res.*, vol. 21, no. April, hal. 101945, 2025, doi: 10.1016/j.jafr.2025.101945.

[3] M. M. Ali, N. I. Khalid, M. H. Wondi, N. I. N. Haris, dan P. N. M. A. Azman, “Exploring the nutritional values, volatile compounds, health benefits, and potential food products of chilli (Capsicum annuum): A comprehensive review,” *Food Chem.*, vol. 490, 2025, doi: 10.1016/j.foodchem.2025.145091.

[4] A. F. Faisal dan Y. F. Mustafa, “Chili pepper : A delve into its nutritional values and roles in food-based therapy,” *Food Chem. Adv.*, vol. 6, no. January, hal. 100928, 2025, doi: 10.1016/j.focha.2025.100928.

[5] M. T. Sundari, Darsono, J. Sutisno, dan E. Antriyandarti, “Analysis of chili supply in Indonesia,” *Int. Conf. Food Sci. Eng. 2022*, 2023, doi: 10.1088/1755-1315/1200/1/012036.

[6] N. Islam, M. Rashid, S. Wibowo, C. Xu, dan A. Morshed, “Early Weed Detection Using Image Processing and Machine Learning Techniques in an Australian Chilli Farm,” *Agric. 2021,* vol. 11, no. 387, 2021, doi: https://doi.org/10.3390/ agriculture11050387.

[7] K. Subagyono *et al.*, “Budidaya dan Pascapanen Cabai Merah (Capsicum annuum L.),” *BPTP Jateng / KAN*, 2010. https://repository.pertanian.go.id/handle/123456789/9109.

[8] A. Jafar, N. Bibi, dan R. A. Naqvi, “Revolutionizing agriculture with arti fi cial intelligence : plant disease detection methods , applications , and their limitations,” *Front. Plant Sci.*, vol. 15, no. March, hal. 1–20, 2024, doi: 10.3389/fpls.2024.1356260.

[9] S. Bandopadhyay, A. K. Gaula, I. Haider, dan G. Kumar, “Chili-Net: An Approach for Classifying Chili Leaf Diseases Using Deep Neural Networks,” *Digit. Interact. Mach. Intell.*, hal. 45–55, 2024, doi: https://doi.org/10.1007/978-3-031-66594-3_5.

[10] E. Science, “Traditional and current-prospective methods of agricultural plant diseases detection : A review Traditional and current-prospective methods of agricultural plant diseases detection : A review,” *3rd Int. Conf. Agric. Bio-industry (ICAGRI 2021)*, 2022, doi: 10.1088/1755-1315/951/1/012002.

[11] A. Shrestha, A. Mahmood, dan S. Member, “Review of Deep Learning Algorithms and Architectures,” *IEEE Access*, vol. PP, no. c, hal. 1, 2019, doi: 10.1109/ACCESS.2019.2912200.

[12] J. Xu, Y. Cui, S. Zhang, dan M. Zhang, “The evolution of precision agriculture and food safety : a bibliometric study,” *Front Sustain. food Syst.*, no. December, hal. 1–26, 2024, doi: 10.3389/fsufs.2024.1475602.

[13] G. Natanael, T. Suparyanto, dan G. N. Elwirehardja, “Determining Variables Associated with Annual Oil Palm Yield : An Determining Variables Associated with Annual Oil Palm Yield : An Explainable Gradient Boosting Approach Explainable Gradient Boosting Approach,” *Procedia Comput. Sci.*, vol. 227, hal. 262–271, 2023, doi: 10.1016/j.procs.2023.10.524.

[14] T. Hidayat, I. Asti, A. Yaqin, A. Phuk, dan T. Arifianto, “Grouping of Image Patterns Using Inceptionv3 for Face Shape Classification,” *Int. J. INFORMATICS Vis.*, vol. 7, no. December, hal. 2336–2343, 2023.

[15] B. N. Naik, R. Malmathanraj, dan P. Palanisamy, “Detection and classification of chilli leaf disease using a squeeze-and-excitation-based CNN model,” *Ecol. Inform.*, vol. 69, 2022, doi: https://doi.org/10.1016/j.ecoinf.2022.101663.

[16] R. A. Saputra, F. D. Adhinata, N. R. Ghaniaviyanto, dan R. A. Charisma, “A Model Convolutional Neural Network for Early Detection of Chili Plant Diseases in Small Datasets,” *Proceeding 3rd Int. Conf. Electron. Biomed. Eng. Heal. Informatics. Lect. Notes Electr. Eng.*, vol. 1008, hal. 131–143, 2023, doi: https://doi.org/10.1007/978-981-99-0248-4_10.

[17] A. Kaur, V. Kukreja, S. Chamoli, S. Mehta, dan A. Garg, “A Unified Framework for Chili Leaf Disease Classification Using Federated Learning CNN,” *2023 10th IEEE Uttar Pradesh Sect. Int. Conf. Electr. Electron. Comput. Eng.*, hal. 1398–1403, 2023, doi: 10.1109/UPCON59197.2023.10434686.

[18] S. Winiarti dan A. Pujiyanta, “Identification of Chili Plant Diseases Based on Leaves Using Hyperparameter Optimization Architecture Convolutional Neural Network,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 11, hal. 859–865, 2024, doi: https://doi.org/10.14569/IJACSA.2024.0151185.

[19] R. A. Boukabouya, A. Moussaoui, dan M. Berrimi, “Vision Transformer Based Models for Plant Disease Detection and Diagnosis,” *2022 5th Int. Symp. Informatics its Appl.*, hal. 1–6, 2022, doi: 10.1109/ISIA55826.2022.9993508.

[20] H. Yin, A. Vahdat, J. M. Alvarez, A. Mallya, J. Kautz, dan P. Molchanov, “A-ViT: Adaptive Tokens for Efficient Vision Transformer,” *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, doi: 10.1109/CVPR52688.2022.01054.

[21] L. Yuan *et al.*, “Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet,” *2021 IEEE/CVF Int. Conf. Comput. Vis.*, 2021, doi: 10.1109/ICCV48922.2021.00060.

[22] K. Al, F. Gebali, A. Kanan, dan I. T. Chelvan, “Vision transformer architecture and applications in digital health : a tutorial and survey,” *Vis. Comput. Ind. Biomed. Art*, vol. 9, 2023, doi: 10.1186/s42492-023-00140-9.

[23] R. Kursun dan M. Koklu, “International Conference 2024 ViT-AlexNet : A Hybrid Approach for Chili Leaf Disease International Conference 2024,” *Int. Conf. Sci. Technol. Interdiscip. Res.*, hal. 26–35, 2024, [Daring]. Tersedia pada: https://academicresearchlibrary.com/abstract-detail/986.

[24] J.-Y. Cheng, M.-L. Tham, B.-H. Kwan, dan C. K. F. Wong, “Performance evaluation of vision transformer and YOLOv8 in plant disease classification,” *2024 IEEE Int. Conf. Autom. Control Intell. Syst.*, 2024, doi: 10.1109/I2CACIS61270.2024.10649856.

[25] L. Chen, G. Ailimujiang, dan Z. Zhao, *Swin Transformer Based on Image Enhancement Algorithm*, Advances i. 2024.

[26] X. Zhai *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ICLR 2021*, 2021, doi: https://doi.org/10.48550/arXiv.2010.11929.

[27] Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” *2021 IEEE/CVF Int. Conf. Comput. Vis.*, hal. 9992–10002, 2021, doi: 10.1109/ICCV48922.2021.00986.

[28] Z. Tu, H. Talebi, H. Zhang, dan F. Yang, “MaxViT : Multi-Axis Vision Transformer,” *Springer*, hal. 1–28, 2022, doi: 10.48550/arXiv.2204.01697.

[29] M. Oquab *et al.*, “DINOv2 : Learning Robust Visual Features without Supervision,” *Trans. Mach. Learn. Res.*, hal. 1–32, 2024, doi: https://doi.org/10.48550/arXiv.2304.07193.

[30] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, dan Y. Cao, “EVA-02: A visual representation for neon genesis,” *Image Vis. Comput.*, vol. 149, 2024, doi: https://doi.org/10.1016/j.imavis.2024.105171.

[31] D. H. Nugraha, “Chili Plant Disease,” *Kaggle [online]*, 2021. https://www.kaggle.com/datasets/dhenyd/chili-plant-disease (diakses Jun 08, 2025).

[32] S. Sulandari, “PENYAKIT DAUN KERITING KUNING CABAI DI INDONESIA,” *J. Perlindungan Tanam. Indones.*, vol. 12, no. 1, hal. 1–12, 2006, doi: https://doi.org/10.22146/jpti.11941.

[33] M. Kammara dan A. Kumar, “Integrated Approaches for Management of Whitefly (Bemisia tabaci Gennadius) to Prevent Chilli Leaf Curl Disease in Chilli Crop : A Review,” *Int. J. Plant Soil Sci.*, vol. 35, no. 19, hal. 1447–1457, 2023, doi: 10.9734/IJPSS/2023/v35i193689.

[34] A. Steiner, R. Wightman, dan J. Uszkoreit, "How to train your ViT ? Data , Augmentation , and Regularization in Vision Transformers," *arXiv*, no. 18, 2022.

[35] A. Laksito, N. Wiranda, S. Nur, dan M. Hayaty, "The COVID-19 Tweets Classification Based on Recurrent Neural Network," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, no. 1, hal. 358–364, 2024, doi: 10.18517/ijaseit.14.1.18832.

[36] S. Sathyanarayanan dan B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," */africanjournalofbiomedicalresearch*, vol. 27, no. 4, hal. 4023–4031, 2024, doi: https://doi.org/10.53555/AJBR.v27i4S.4345.

[37] J. Brownlee, "Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning," in *Machine Learning Mastery*, 2020.

[38] K. M. Sujon, R. Hassan, K. Choi, dan A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 268, 2025, doi: https://doi.org/10.1186/s40537-025-01313-4.

[39] A. Tafvizi, "Attributing AUC-ROC to Analyze Binary Classifier Performance," *arx*, 2022.

[40] Y. D. Wulandari, L. C. Munggaran, F. A. Setiawan, dan I. Atman, "Chili Leaf Health Classification using Xception Pretrained Model," *Sist. J. Sist. Inf.*, vol. 13, hal. 1084–1093, 2024.

[41] V. K. Pratap dan N. S. Kumar, "High-precision multiclass classification of chili leaf disease through customized EffecientNetB4 from chili leaf images," *Smart Agric. Technol.*, 2023, doi: https://doi.org/10.1016/j.atech.2023.100295.

[42] R. Sistem, A. Naila, M. Ramadhani, G. W. Saraswati, R. T. Agung, dan H. A. Santoso, "Performance Comparison of Convolutional Neural Network and MobileNetV2 for Chili Diseases Classification," *J. RESTI*, vol. 5, no. 158, hal. 940–946, 2026, doi: 10.29207/resti.v7i4.5028.

.