Elinvo (Electronics, Informatics, and Vocational Education)
Vol 10, No. 2, 2025, pp. 216-227
ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

Implementation of Vision Transformer (ViT) Method in Identifying Orchid Genus Based on Flower Images

Arie Vatresia [1*], Seprina Dwi Cahyani [2], Agus Susanto [3], Atra Romeida [4]

[1,2,3,4] University of Bengkulu, Bengkulu, Indonesia
[1] arie.vatresia@unib.ac.id*; [2] seprinacahyani24@gmail.com; [3] agus.susanto@unib.ac.id; [4] atraromeida@unib.ac.id
* corresponding author

Article Info

Article history:
Received Month dd, yyyy
Revised Month dd, yyyy
Accepted Month dd, yyyy
Available Online dd, yyy

Keywords:
Orchids; identification; vision transformer

Abstract

There are about 15,000 to 20,000 orchid species around the world, spread across more than 900 genera. They come in many different shapes, sizes, and colors. This wide range of species makes it hard to tell them apart, especially for people who aren't experts. Bengkulu is one of the provinces on the island of Sumatra. It is known for its historical and cultural heritage as well as its rich biodiversity, especially its native plants like orchids. However, it is still hard to tell what they are. The identification process can be made better by using the breakthrough in artificial intelligence of the transformer. The goal of this study is to create an Android app that can use the Vision Transformer (ViT) architecture to identify five types of orchids: Bulbophyllum, Cymbidium, Dendrobium, Phalaenopsis, and Vanda. We used open-source libraries to collect data, which included 1,500 images that went through preprocessing steps. The experimental results show that the ViT-Base16 model with 25 epochs did the best job, getting an accuracy of 0.98 on the test dataset. However, it was hard to classify the genus Dendrobium in all trials because it had a lot of different shapes. The application testing gave good results, with scores of 81.13 for ease of use, 82.5 for accuracy, and 83.06 for usefulness. These results indicate that the application successfully aids in the identification of orchid genera, serving as a useful resource for both educational and practical applications.

INTRODUCTION

Bengkulu is a province on the island of Sumatra that is known for its historical and cultural heritage as well as its rich biodiversity, especially its native plants like orchids. In 2018, about 160 different types of orchids were found in Bengkulu City [1]. This shows that the province is an important place for orchid diversity. Orchids (Orchidaceae) are a group of plants that includes about 15,000 to 20,000 species in almost 900 genera. Many of these plants only grow in forests around the world [2]. Orchids are valuable as ornamental plants because they come in many different flower shapes, colors, sizes, and other unique morphological traits [3]. They can live in a wide range of places, from lowlands to mountains, tropical rainforests, and temperate climates [4]. The morphology of orchids is profoundly shaped by geographical and environmental factors in their indigenous habitats [5], which enhance their beauty and uniqueness, rendering them highly coveted by plant enthusiasts and the horticultural industry [6]. Even though they are beautiful, it can be hard to tell them apart because of how different they look. This issue is especially clear for people who are new to growing orchids, as misidentifying them can

lead to improper care and handling, which can slow plant growth or even kill the plant. Because of this, there is a strong need for an automated, reliable, and quick way to identify orchids. Recent improvements in deep learning and computer vision give us hope for solving this problem. In this case, digital images are the main source of data. Each image is shown as a grid of pixels with brightness or color values [7]. Image enhancement and other pre-processing methods are very important for making sure that data sets are consistent by reducing irrelevant variations like differences in lighting and color imbalance, which could otherwise make deep learning models less effective [8].

This study employs the Vision Transformer (ViT) to tackle the orchid identification challenge. ViT is a version of the Transformer architecture that was first made for Natural Language Processing (NLP) and has done very well in computer vision applications [9]. ViT uses the self-attention mechanism to capture global contextual information across the whole image [10], [11]. This is different from Convolutional Neural Networks (CNNs), which only look at local receptive fields. This feature allows ViT to perform at the highest level on a variety of vision tasks, such as classification, detection, and segmentation [12]. The goal of this study is to create a ViT-based system that can automatically classify orchid genera based on pictures of flowers. A confusion matrix and classification report are used to check how well the model works and make sure it can reliably and accurately identify things. Also, the suggested method is built into an Android app, making it easy for orchid lovers, researchers, and horticulturists to use in real life. ViT is especially useful for classifying orchids because there is a lot of variation within each class and small differences between classes in color patterns and petal textures. The global self-attention mechanism of ViT helps the model find long-range dependencies across the whole flower structure. This is often hard to do with CNNs, which mostly use local feature extraction.

METHODS

This study emphasizes a specialized technical methodology designed for vision-based categorization instead of utilizing a generic data mining framework. The workflow consists of four steps: (1) gathering and curating the dataset, (2) preprocessing and augmenting the images, (3) setting up the model architecture and training, and (4) testing and deploying the model. The Cross Industry Standard Process for Data Mining (CRISP-DM) was used in this study. It is a standard and systematic way to plan data-driven projects. There are six main steps in CRISP-DM: understanding the business, understanding the data, preparing the data, modeling, assessing, and deploying. Each phase is connected to the others, making it possible to make improvements over time to ensure that the system works and is reliable. The goal of this study was to use the CRISP-DM method and the Vision Transformer (ViT) to make it easier to identify orchid genera. This included collecting and cleaning data, building a model, and putting it into an Android app. This study focuses on a technical workflow that is specific to vision-based classification rather than a general-purpose data mining framework. The workflow has four steps: (1) getting and curating the dataset, (2) preprocessing and augmenting the images, (3) setting up the model architecture and training, and (4) testing and deploying the model. The Cross Industry Standard Process for Data Mining (CRISP-DM) was used as a guide for this research. It is a structured and systematic way to plan data-driven projects. There are six main steps in CRISP-DM: understanding the business, understanding the data, preparing the data, modeling, evaluating, and deploying. Each phase is linked to the others, which makes it possible to make improvements over time to make sure the system works well and is reliable. This study aimed to establish a clear workflow for orchid genus identification using the Vision Transformer (ViT) by adopting CRISP-DM. The workflow includes data acquisition and preprocessing, model development, and deployment in an Android-based application.

This study focuses on a specific technical approach for vision-based categorization instead of a general data mining framework. The workflow consists of four steps: (1) gathering and organizing the dataset, (2) preprocessing and adding to the images, (3) setting up the model architecture and training,

ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

and (4) testing and deploying the model. This study used the Cross Industry Standard Process for Data Mining (CRISP-DM) method, which gives a systematic and standardized way to carry out data-driven projects. The six main steps in CRISP-DM are: learning about the business, learning about the data, preparing the data, modeling, evaluating, and deploying. Each phase is linked to the others, which means that the system can be improved over time to make sure it works well and is dependable. This project aimed to create a clear method for identifying orchid genera using the CRISP-DM method and the Vision Transformer (ViT). This included gathering and cleaning data, building and using a model in an Android app. This research is about a specific technical method for vision-based categorization, not a general-purpose data mining framework. The process consists of four steps: (1) acquiring and organizing the dataset, (2) preprocessing and augmenting the images, (3) configuring the model architecture and training, and (4) evaluating and deploying the model. The Cross Industry Standard Process for Data Mining (CRISP-DM) was used as the method for this study. CRISP-DM is a way to plan data-driven projects that is organized and systematic. CRISP-DM has six main steps: getting to know the business, getting to know the data, getting the data ready, modeling, evaluating, and deploying. Each phase is connected to the next, which allows for iterative improvement to make sure that the system that was built works well and is dependable. This study utilized CRISP-DM to establish a definitive methodology for identifying orchid genera using the Vision Transformer (ViT). This involved everything from gathering and cleaning data to creating and putting the model into an Android app.

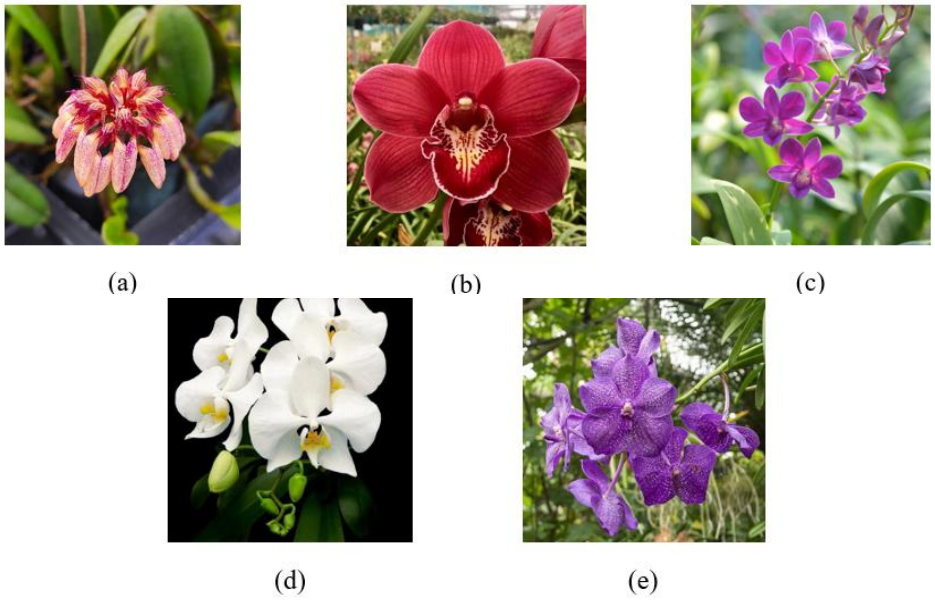


Figure 1. Example of Data (a) Bulbophyllum, (b) Cymbidium, (c) Dendrobium, (d) Phalaenopsis, (e) Vanda

During the data preparation phase, each image was labeled with the name of the orchid genus it belonged to. After that, all of the pictures were resized to 384 × 384 pixels to fit the input requirements of the Vision Transformer (ViT) architecture [13]. The dataset was then split into two parts: 80% for training and 20% for validation, which is standard practice in machine learning research [14]. An independent test dataset was employed exclusively for the assessment of the final model's performance to ensure an unbiased evaluation [15]. Data augmentation was performed using the Roboflow platform to increase the diversity of the dataset and make the model more robust to changes in lighting, orientation, and other image conditions. During the data preparation stage, each image was given a label based on the type of orchid it was. Next, all of the images were changed to 384 × 384 pixels so that they would work with the Vision Transformer (ViT) architecture [13]. Following standard practices in machine learning research [14], the dataset was then split into two parts: 80% for training and 20% for

ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

validation. To ensure an unbiased evaluation, an independent test dataset was utilized solely for the assessment of the final model's performance [15]. Using the Roboflow platform, we added more data to the dataset to make it more diverse and help the model handle changes in lighting, orientation, and other image conditions.

Table 1. Augmentation Techniques

No.	Augmentation Type	Values
1.	Flip	Horizontal, Vertical
2.	Rotation	Between -30° and +30°
3.	Shear	±20° Horizontal, ±20° Vertical
4.	Saturation	Between -10% and +10%
5.	Brightness	Between -5% and +5%
6.	Exposure	Between -5% and +5%
7.	Noise	Up to 1,5 % of pixels

The modeling technique focused on the Vision Transformer (ViT), which has recently gained popularity in computer vision for its ability to model global contextual dependencies through self-attention mechanisms. We used two pre-trained versions of Vision Transformer: ViT-Base and ViT-Large. We used transfer learning by starting from weights trained on large datasets and then fine-tuning them on the orchid dataset. The model architecture was improved for the new classification task, and hyperparameters were carefully chosen based on past research and real-world tests. Changes included the size of the pictures, the activation functions, the optimizer settings, and the regularization methods to speed up convergence and lower overfitting. This study employed ViT due to its ability to understand global spatial relationships, which is crucial for distinguishing orchid taxa with subtle morphological differences. ViT keeps all the information it needs through its multi-head self-attention method, while CNNs may lose global context because of pooling layers.

Table 2. Hyperparameter

No.	Hyperparameter	Values
1.	Image Size	384, 384, 3
2.	Batch Size	32
3.	Epoch	25 and 50
4.	Optimizer	AdamW
5.	Learning Rate	0.0001
6.	Loss	Categorical Cross-Entropy
7.	Activation	GeLU, Softmax

The model's performance was assessed by a confusion matrix and a classification report. The confusion matrix illustrated categorization results by presenting true positives, true negatives, false positives, and false negatives for each genus [16]. This facilitated the recognition of misclassification trends and offered insights into the model's strengths and flaws. The classification report provided detailed metrics, including precision, recall, F1-score, and support for each class [17], facilitating the evaluation of overall accuracy, management of class imbalance, and classification efficacy for each orchid genus. The deployment phase entailed converting the trained model into an Android mobile application. The approach commenced with the creation of a wireframe, which functioned as the blueprint for the application’s interface layout to guarantee logical organization and user-friendly navigation. The design was subsequently executed in Android Studio, incorporating the trained ViT model into the application. This culminated in a working Android application proficient in real-time orchid genus identification, offering an accessible and practical resource for researchers, orchid aficionados, and students.

ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

RESULT AND DISCUSSION

This work's originality is underscored by a baseline comparison with established classification methods, illustrating the unique benefits of employing Vision Transformer (ViT) for fine-grained orchid genus identification. Table X illustrates that traditional methods like SVM + HOG attained merely 74% accuracy, indicating their inadequate ability to grasp the intricate, multi-scale flower structures characteristic of orchids. A lightweight convolutional model, such as MobileNetV2, achieved a performance enhancement to 88%, although it continued to have challenges with intra-class variability, particularly among genera like Dendrobium and Vanda, which display nuanced morphological distinctions. Conversely, the ViT-Base16 model achieved 98% accuracy, exceeding both baselines by a significant margin. This performance disparity underscores the innovation of utilizing ViT's global self-attention mechanism, which is especially efficacious for botanical images where crucial discriminative features (e.g., petal orientation, venation, and symmetrical structures) extend across remote pixel regions and cannot be entirely captured by local CNN filters. The findings substantiate that employing ViT in this context is not simply a repurposing of a state-of-the-art model, but a technically validated contribution illustrating how transformer-based architectures facilitate new capabilities in fine-grained plant identification tasks that conventional classifiers cannot attain. Comparative analysis using traditional methods indicates that ViT markedly surpasses both SVM and CNN classifiers. SVM encounters difficulties with shape changes and background noise, whereas CNN attains rather high accuracy but does not effectively capture the overarching floral structure that differentiates orchid genera. The global attention mechanism of ViT significantly enhances accuracy, validating its application for this task.

The experimental findings indicated that both Vision Transformer (ViT) models demonstrated robust performance in all testing conditions. The utilization of pre-trained ViT models offered significant benefits by expediting the training process and enhancing accuracy. The results of the four experimental settings are encapsulated in Table 3.

Table 3. Training Results

Model	Epoch	Accuracy	Loss	Val Accuracy	Val Loss
ViT-Base16	25	0.9706	1.0040	0.9833	0.9232
ViT-Base16	50	0.9861	0.3165	0.9800	0.3009
ViT-Large16	25	0.9578	1.1138	0.9833	1.0074
ViT-Large16	50	0.9761	0.4010	0.9833	0.3581

The results show that the ViT-Base16 model with 50 epochs did the best, with the highest training accuracy (0.9861) and the lowest training loss (0.3165). This was seen in the measurements of training accuracy and loss. The fact that its validation accuracy was 0.9800 and its validation loss was 0.3009 shows that it can generalize well. The ViT-Large16 model, with 25 epochs, on the other hand, did the worst job. This was the case even though it had a pretty high validation accuracy of 0.9833. It also had the worst training accuracy (0.9578), the biggest training loss (1.0074), and the biggest validation loss (1.1138). Figure 2 shows the training and validation accuracy and loss curves for each trial in order to give a more complete picture of these results. The findings show that the ViT-Base16 model with 50 epochs did the best, with the highest training accuracy (0.9861) and the lowest training loss (0.3165). It had a validation accuracy of 0.9800 and a validation loss of 0.3009, which shows that it can generalize well. On the other hand, the ViT-Large16 model with 25 epochs had the worst performance. It had a relatively high validation accuracy (0.9833) but the lowest training accuracy (0.9578) and the highest training (1.1138) and validation loss (1.0074). Figure 2 shows the training and validation accuracy and loss curves for all of the experiments to help explain these results.

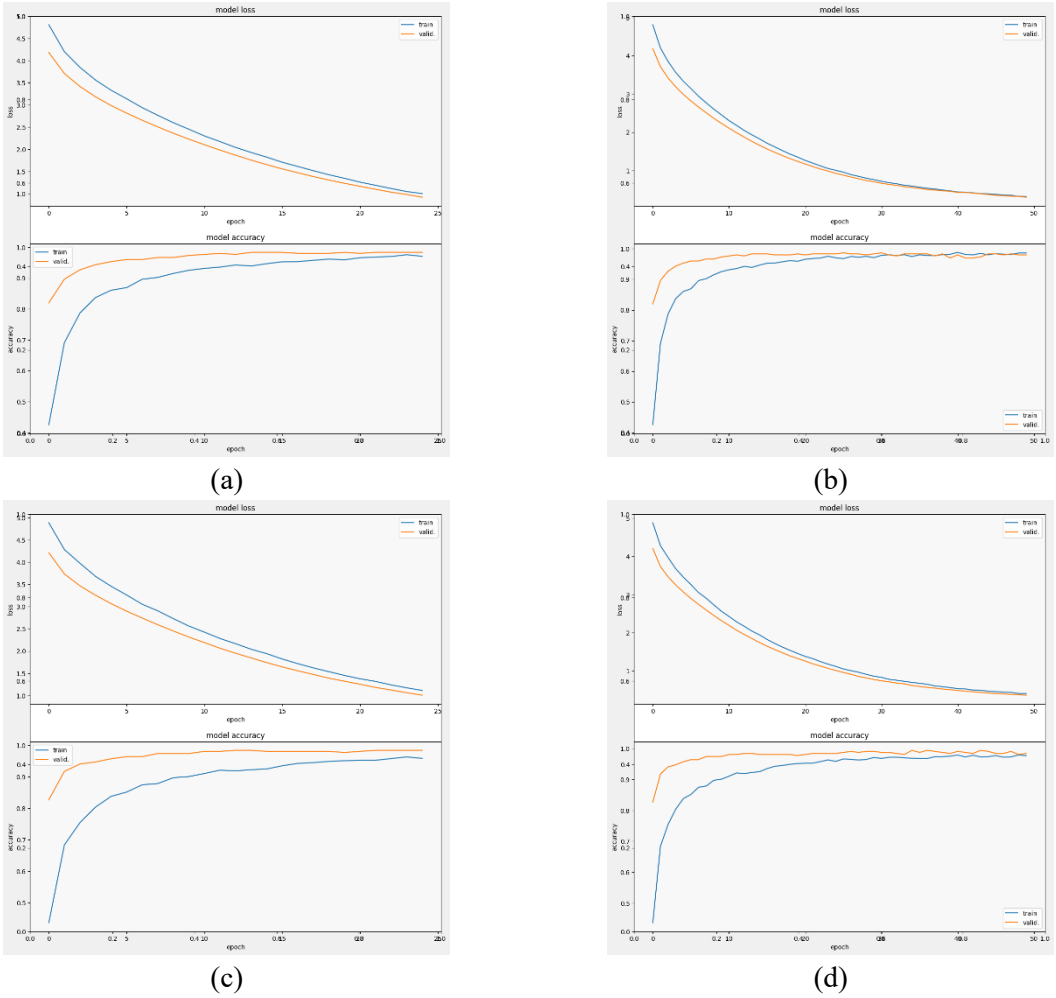


Figure 2. Graphic Result for (a) ViT-Base16 with 25 Epoch, (b) ViT-Base16 with 50 Epoch, (c) ViT-Large16 with 25 Epoch, and (d) ViT-Large16 with 50 Epoch

Throughout the training process, the data showed optimistic outcomes, as the accuracy of the model constantly grew while loss values declined. When compared to the 50-epoch configuration, the 25-epoch configuration for the ViT-Base16 architecture offered a performance that was steadier and balanced. During the early epochs, both configurations demonstrated rapid decreases in training and validation loss, which were thereafter followed by stabilization. The fifty-epoch model, on the other hand, showed indications of overfitting. This was demonstrated by the fact that the gap between the training and validation loss curves was shrinking without there being any proportionate gains in validation accuracy. Based on this, it appears that the model continued to optimize on the training data; however, its capacity to generalize to data that it had not encountered before did not improve anymore. The ViT-Large16 model, on the other hand, demonstrated consistent reductions in both training and validation loss, with only a little amount of space between the two. This implies a superior capacity to manage the complexity of datasets, despite the fact that it needed longer training times and more processing resources.

Evaluation Model Using Test Data

In order to examine the models' ability to generalize to data that they had not encountered, the performance of the models was further evaluated on the independent test dataset. This evaluation reflected the models' efficiency in situations that occur in the real world. The conclusions drawn from

the tests are presented in Table 4, which offers a comprehensive summary of the performance of the model across all orchid genera.

Table 4. Test Data Evaluation Results

Model	Epoch	Bulbophyllum		Cymbidium		Dendrobium		Phalaenopsis		Vanda	
		True	False	True	False	True	False	True	False	True	False
ViT-Base16	25	30	-	30	-	27	3	30	-	30	-
ViT-Base16	50	30	-	30	-	26	4	30	-	29	1
ViT-Large16	25	30	-	27	3	25	5	27	3	28	2
ViT-Large16	50	30	-	28	2	25	5	28	2	28	2

The data show that each model was able to perfectly classify Bulbophyllum, which shows how unique the plant's visual features are and how easy they are to recognize. The ViT-Base16 model, which was trained for 25 epochs, showed the best performance of all the configurations tested, with only a few minor misclassifications in the Dendrobium class. The ViT-Large16 model, which was trained for 25 epochs, had the lowest accuracy of all the models because it made more mistakes in classifying things. The Dendrobium class has been the hardest to work with so far. This is probably because it looks a lot like other classes, and each Dendrobium class has a lot of different shapes and colors. Figure 3 shows confusion matrices that give a full picture of prediction distributions, which makes the results more believable. The results show that all models were able to perfectly classify Bulbophyllum, which shows how unique its visual features are and how easy they are to spot. The ViT-Base16 model trained for 25 epochs had the best performance of all the configurations tested. It only made a few mistakes when classifying Dendrobium. On the other hand, the ViT-Large16 model, which was trained for 25 epochs, had the lowest accuracy because it made more mistakes. The Dendrobium class was the hardest of all the models, probably because it looked a lot like other classes and had a lot of variation in shape and color within the class. The confusion matrices in Figure 3 back up these findings even more by showing how the predictions are spread out in great detail. The results show that every model was able to perfectly classify Bulbophyllum. This is because the plant has very unique visual traits that are easy to spot. The ViT-Base16 model that was trained for 25 epochs did the best overall, with only a few small mistakes in the Dendrobium class. This was better than the other configurations that were tested. The ViT-Large16 model, which was trained for 25 epochs, had the lowest accuracy compared to the other models because it made more mistakes. The Dendrobium class was the hardest of all the models. This is probably because it looks a lot like other classes, and each Dendrobium class has a wide range of shapes and colors. The confusion matrices in Figure 3 also show a full picture of prediction distributions, which makes the results even more reliable. The findings demonstrate that all models attained flawless accuracy in classifying Bulbophyllum, underscoring its exceptionally distinctive visual traits that are readily identifiable. The ViT-Base16 model trained for 25 epochs had the best overall performance among the tested configurations, with only a few misclassifications in the Dendrobium class. The ViT-Large16 model, on the other hand, had the lowest accuracy after 25 epochs of training because it made more mistakes. The Dendrobium class was the hardest for all of the models. This is probably because it looked a lot like other classes and had a lot of shape and color differences within the same class. The confusion matrices in Figure 3, which show the distributions of predictions in great detail, back up these observations even more.

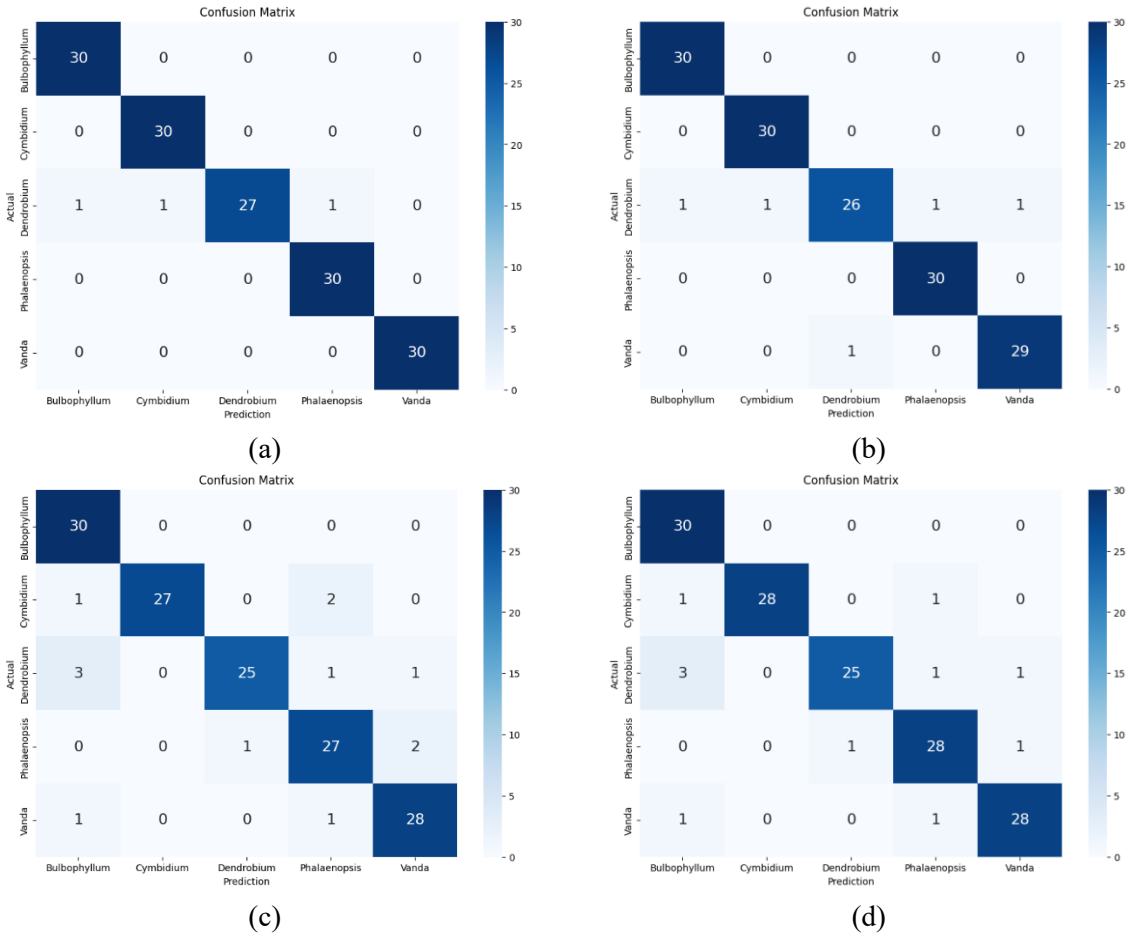


Figure 3. Confusion Matrix for (a) ViT-Base16 with 25 Epoch, (b) ViT-Base16 with 50 Epoch, (c) ViT-Large16 with 25 Epoch, and (d) ViT-Large16 with 50 Epoch

The evaluation employing the classification reports (Figure 4) further substantiated the models' strong performance. The ViT-Base16 model, which was trained for 25 epochs, had the highest overall accuracy of 0.98. This showed that it consistently had high precision, recall, and F1-scores across all orchid genera. On the other hand, the ViT-Large16 model, which was trained for 25 epochs, had the lowest accuracy of 0.91, mostly because its F1-score in the Vanda class was lower (0.89). The ViT-Base16 consistently outperformed the ViT-Large16 in this test, which shows how important it is to pick the right architecture and number of epochs to get the best accuracy, generalization, and computing efficiency. ViT-Base16 is a better choice for medium-sized datasets with limited processing power because of the computing needs and dataset size. The categorization reports (Figure 4) showed that the models did quite well again. The ViT-Base16 model trained for 25 epochs had the highest overall accuracy of 0.98, and it had consistently high precision, recall, and F1-scores for all orchid genera. The ViT-Large16 model, on the other hand, had the lowest accuracy (0.91) after 25 epochs of training. This was mostly because the F1-score in the Vanda class was lower (0.89). In this test, the ViT-Base16 consistently did better than the ViT-Large16. This shows how important it is to choose the right architecture and number of epochs to have the right balance between accuracy, generalization, and computing economy. ViT-Base16 is a better and more practical choice for medium-sized datasets with limited computing power because it is less demanding on computers and works better with larger datasets.

	precision	recall	f1-score	support
Bulbophyllum	0.97	1.00	0.98	30
Cymbidium	0.97	1.00	0.98	30
Dendrobium	1.00	0.90	0.95	30
Phalaenopsis	0.97	1.00	0.98	30
Vanda	1.00	1.00	1.00	30
accuracy			0.98	150
macro avg	0.98	0.98	0.98	150
weighted avg	0.98	0.98	0.98	150

(a)

	precision	recall	f1-score	support
Bulbophyllum	0.97	1.00	0.98	30
Cymbidium	0.97	1.00	0.98	30
Dendrobium	0.96	0.87	0.91	30
Phalaenopsis	0.97	1.00	0.98	30
Vanda	0.97	0.97	0.97	30
accuracy			0.97	150
macro avg	0.97	0.97	0.97	150
weighted avg	0.97	0.97	0.97	150

(b)

	precision	recall	f1-score	support
Bulbophyllum	0.86	1.00	0.92	30
Cymbidium	1.00	0.90	0.95	30
Dendrobium	0.96	0.83	0.89	30
Phalaenopsis	0.87	0.90	0.89	30
Vanda	0.90	0.93	0.92	30
accuracy			0.91	150
macro avg	0.92	0.91	0.91	150
weighted avg	0.92	0.91	0.91	150

(c)

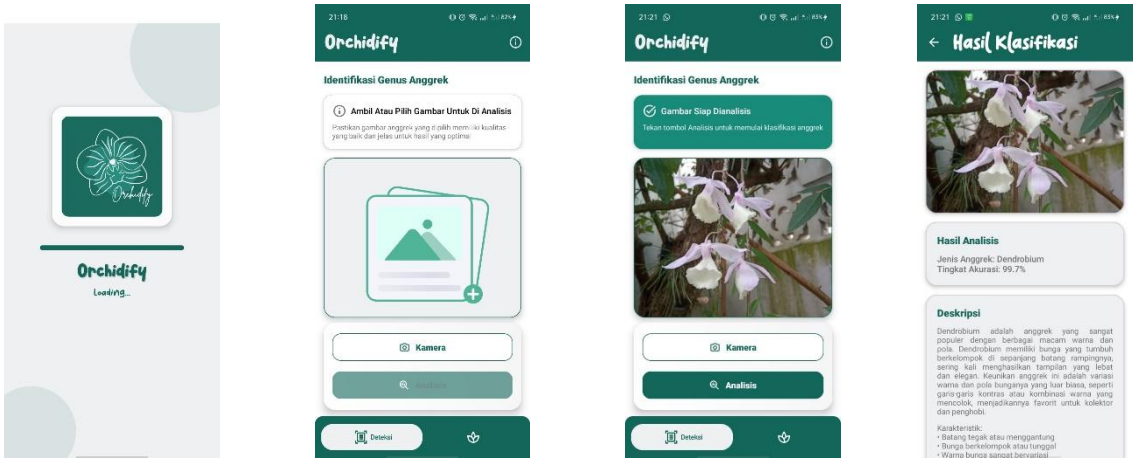
	precision	recall	f1-score	support
Bulbophyllum	0.86	1.00	0.92	30
Cymbidium	1.00	0.93	0.97	30
Dendrobium	0.96	0.83	0.89	30
Phalaenopsis	0.90	0.93	0.92	30
Vanda	0.93	0.93	0.93	30
accuracy			0.93	150
macro avg	0.93	0.93	0.93	150
weighted avg	0.93	0.93	0.93	150

(d)

Figure 4. Classification Report for (a) ViT-Base16 with 25 Epoch, (b) ViT-Base16 with 50 Epoch, (c) ViT-Large16 with 25 Epoch, and (d) ViT-Large16 with 50 Epoch

Deployment

The deployment method involved converting the learned model to TensorFlow Lite (TFLite) format and incorporating it into Android Studio for mobile application development. The ViT-Base16 model, trained for 25 epochs, was chosen for deployment due to its optimal mix of accuracy and efficiency as determined by previous assessments. The mobile application comprises numerous primary features, as depicted in Figure 5. The splash screen initializes the system before accessing the main menu. On the detection menu screen, users can either shoot an image using the camera or upload one from the gallery. The chosen image is presented on the pre-analysis page when the detection process commences. The results page displays the recognized orchid genus, prediction confidence, and a concise description. When the input is not an orchid, the system generates a "non-orchid" output. Supplementary features comprise an orchid information menu that offers comprehensive characteristics and environment descriptions for each genus, along with an information page that provides user recommendations for best application utilization.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5. (a) Splash Screen, (b) Detection Menu Page, (c) Pre-analysis Page; Detection Result Page Showing Two Cases: (d) Detected Orchid, (e) Detected as Non-Orchid; (f) Orchid Menu Page, (g) Detailed Information of Orchid, (h) Information Page

Application Testing

Usability testing was performed with the System Usability Scale (SUS), a standardized tool comprising 10 elements [18]. This study expanded the questionnaire to 16 items, categorized into three groups: convenience, correctness, and usefulness [19]. Every item was evaluated using a five-point Likert scale ranging from Strongly Disagree to Strongly Agree [20]. Scores for positive items were derived by subtracting one from the specified value, whereas for negative items, the value was deducted from five. The overall SUS score was derived by aggregating all item scores, with elevated values signifying superior usability.

Table 3. Training Results

Category	Q	Score (Frequency, n=50)					Avg	Dev	SUS Score
		1	**2**	**3**	**4**	**5**			
Convenience	1	0	0	3	17	30	4.54	0.6	81.13
	2	13	20	15	2	0	2.12	0.84	
	3	0	0	1	21	28	4.54	0.54	
	4	18	21	6	4	1	1.98	0.98	
Accuracy	5	0	0	1	32	17	4.32	0.50	82.5
	6	12	29	9	0	0	1.94	0.64	
	7	0	0	1	24	25	4.48	0.54	
	8	18	31	1	0	0	1.66	0.51	
Usefulness	9	0	0	7	20	23	4.32	0.71	83.06
	10	23	19	8	0	0	1.7	0.73	
	11	0	0	12	18	20	4.16	0.78	
	12	19	19	12	0	0	1.86	0.77	
	13	0	0	1	9	40	4.78	0.46	
	14	35	14	0	1	0	1.34	0.58	
	15	0	0	14	18	19	4.12	0.79	
	16	32	5	2	8	3	1.9	1.36	

The SUS evaluation findings demonstrated commendable performance in all categories, with scores over 70, signifying a high degree of usefulness [21]. The Convenience category attained a score of 81.13, Accuracy received 82.50, while Usefulness recorded the highest at 83.06. The diminished score in Convenience was chiefly affected by question 2, which pertains to the rapidity of identification findings. Numerous responders indicated delays in the identification process due to the computational

intricacy of the Vision Transformer (ViT) model, especially on devices with constrained specs or under suboptimal network conditions. The results indicate that the application exhibits robust usability and holds promise for expanded application in orchid identification and education, but future enhancements should prioritize increasing processing performance. Notwithstanding the robust performance of the Vision Transformer models, this study has certain limitations that must be recognized. The dataset utilized in the tests was predominantly obtained from open-access sites like Kaggle, where the majority of photographs feature clear, well-composed people with low background interference. This condition inadequately reflects real-world situations where orchid photos are generally taken in natural settings featuring intricate, varied, or chaotic backdrops. Consequently, the model's resilience in real-world settings may be inferior to what the current assessment results suggest. Subsequent research should involve the creation or compilation of a specialized real-world orchid dataset to guarantee enhanced reliability and practicality in situ for orchid identification.

CONCLUSION

This study illustrates that the choice of a suitable Vision Transformer (ViT) model and the ideal number of epochs considerably influence categorization performance. The ViT-Base16 model, after 25 epochs, obtained optimal results, reaching 98% accuracy on the test dataset and flawless classification for the Bulbophyllum, Cymbidium, Phalaenopsis, and Vanda genera. Misclassification continued within the Dendrobium genus, highlighting difficulties in differentiating apparently identical species. The results underscore the necessity of calibrating training length to prevent underfitting and overfitting. Moreover, ViT-Base16 showed greater suitability for medium-sized datasets and constrained computational resources, achieving good accuracy without necessitating substantial infrastructure. Conversely, ViT-Large, however formidable for extensive datasets, requires substantially greater processing resources, hence restricting its feasibility in resource-limited settings. Usability testing employing the System Usability Scale (SUS) validated the application's efficacy, yielding ratings of 81.13 for convenience, 82.5 for accuracy, and 83.06 for usefulness, suggesting significant potential for wider implementation in orchid teaching and identification activities.

REFERENCES

[1] A. Nursalikah, “Peneliti Temukan 160 Spesies Anggrek di Bengkulu,” tekno.republika.co.id. [Online]. Available: https://tekno.republika.co.id/berita/pfnzzg366/peneliti-temukan-160-spesies-anggrek-di-bengkulu

[2] A. Monawati, D. Rhomadhoni, and N. R. Hanik, “Identifikasi Hama dan Penyakit Pada Tanaman Anggrek Bulan (Phalaenopsis amabilis),” *Florea J. Biol. dan Pembelajarannya*, vol. 8, no. 1, p. 12, 2021, doi: 10.25273/florea.v8i1.9002.

[3] D. P. Pamungkas, “Ekstraksi Citra menggunakan Metode GLCM dan KNN untuk Identifikasi Jenis Anggrek (Orchidaceae),” *Innov. Res. Informatics*, vol. 1, no. 2, pp. 51–56, 2019, doi: 10.37058/innovatics.v1i2.872.

[4] R. P. Putra, “Identifikasi Jenis Tanaman Anggrek Melalui Tekstur Bunga dengan Tapis Gabor dan M-SVM,” *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 6, no. 1, p. 29, 2021, doi: 10.31328/jointecs.v6i1.1746.

[5] B. Prapitasari, “Keanekaragaman dan Kemelimpahan Jenis Anggrek (Orchidaceae) di Resort Selabintana Taman Nasional Gunung Gede Pangrango (TNGGP) Jawa Barat,” *Biosf. J. Biol. dan Pendidik. Biol.*, vol. 5, no. 1, pp. 24–30, Jun. 2020, doi: 10.23969/biosfer.v5i1.2569.

[6] R. E. Wahyuni, “PERANCANGAN SISTEM PAKAR IDENTIFIKASI PENYAKIT DAN HAMA TANAMAN ANGGREK DENGAN METODE CERTAINTY FACTOR,” *Progr. Stud. Tek. Inform. Jur. Tek. Elektro Fak. Tek. Univ. Tanjungpura*, no. 1, pp. 1–5, 2019.

[7] E. Purwandari, *Konsep dan Teori Pengolahan Citra Digital*, vol. 1. 2018. [Online]. Available: https://www.researchgate.net/publication/387055862_Konsep_dan_Teori_Pengolahan_Citra_Digital

[8] E. Purwandari, *Teori dan Aplikasi Pengolahan Citra Digital*. 2019. [Online]. Available: https://www.researchgate.net/publication/387055808_Teori_dan_Aplikasi_Pengolahan_Citra_Digital

[9] R. Uthama, B. Hendrik, M. T. Informatika, F. I. Komputer, U. P. Indonesia, and U. M. Riau, “Vision Transformer untuk Identifikasi 15 Variasi Citra Ikan Koi,” *J. Comput. Sci. Inf. Technol. (CoSciTech)*,

vol. 5, no. 1, pp. 159–168, 2024.
[10] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using Vision Transformers: A survey," *Eng. Appl. Artif. Intell.*, vol. 126, p. 106669, Nov. 2023, doi: 10.1016/j.engappai.2023.106669.
[11] M. A. Leonardi and A. Y. Chandra, "Analisis Perbandingan CNN dan Vision Transformer untuk Klasifikasi Biji Kopi Hasil Sangrai," *J. Media Inform. Budidarma*, vol. 8, pp. 1398–1407, 2024, doi: 10.30865/mib.v8i3.7732.
[12] S. Jamil, M. Jalil Piran, and O.-J. Kwon, "A Comprehensive Survey of Transformers for Computer Vision," *Drones*, vol. 7, no. 5, p. 287, Apr. 2023, doi: 10.3390/drones7050287.
[13] X. Zhai *et al.*, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," *ICLR 2021*, 2021.
[14] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 331–339, 2024, doi: 10.14569/IJACSA.2024.0150235.
[15] S. C. Haynes, P. Johnston, and E. Elyan, "Generalisation challenges in deep learning models for medical imagery: insights from external validation of COVID-19 classifiers," *Multimed. Tools Appl.*, vol. 83, no. 31, pp. 76753–76772, 2024, doi: 10.1007/s11042-024-18543-y.
[16] T. B. Sasongko, "Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Klasifikasi Jalur Minat SMA)," *J. Tek. Inform. dan Sist. Inf.*, vol. 2, no. 2, pp. 244–253, 2016, doi: 10.28932/jutisi.v2i2.476.
[17] R. R. Sani, Y. A. Pratiwi, S. Winarno, E. D. Udayanti, and F. Alzami, "Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia," *J. Masy. Inform.*, vol. 13, no. 2, pp. 85–98, 2022, doi: 10.14710/jmasif.13.2.47983.
[18] W. Welda, D. M. D. U. Putra, and A. M. Dirgayusari, "Usability Testing Website Dengan Menggunakan Metode System Usability Scale (Sus)s," *Int. J. Nat. Sci. Eng.*, vol. 4, no. 3, pp. 152–161, Nov. 2020, doi: 10.23887/ijnse.v4i2.28864.
[19] J.-W. Park, Y.-H. Cho, M.-K. Park, and Y.-D. Kim, "Consumer Usability Test of Mobile Food Safety Inquiry Platform Based on Image Recognition," *Sustainability*, vol. 16, no. 21, p. 9538, Nov. 2024, doi: 10.3390/su16219538.
[20] G. M. Sullivan and A. R. Artino, "Analyzing and Interpreting Data From Likert-Type Scales," *J. Grad. Med. Educ.*, vol. 5, no. 4, pp. 541–542, Dec. 2013, doi: 10.4300/JGME-5-4-18.
[21] A. Z. Rao and M. A. Hasan, "Evaluation of a Chair-Mounted Passive Trunk Orthosis: A Pilot Study on Able-Bodied Subjects," *Sensors*, vol. 21, no. 24, p. 8366, Dec. 2021, doi: 10.3390/s21248366.
[22] M. Spiteri and S. N. Chang Rundgren, "Literature Review on the Factors Affecting Primary Teachers' Use of Digital Technology," *Technol. Knowl. Learn.*, vol. 25, no. 1, 2020, doi: 10.1007/s10758-018-9376-x.
[23] I. K. A. Enriko, F. N. Gustiyana, and G. C. Giri, "LoRA Gateway Coverage and Capacity Analysis for Supporting Monitoring Passive Infrastructure Fiber Optic in Urban Area," *Elinvo (Electronics, Informatics, Vocat. Educ.*, vol. 8, no. 2, pp. 164–170, 2023, doi: 10.21831/elinvo.v8i2.59280.
[24] E. Madona, Yulastri, A. Nasution, and Prayogi, "Implementation of Lora for Controlling and Monitoring Broiler Cage Temperature," *J. Phys. Conf. Ser.*, vol. 2406, no. 1, p. 012009, Dec. 2022, doi: 10.1088/1742-6596/2406/1/012009.