International Journal of Management Science and Information Technology
IJMSIT

E-ISSN: 2774-5694 P-ISSN: 2776-7388 OPEN ACCESS

Volume 6 (2), July-December 2026, 1434-1445
DOI: https://doi.org/10.35870/ijmsit.v6i2.7992

Classification of Ceremonial Plants with Vision Transformer

Ida Bagus Kade Dwi Suta Negara [1*], I Ketut Gede Darma Putra [2], Made Sudarma [3], I Made Sukarsa [4]
[1*, 3] Doctoral Study Program in Engineering Sciences, Faculty of Engineering, Universitas Udayana, Denpasar City, Bali Province, Indonesia
[2, 4] Information Technology Study Program, Faculty of Engineering, Universitas Udayana, Denpasar City, Bali Province, Indonesia

Email: negara.2291011017@student.unud.ac.id [1*], ikgdarmaputra@unud.ac.id [2], sudarmaee@unud.ac.id [3], sukarsa@unud.ac.id [4]

Article history:
Received July 13, 2026
Revised July 24, 2026
Accepted July 28, 2026

Abstract
The rapid advancement of computer vision technology has accelerated the adoption of artificial intelligence in agriculture, particularly for plant image classification tasks. However, the identification of ceremonial plants remains challenging due to the high visual similarity among species and the continued reliance on manual identification methods, which are time-consuming and require expert knowledge. Unlike previous studies that primarily focused on general crop species or plant disease classification, this study specifically investigates the application of the Vision Transformer (ViT) model for the classification of ceremonial plants, which represent culturally significant plant species with distinctive yet visually similar characteristics. An experimental approach was employed using a dataset of 1,244 ceremonial plant images representing seven classes, with the data divided into training, validation, and testing sets at proportions of 70%, 15%, and 15%, respectively. A pretrained Vision Transformer model was fine-tuned by adapting its classification head to the target classes and evaluated using accuracy, precision, recall, F1-score, and confusion matrix metrics. The experimental results demonstrate that the proposed model achieved a test accuracy of 97.33% and an average class accuracy of 97.01%, indicating its effectiveness in learning complex visual representations and accurately distinguishing visually similar ceremonial plant species. These findings demonstrate the feasibility of Vision Transformer for culturally specific plant recognition and provide a reliable baseline for the development of intelligent ceremonial plant identification systems, contributing to the digital preservation of traditional botanical knowledge and AI-based plant recognition applications.

Keywords:
Deep Learning; Image Classification; Ceremonial Plants; Vision Transformer.

1. INTRODUCTION

Recent advancements in computer vision have led to its widespread adoption across numerous domains, including agriculture, biodiversity conservation, and automated plant recognition. One emerging application is the identification of ceremonial plants, which has gained increasing interest because of their cultural significance and their indispensable role in traditional religious ceremonies. Nevertheless, accurately classifying ceremonial plants remains a challenging problem due to the high degree of visual similarity among species, together with variations in leaf morphology, color, and texture, and the scarcity of specialized image datasets. As a result, the identification process is still predominantly performed manually, relying heavily on expert knowledge while requiring considerable time and effort, particularly when large image collections must be analyzed. To overcome these limitations, deep learning has become a promising approach because it can automatically learn hierarchical and discriminative visual representations from digital images, enabling more accurate and efficient plant classification (Ali et al., 2025). To overcome these limitations, deep learning

approaches have become a popular alternative because of their capability to automatically learn and extract complex, distinctive visual features from digital images (Yuardi et al., 2025).

Among the many deep learning architectures available, the Vision Transformer (ViT) has shown outstanding performance in image classification by utilizing a self-attention mechanism that effectively models global spatial relationships across an image. This capability enables ViT to provide a more comprehensive visual representation compared to traditional convolutional models that primarily focus on local features (Bhowmik et al., 2024). In the context of plant identification both for agricultural purposes and for visually similar species such as ceremonial plants Vision Transformer has been shown to significantly improve classification performance over conventional approaches (Miryala & Rasane, 2025). By modeling long-range dependencies across image regions, ViT offers a more holistic understanding of visual patterns, which is particularly beneficial for complex classification tasks (Yuardi et al., 2025; Nabila, L. R., & Wicaksono, A. D. P. 2025).

In research in the field of plants, Vision Transformer has been applied to handle various complex image classification problems. The findings indicate that the Vision Transformer model, when configured with optimal parameters, can accurately classify multiple plant species, particularly in datasets where the visual distinctions between classes are subtle (Yuardi et al., 2025). This finding confirms that Vision Transformer has a good ability to capture complex visual details, which is very needed in advanced classification tasks such as the recognition of ceremonial plant types.

In addition, the Vision Transformer approach has also been used in the classification of plant leaf diseases. Several studies show that Vision Transformer-based models are able to distinguish the condition of healthy leaves and infected leaves with competitive accuracy and precision values, although the visual differences between classes are relatively small (Nabila, L. R., & Wicaksono, A. D. P. (2025). These results show that Vision Transformer is not only effective in the classification of plant species, but also in tasks that require more detailed visual differentiation of high complexity plant images.

Other research revealed that the combination of Vision Transformer with the Convolutional Neural Networks (CNN) model can significantly improve plant classification performance. This hybrid approach is able to utilize CNN's advantages in extracting local features as well as Vision Transformer's ability to capture global relationships, resulting in a more accurate and efficient plant image classification system. (Lee et al., 2023a). Subsequent studies have demonstrated that hybrid deep learning models integrating Convolutional Neural Networks (CNNs) with Vision Transformers can improve plant leaf disease detection performance by combining local feature extraction strengths with global attention mechanisms ("Fine-Grained Plant Classification Using Vision Transformers with Optimized MLP Heads," 2023). Moreover, hybrid CNN–ViT architectures have been introduced to enhance plant disease classification, attaining high accuracy on multi-class leaf disease datasets by combining CNN-based spatial feature extraction with the global contextual representation capabilities of the Vision Transformer (Jawed et al., 2026), (Demir et al., 2025). The flexibility of Vision Transformer architectures has also been demonstrated in agricultural applications beyond species classification, such as fruit maturity detection, where ViT models achieved competitive results over conventional methods (Aboelenin et al., 2025; Esaki et al., 2025).

Recent advances in deep learning have highlighted the growing potential of Vision Transformer (ViT) architectures for plant image classification. Studies have shown that ensemble models integrating Vision Transformers with multiple CNN backbones can achieve high classification performance across diverse plant leaf datasets, demonstrating excellent generalization capabilities for recognizing plant species with varying visual characteristics (Hossain et al., 2024). Moreover, Vision Transformer has proven effective in handling fine-grained classification tasks by capturing subtle differences between visually similar plant classes through its ability to model long-range dependencies and global contextual information, which are often difficult for conventional convolutional networks to represent (Murugavalli & Gopi, 2025). In addition, hybrid CNN–Vision Transformer architectures have consistently delivered superior performance over standalone models by combining the strengths of CNNs in extracting local features with the global attention mechanism of Vision Transformers. This complementary learning strategy enhances classification accuracy, particularly for plant species exhibiting high visual similarity and complex morphological characteristics (Ait Nasser & Akhloufi, 2024). Collectively, these findings demonstrate that Vision Transformer-based models provide a reliable and effective framework for developing accurate and robust automated plant classification systems.

Moreover, recent advancements in lightweight and resource-efficient Vision Transformer hybrids, such as MobilePlantViT, show that transformer-based models can be adapted for deployment on mobile or edge devices while maintaining competitive classification performance. These interconnected findings collectively highlight the growing importance and effectiveness of Vision Transformer and related hybrid approaches in plant image classification problems.

Based on these findings, it can be inferred that the Vision Transformer holds significant potential for automatic plant classification with a high degree of accuracy. Accordingly, this study seeks to investigate the implementation of the Vision Transformer in classifying ceremonial plants, which possess unique visual features and cultural significance, with the expectation of contributing to the advancement of digital agriculture and AI-driven plant recognition systems.

2. RESEARCH METHOD

This section describes the materials and methodology adopted in this study, with a particular focus on the implementation of the Vision Transformer model for ceremonial plant image classification. It presents the overall research workflow, including dataset preparation, image preprocessing, model development and training, as well as the evaluation procedures used to assess classification performance.

2.1. Research Design

This research employs an experimental methodology to investigate the effectiveness of the Vision Transformer model for classifying ceremonial plants from digital images. An experimental design was selected because it provides a systematic framework for training and evaluating the model using a labeled image dataset, allowing its classification performance to be assessed objectively. Similar experimental approaches have been widely adopted in plant image classification research to examine the capability of transformer-based architectures in learning complex visual representations and recognizing intricate patterns from botanical image data. (Lee et al., 2023b; Dhakyanaikabdel, n.d.-a, n.d.-b). Through this approach, the learning behavior of the model can be observed quantitatively using performance metrics derived from training, validation, and testing phases.

All research stages are systematically structured, starting from dataset collection and preparation, proceeding to image preprocessing, model training, and ultimately performance evaluation. This organized workflow aligns with prior studies that highlight the importance of a clearly defined experimental pipeline to ensure reliable evaluation outcomes and strong model generalization in plant classification tasks (Dhakyanaikabdel, n.d.-a). By following these stages, this study aims to provide an objective assessment of the Vision Transformer's ability to recognize and distinguish ceremonial plant categories accurately.

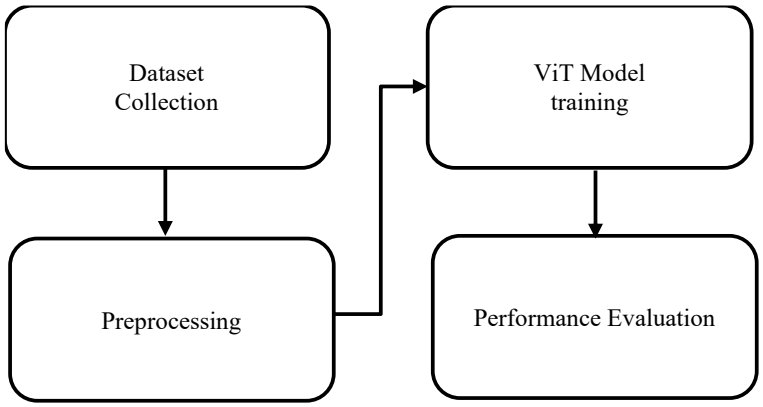

Figure 1. Reaserch Desaign

2.2. Dataset dan Data Source

The dataset employed in this study comprises images of ceremonial plants acquired through field documentation and publicly available sources. These images are organized into multiple categories according to the corresponding ceremonial plant species. To facilitate effective model training and ensure reliable performance assessment, the dataset is partitioned into training, validation, and testing subsets using balanced proportions. This data-splitting strategy enables a comprehensive evaluation of the model's ability to generalize when classifying previously unseen ceremonial plant images.

Table 1. Dataset

Plant Species	Total Number of Images
Yellow Coconut	197
Green Coconut	175
Red Coconut	141
Betel Leaf	205
Nutmeg	190
Sacred Wood Leaf	167
Nagasari Leaf	169
	1244

The dataset used in this study comprises 1,244 images of ceremonial plants collected from both field observations and publicly accessible sources. The images are categorized into seven distinct plant classes and subsequently divided into training, validation, and testing sets with proportions of 70%, 15%, and 15%,

respectively. This data partitioning strategy was adopted to ensure effective model training while providing a reliable assessment of the Vision Transformer's ability to generalize to previously unseen images.

2.3. Data Preprocessing

Before being used in the training process, all images undergo preprocessing to adjust their format and quality according to the requirements of the Vision Transformer model. Image preprocessing typically includes resizing image to a uniform resolution and standardizing pixel values through normalization to ensure consistent input dimension and pixel scale which improves convergence and learning stability in deep learning models (Yenni Research Scholar & Kumar Professor, 2025), (Salabi & Manthila, n.d.). During the preprocessing stage, data augmentation techniques such as rotation, flipping, and contrast adjustment are applied to artificially expand the dataset and increase visual diversity, thereby improving generalization capability and minimizing the risk of overfitting. In addition, contemporary studies have shown that advanced augmentation strategies beyond basic geometric transformations—such as color jittering, random cropping, and GAN-based synthetic augmentations—can further diversify training data and improve classification accuracy for agricultural leaf and plant image datasets (Antwi et al., 2024). These preprocessing steps ensure that the model is expected to recognize the visual patterns of ceremonial plants under various conditions and viewpoints, leading to more robust performance.

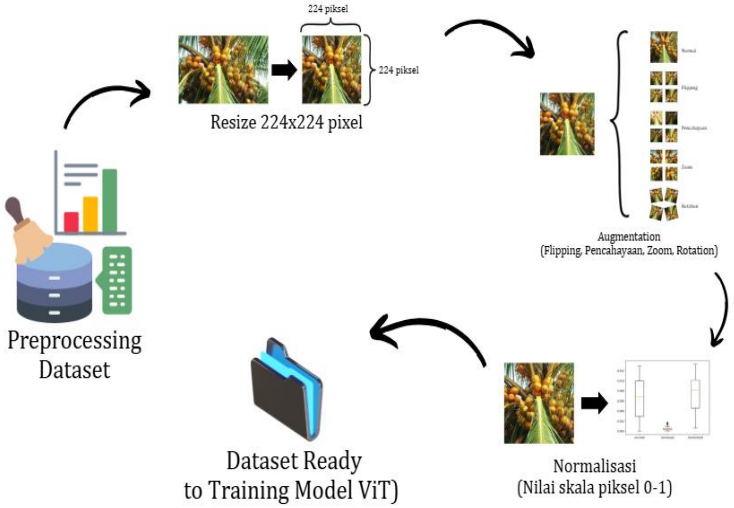

Figure 2. Preprocessing Data

2.4. Vision Transformer Architecture

The model used in this study is the Vision Transformer, configures or plant classification in ceremonies. The model uses pretrained initial weights to accelerate convergence during training. At the end of the architecture, the classification layer is adjusted to match the number of plant classes used in the study. The selection of the Vision Transformer is motivated by its capability to capture global relationships among different regions of an image through its self-attention mechanism.

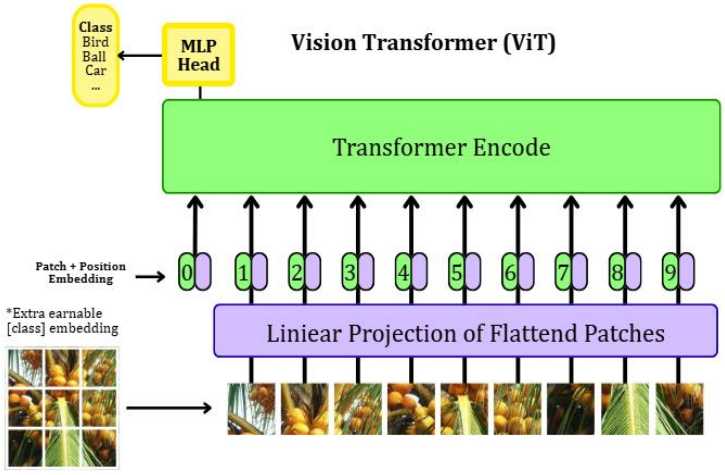

Figure 3. Vision Transformer

2.5. Model Training Process

The model training process utilizes the training dataset along with predefined hyperparameters. The model is trained for a specified number of epochs using an appropriate optimizer and loss function designed for multi-class classification tasks. Throughout the training phase, performance is continuously monitored on the validation set to prevent overfitting and to ensure consistent improvement in classification performance.

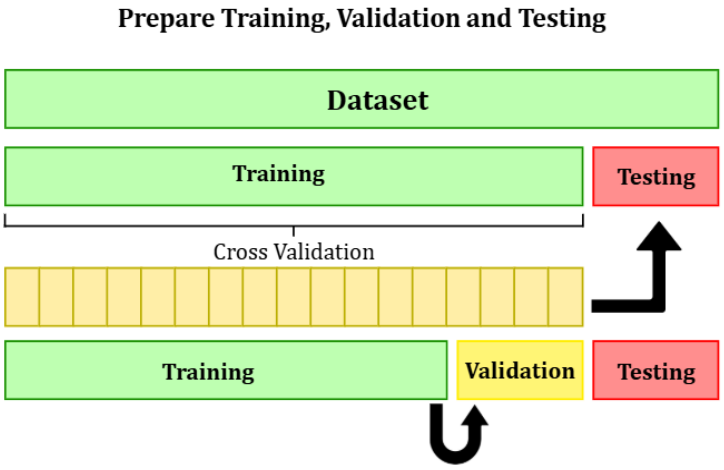

Figure 4. Model Training Process

The Vision Transformer model was fine-tuned using the Adam optimization algorithm with an initial learning rate of 0.0001 and a batch size of 32. Multi-class classification was performed using the CrossEntropyLoss function as the optimization objective. The training process was conducted over 100 epochs with all input images resized to 224 × 224 pixels prior to model training. Throughout the training phase, model performance on the validation set was continuously monitored to assess convergence and reduce the risk of overfitting. To enhance the diversity of the training data and improve the model's ability to generalize to unseen samples, several data augmentation techniques, including random rotation, horizontal flipping, and contrast enhancement, were applied.

2.6. Evaluation Metrics

Following the completion of the training process, the Vision Transformer model was evaluated using the test dataset to measure its overall classification performance. The evaluation was conducted using multiple performance metrics, including accuracy, precision, recall, and F1-score, while a confusion matrix was employed to analyze classification errors among the ceremonial plant classes. The combination of these evaluation metrics provides a comprehensive assessment of the model's predictive capability and class-wise performance. The entire experimental workflow, encompassing data preprocessing, model training, and performance evaluation, was implemented in Python using several deep learning libraries and executed within the Google Colaboratory environment. This cloud-based platform facilitates efficient computational resource management, supports reproducible experiments, and provides an effective environment for developing and evaluating Vision Transformer–based deep learning models.

3. RESULTS AND DISCUSSION

This section presents the experimental results obtained from the implementation of the Vision Transformer model for ceremonial plant classification and provides a detailed discussion of the findings. The analysis focuses on evaluating the model's classification performance, interpreting the obtained results, examining the distribution of correctly and incorrectly classified samples, and comparing the proposed approach with findings reported in previous studies. Through this comprehensive evaluation, the effectiveness of the Vision Transformer model in recognizing ceremonial plant images can be thoroughly assessed.

3.1. Vision Transformer Training Results

This section presents the training outcomes of the Vision Transformer (ViT) model developed for ceremonial plant classification. The reported results provide an initial evaluation of the model's ability to learn discriminative visual features and capture representative patterns from the ceremonial plant image dataset. To achieve stable convergence and reliable learning performance, the model was trained for 100 epochs using both the training and validation datasets, allowing continuous monitoring of the learning process and optimization of the model throughout the training phase.

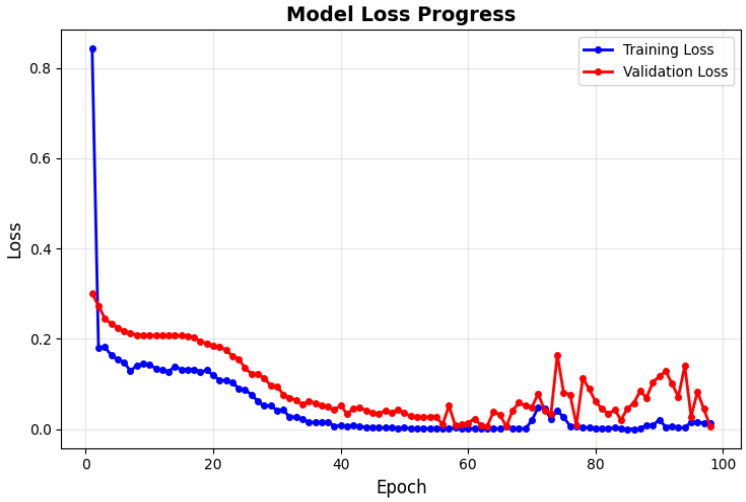

Figure 5. Loss Model Accuracy

Figure 5 illustrates the progression of training and validation loss over the course of 100 training epochs for the Vision Transformer (ViT) model. At the beginning of the training process, the training loss is comparatively high; however, it decreases steadily as training advances, demonstrating the model's ability to effectively learn representative visual features from the ceremonial plant images. A similar trend is observed for the validation loss, which also declines consistently throughout the training process. This behavior indicates that the model generalizes well to previously unseen data and suggests that the learning process remains stable without evidence of significant overfitting.

In the later epochs, the training loss stabilizes at a low value, while the validation loss exhibits minor fluctuations. These variations remain within an acceptable range and do not indicate instability or overfitting, confirming that the training process is stable and effective for the ceremonial plant classification task.

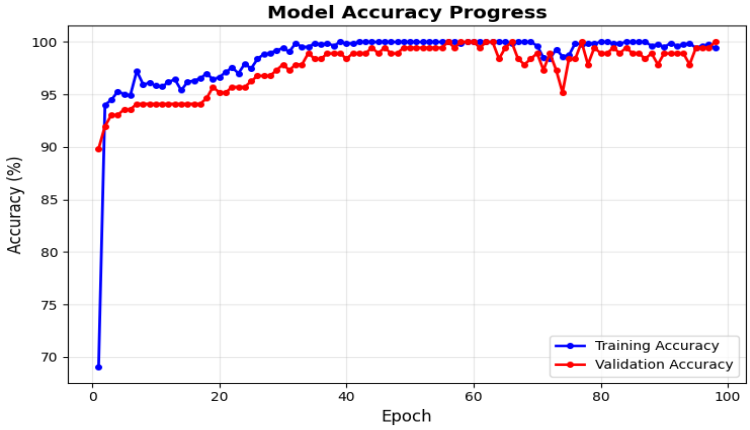

Figure 6 Model Accuracy Progress

Figure 6 presents the progression of training and validation accuracy throughout the 100 training epochs of the Vision Transformer (ViT) model. During the initial training stages, both training and validation accuracies increase rapidly, indicating that the model effectively learns the fundamental visual features of ceremonial plant images. This early improvement demonstrates that the selected Vision Transformer architecture and training configuration are well suited to the characteristics of the dataset. As training continues, both accuracy curves gradually converge to high and stable values, reflecting a consistent learning process and improved classification performance. The relatively small gap between the training and validation accuracy curves suggests that the model generalizes well to unseen data and does not exhibit significant overfitting. Although slight fluctuations in validation accuracy are observed at several epochs, these variations remain within an acceptable range and do not adversely affect the overall learning stability. The final results indicate that the Vision Transformer model achieves high and consistent classification accuracy after 100 training epochs, demonstrating its effectiveness and robustness in recognizing ceremonial plant images.

Table 2. Accuracy Table

Metric	Value
Test Accuracy	97.33%
Best Validation Accuray	100.00%
Aveerage Class Accuracy	97.01%

Table 2 presents the overall classification performance of the Vision Transformer (ViT) model on the ceremonial plant dataset. The model achieved a test accuracy of 97.33%, indicating its strong ability to accurately classify previously unseen ceremonial plant images and demonstrating excellent generalization performance. In addition, the highest validation accuracy reached 100.00%, showing that the model was capable of achieving perfect classification on the validation dataset during specific stages of the training process. This result reflects the effectiveness of the training strategy and suggests that the model successfully learned discriminative visual representations from the input images. Furthermore, the average class accuracy of 97.01% highlights the model's consistent performance across all ceremonial plant categories. The relatively high accuracy achieved for each class indicates that the model performs reliably not only on the dominant classes but also on the remaining categories, confirming its robustness and balanced classification capability across the entire dataset.

3.2. Model Evaluation Result

The evaluation was conducted to assess the ability of the Vision Transformer (ViT) model in classifying seven classes of ceremonial plants based on digital imagery. The dataset used consisted of 1,244 images, with the division of training data as many as 870 images, validation data of 187 images, and test data of 187 images. The evaluation process was focused on measuring the generalization capabilities of the model using test data that were not involved during the training.

Based on the test results, the Vision Transformer model managed to achieve a test accuracy of 97.33%. This value indicates that most of the plant imagery on the test data can be correctly classified by the model. This high performance is also supported by the best validation accuracy of 100%, which indicates that the training process takes place effectively and stably.

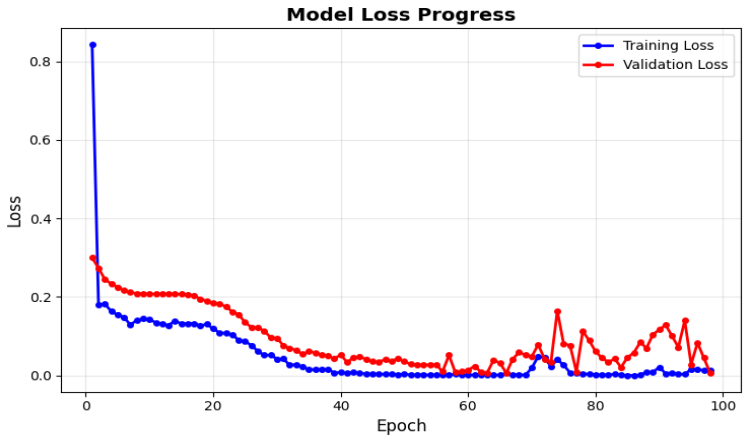

Figure 7. Model Loss Progress

The progression of training loss and validation loss throughout the training process is illustrated in Figure 7 (Model Loss Progress). The graph shows a significant decrease in losses in the early phases of training, then stabilized at very low values until the end of the season. This pattern shows that the model is able to learn the visual features of the plant well without experiencing significant indications of overfitting.

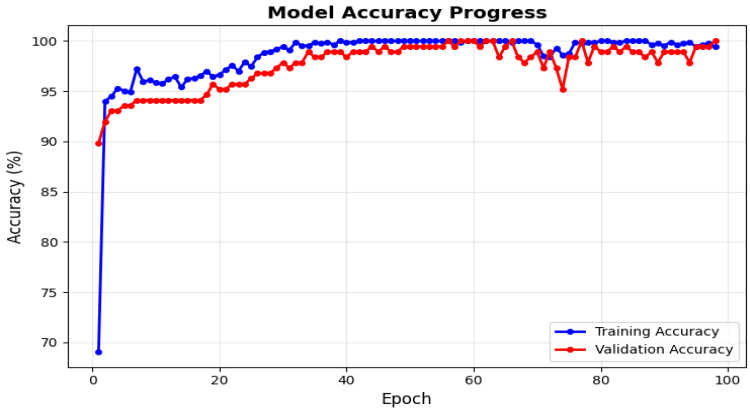

Figure 8. Model Accuracy Progress

Consistent with this observation, Figure 8 (Model Accuracy Progress) demonstrates a rapid increase in both training and validation accuracy to values approaching 100%, with only a minimal gap between them, indicating strong generalization capability of the model.

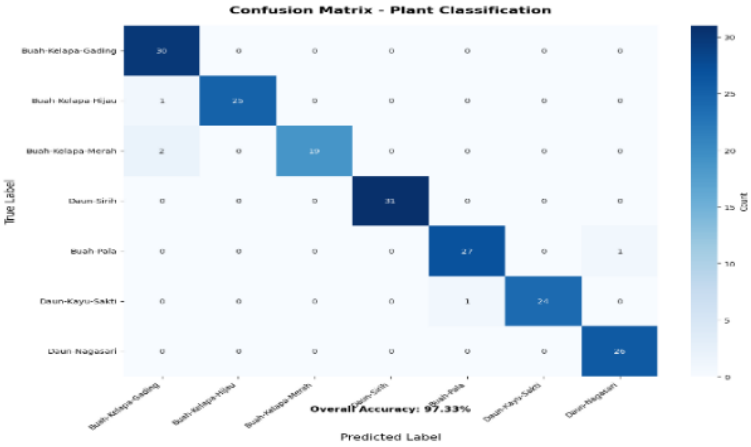

Figure 9. Confusion Matrix - Plan Classification

To further assess the classification performance of the proposed model, a confusion matrix analysis was conducted, as illustrated in Figure 9. The matrix shows that most predictions are located along the principal diagonal, indicating that the Vision Transformer correctly classified the majority of ceremonial plant images into their respective categories. The limited number of misclassified samples can largely be attributed to the substantial visual resemblance among several plant classes. For example, Red Coconut and Yellow Coconut exhibit similar color characteristics, fruit morphology, and surface textures, making them challenging to distinguish, particularly under varying lighting conditions. Furthermore, variations in image acquisition, including differences in background composition, camera perspective, and illumination, may introduce visual inconsistencies that increase classification difficulty. Despite these challenges, the confusion matrix demonstrates that the proposed Vision Transformer model preserves a high level of discriminative capability across all plant classes, confirming its robustness and effectiveness in recognizing ceremonial plant images.

Table 3. Detailed Classification Report

	Precision	Recall	F1-score	Support
Yellow Coconut	0.9091	1.0000	0.9524	30
Green Coconut	1.0000	0.9615	0.9804	26
Red Coconut	1.0000	0.9048	0.9500	21
Betel Leaf	1.0000	1.0000	1.0000	31
NutMeg	0.9643	0.9643	0.9643	28
Sacred Wood Leaf	1.0000	0.9600	0.9796	25
Nagasari Leaf	0.9630	1.0000	0.9811	26
Accuracy			0.9733	187
Macro avg	0.9766	0.9701	0.9725	187
Weight avg	0.9749	0.9733	0.9733	187

To further evaluate the classification performance of the proposed Vision Transformer model, a detailed classification report is presented in Table 3. The results indicate that the model achieved consistently high precision, recall, and F1-score values across nearly all ceremonial plant classes, demonstrating its strong capability to distinguish between different categories. Notably, the Betel Leaf and Nagasari Leaf classes achieved perfect precision and recall scores of 1.00, reflecting flawless classification performance for these categories. In comparison, the Red Coconut class recorded the lowest recall value of 0.9048, suggesting that a small proportion of Red Coconut images were misclassified as other classes. This outcome is likely associated with the substantial visual similarity between Red Coconut and other coconut varieties, particularly Yellow Coconut, which share comparable color characteristics, shape, and texture. Nevertheless, the overall classification report confirms that the proposed model delivers accurate, balanced, and reliable performance across all ceremonial plant classes.

Table 4. Per-Class Accuracy Analysis

	Accuracy	Samples
Yellow Coconut	100.00%	30
Green Coconut	96.15%	21
Red Coconut	90.48%	21
Betel Leaf	100.00%	31
Nutmeg	96.43%	28
Sacred Wood Leaf	96.00%	25
Nagasari Leaf	100.00%	26

The per-class accuracy analysis shown in Table 4 (Per-Class Accuracy Analysis) shows that three classes, namely Yellow Coconut, Betel Leaf, and Nagasari Leaf, achieved a perfect accuracy of 100%. Meanwhile, the Red Coconut class had the lowest accuracy of 90.48%, which indicates that there are challenges in distinguishing this class from other classes that have similar visual characteristics. The average accuracy per class obtained was 97.01%, which indicates consistent model performance across all classes.

Figure 10. Sample Predictions on Test Set

To complement the quantitative evaluation, Figure 8 presents representative prediction results generated by the Vision Transformer model on the test dataset. The visualization demonstrates that the model is capable of producing accurate predictions with high confidence across images acquired under different viewing angles and background conditions, highlighting its robustness in handling variations in image acquisition. The few misclassified examples further illustrate the model's limitations, particularly when distinguishing between ceremonial plant classes with highly similar visual characteristics. Overall, the experimental findings indicate that the proposed Vision Transformer model achieves excellent classification performance, supported by high predictive accuracy, a stable training process, and reliable visual prediction results. These outcomes confirm the strong potential of transformer-based architectures for ceremonial plant image classification. Nevertheless, the relatively lower performance observed in a small number of visually similar classes suggests that future improvements could be achieved by expanding the dataset, increasing the number of samples for challenging categories, and applying more advanced data augmentation techniques to further enhance the model's discriminative capability and generalization performance.

3.3. Comparison with Previous Studies

When compared with previous studies based on conventional Convolutional Neural Network (CNN) architectures, the proposed Vision Transformer (ViT) model demonstrates competitive classification performance. This superior performance can be attributed to the self-attention mechanism, which enables the model to effectively capture long-range feature relationships and global contextual information within the input images. Such capability is particularly advantageous for differentiating ceremonial plant classes that exhibit subtle visual differences and high inter-class similarity. These results suggest that the Vision Transformer represents a promising and effective alternative to traditional CNN-based approaches, especially for fine-grained plant image classification tasks that require robust feature representation and enhanced discriminative performance.

Table 5. Comparison of Classification Accuracy between the Proposed Vision Transformer Model and Previous Deep Learning Models

Model	Accuracy	Source
ResNet50	94.00%	(Syihad et al., 2023)
EfficientNet	96.60%	(Farman et al., 2022)
CNN	95.62%	(Rahman et al., 2025)
ViT	97.33%	This Study

The comparison results demonstrate that the Vision Transformer (ViT) outperforms conventional Convolutional Neural Network (CNN)-based architectures in the task of ceremonial plant classification. This enhanced performance is primarily attributed to the self-attention mechanism, which enables the model to effectively capture long-range feature dependencies and global contextual relationships across the entire image. Such capability allows the Vision Transformer to distinguish between ceremonial plant classes with subtle visual differences, including similarities in color, shape, and texture, more effectively than traditional convolution-based approaches. Consequently, the proposed model exhibits superior discriminative performance for fine-grained ceremonial plant image classification and represents a promising solution for the development of intelligent plant recognition systems, particularly in applications related to digital agriculture, biodiversity conservation, and cultural heritage preservation.

4. CONCLUSION

The findings of this study demonstrate that the Vision Transformer (ViT) model is highly effective for the classification of ceremonial plant images, achieving high accuracy and consistent performance across all evaluation metrics. The experimental results indicate excellent generalization capability, as evidenced by the high-test accuracy and balanced classification performance across all plant categories. The use of pretrained weights, together with the adaptation of the final classification layer to the target number of ceremonial plant classes, enables the model to learn rich and discriminative visual representations. This architectural configuration allows the model to capture both global contextual information and subtle inter-class differences, thereby improving its ability to distinguish visually similar ceremonial plant species.

A key contribution of this research is the application of the Vision Transformer architecture to the classification of ceremonial plants using a culturally significant image dataset. While previous studies have predominantly focused on general plant species or plant disease recognition, this work addresses the more specialized task of ceremonial plant identification, which is inherently challenging because of the high degree of visual similarity among classes and the limited availability of dedicated datasets. The experimental results confirm that the Vision Transformer effectively extracts discriminative features for fine-grained classification and achieves competitive performance compared with previously reported deep learning approaches.

The evaluation further demonstrates that the proposed model performs reliably across both majority and minority classes, highlighting its robustness in multi-class classification tasks. The consistently high precision, recall, and F1-score values indicate a stable learning process and balanced predictive capability for all ceremonial plant categories. Collectively, these findings confirm that the Vision Transformer is a reliable and scalable solution for automated ceremonial plant recognition, with considerable potential for supporting intelligent plant identification systems in digital agriculture, biodiversity conservation, and AI-assisted environmental monitoring.

Future work will focus on extending the dataset by incorporating additional ceremonial plant species and images acquired under more diverse environmental conditions to further improve model robustness and generalization. Moreover, subsequent studies will investigate the performance of hybrid deep learning architectures, particularly CNN–Vision Transformer models, and explore the deployment of the proposed model on mobile and edge computing platforms to enable real-time ceremonial plant recognition and facilitate practical applications in digital agriculture and the preservation of cultural heritage.

REFERENCES

Aboelenin, S., Elbasheer, F. A., Eltoukhy, M. M., El-Hady, W. M., & Hosny, K. M. (2025). A hybrid Framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer. Complex and Intelligent Systems, 11(2). https://doi.org/10.1007/s40747-024-01764-x

Ait Nasser, A., & Akhloufi, M. A. (2024). A Hybrid Deep Learning Architecture for Apple Foliar Disease Detection. Computers, 13(5). https://doi.org/10.3390/computers13050116

Ali, M., Salma, M., Haji, M. El, & Jamal, B. (2025). Plant disease detection using vision transformers. International Journal of Electrical and Computer Engineering (IJECE), 15(2), 2334. https://doi.org/10.11591/ijece.v15i2.pp2334-2344

Antwi, K., Bennin, K. E., Pobi Asiedu, D. K., & Tekinerdogan, B. (2024). On the application of image augmentation for plant disease detection: A systematic literature review. In Smart Agricultural Technology (Vol. 9). Elsevier B.V. https://doi.org/10.1016/j.atech.2024.100590

Bhowmik, A. C., Ahad, M. T., Emon, Y. R., Ahmed, F., Song, B., & Li, Y. (2024). A customised vision transformer for accurate detection and classification of Java Plum leaf disease. Smart Agricultural Technology, 8. https://doi.org/10.1016/j.atech.2024.100500

Demir, A., Sarı, F., Arzu, M., & Kaya, M. (2025). Artificial Intelligence-Aided Diagnosis in Agriculture: Plant Disease Classification with Vision Transformer and CNN Models. NATURENGS MTU Journal of Engineering and Natural Sciences Malatya Turgut Ozal University, 6(2), 22–31. https://doi.org/10.46572/naturengs.1832476

Dhakyanaikabdel, K. (n.d.-a). A Hybrid CNN-Vision Transformer Framework for Optimized Leaf Disease Detection Using Feature Fusion and Transfer Learning. Journal of Robotics and Control (JRC), 6(6), 2025. https://doi.org/10.18196/jrc.v6i6.28426

Dhakyanaikabdel, K. (n.d.-b). A Hybrid CNN-Vision Transformer Framework for Optimized Leaf Disease Detection Using Feature Fusion and Transfer Learning. Journal of Robotics and Control (JRC), 6(6), 2025. https://doi.org/10.18196/jrc.v6i6.28426

Esaki, I., Noma, S., Ban, T., Sultana, R., & Shimizu, I. (2025). Maturity Classification of Blueberry Fruit Using YOLO and Vision Transformer for Agricultural Assistance †. Horticulturae, 11(10). https://doi.org/10.3390/horticulturae11101272

Farman, H., Ahmad, J., Jan, B., Shahzad, Y., Abdullah, M., & Ullah, A. (2022). Efficientnet-based robust recognition of peach plant diseases in field images. Computers, Materials and Continua, 71(1). https://doi.org/10.32604/cmc.2022.018961

Fine-Grained Plant Classification using Vision Transformers with Optimized MLP Heads. (2023). Eltikom.

Hossain, M. A., Sakib, S., Abdullah, H. M., & Arman, S. E. (2024). Deep learning for mango leaf disease identification: A vision transformer perspective. Heliyon, 10(17). https://doi.org/10.1016/j.heliyon.2024.e36361

Jawed, M. M., Tufail, F. A., Ahmed, M. Z., R, A. S., Nallusamy, P., & Raja, K. T. (2026). A hybrid deep learning framework using convolutional and transformer models for robust plant disease classification. Scientific Reports. https://doi.org/10.1038/s41598-026-38209-z

Lee, C. P., Lim, K. M., Song, Y. X., & Alqahtani, A. (2023a). Plant-CNN-ViT: Plant Classification with Ensemble of Convolutional Neural Networks and Vision Transformer. Plants, 12(14). https://doi.org/10.3390/plants12142642

Lee, C. P., Lim, K. M., Song, Y. X., & Alqahtani, A. (2023b). Plant-CNN-ViT: Plant Classification with Ensemble of Convolutional Neural Networks and Vision Transformer. Plants, 12(14). https://doi.org/10.3390/plants12142642

Miryala, S., & Rasane, K. (2025). Enhancing sugarcane leaf disease classification using vision transformers over CNNs. Discover Artificial Intelligence, 5(1). https://doi.org/10.1007/s44163-025-00340-7

Murugavalli, S., & Gopi, R. (2025). Plant leaf disease detection using vision transformers for precision agriculture. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-05102-0

Rahman, K. N., Banik, S. C., Islam, R., & Fahim, A. Al. (2025). A real time monitoring system for accurate plant leaves disease detection using deep learning. Crop Design, 4(1). https://doi.org/10.1016/j.cropd.2024.100092

Nabila, L. R., & Wicaksono, A. D. P. (2025). Klasifikasi Penyakit Pada Tanaman Daun Singkong Menggunakan Vision Transformer: Klasifikasi Penyakit Pada Tanaman Daun Singkong Menggunakan Vision Transformer. POSITIF: Jurnal Sistem dan Teknologi Informasi, 11(2). https://doi.org/10.31961/positif.v11i2.14953

Salabi, L., & Manthila, P. (n.d.). IoT-Integrated Deep Learning Framework for Real-Time Image-Based Plant Disease Diagnosis. National Journal of Signal and Image Processing, 6(1), 19–24. https://doi.org/10.31838/jvcs/06.01

Syihad, I. R., Rizal, M., Sari, Z., & Azhar, Y. (2023). CNN Method to Identify the Banana Plant Diseases based on Banana Leaf Images by Giving Models of ResNet50 and VGG-19. Jurnal RESTI, 7(6). https://doi.org/10.29207/resti.v7i6.5000

Yenni Research Scholar, K., & Kumar Professor, K. V. (2025). PLANT DISEASE DETECTS BASED ON MACHINE LEARNING ALOGRITHMS (Vol. 10). www.ijnrd.org

Yuardi, K., Alfarisy, G. A. F., & Ramadhan Paninggalih. (2025). Fine-Grained Plant Classification using Vision Transformers with Optimized MLP Heads. Jurnal ELTIKOM: Jurnal Teknik Elektro, Teknologi Informasi Dan Komputer, 9(2), 130–139. https://doi.org/10.31961/eltikom.v9i2.1500