Jurnal ELTIKOM :
Jurnal Teknik Elektro, Teknologi Informasi dan Komputer

Vol. 9, No. 2, December 2025, page. 130-139
ISSN 2598-3245 (Print), ISSN 2598-3288 (Online)
DOI: http://doi.org/10.31961/eltikom.v9i2.1500
Available online at http://eltikom.poliban.ac.id

FINE-GRAINED PLANT CLASSIFICATION USING VISION TRANSFORMERS WITH OPTIMIZED MLP HEADS

Koko Yuardi, Gusti Ahmad Fanshuri Alfarisy*, Ramadhan Paninggalih
Faculty of Science and Information Technology, Institut Teknologi Kalimantan, Balikpapan, Indonesia
e-mail: 11201048@student.itk.ac.id, gusti.alfarisy@lecturer.itk.ac.id, ramadhanpaninggalih@lecturer.itk.ac.id

Received: 11 April 2025 – Revised: 5 June 2025 – Accepted: 13 June 2025

ABSTRACT

Automatic plant species classification is crucial for advancing education and biodiversity conservation. Deep learning models, such as Vision Transformer (ViT), have demonstrated strong performance in plant species classification tasks. However, limited research explored the impact of hyperparameters in the Multi-Layer Perceptron (MLP) head of ViT models for plant-species classification. This study investigated the influence of learning rates, number of neurons, and activation functions on model performance. It also evaluated efficiency in both CPU and GPU environments. The objective was to determine the optimal configuration by analyzing accuracy, F1-score, and computation time. Two ViT models, ViT-B/16 and ViT-L/16, were tested using the VNPlant-200 dataset, which contains 200 plant species. Thirteen activation functions, multiple learning rates, and neuron configurations were examined. The results showed that the Tanh activation function, combined with a learning rate of 10^{-4} and 1024 neurons, yielded the best performance on the ViT-B/16 model, achieving an accuracy of 0.9692 and F1-score of 0.9684. Meanwhile, the Hard Tanh activation function, with a learning rate of 10^{-4} and 256 neurons, delivered the best results on the ViT-L/16 model, achieving an accuracy of 0.9855 and an F1-score of 0.9854. Computational analysis showed that ViT-B/16 achieved an average inference time of 0.0159 seconds on a GPU and 0.8902 seconds on a CPU, while ViT-L/16 took 0.0492 seconds on a GPU and 2.8335 seconds on a CPU. These findings highlight the importance of selecting suitable activation functions, learning rates, and neuron configurations to optimize model performance while maintaining computational efficiency.

Keywords*: deep learning, fine-grained classification, plant species classification, transfer learning, vision transformer.*

I. INTRODUCTION

PLANT classification is a fundamental task in agriculture, biodiversity monitoring, and environmental science. Accurate systems for classifying plant species are imperative for researchers, farmers, and policymakers, with applications ranging from identifying crop diseases to conserving endangered species [1]. However, fine-grained plant classification remains challenging due to subtle morphological differences between species, variations in lighting, occlusion, and the presence of complexity of natural backgrounds [2].

Conventional approaches to automatic plant classification rely heavily on Convolutional Neural Networks (CNNs), which have shown remarkable results in general image recognition tasks [3]. Deep models like ResNet and DenseNet are widely used to extract features from plant datasets. Despite their success, CNN-based methods often struggle to capture global dependencies within an image, which are essential for fine-grained classification tasks such as distinguishing plant species with highly similar morphological features [4].

The advent of the Vision Transformer (ViT) has introduced a transformative shift in computer vision. Unlike CNNs, ViTs process images by dividing them into smaller patches and applying self-attention mechanisms, allowing them to model both local and global relationships within the data [5]. This capability makes ViTs well-suited for fine-grained classification tasks, including plant species identification. However, the performance of the ViT model is sensitive to parameter configurations, including learning rate, the number of neurons in the hidden layer, and the choice of activation functions.

Proper hyperparameters optimization is critical for maximizing performance.

Recent advancement in ViT architectures have introduced variants such as Swin Transformer [5], DeiT [6], and MobileViT [7], which aim to improve computational efficiency and scalability across different visual tasks. These developments highlight the growing relevance of ViT not only for large-scale vision problems but also for resource-constrained environments and mobile deployments.

Despite significant advances in plant classification, much of the existing research focuses on CNN-based models. For example, Yang et al. [8] utilized a NASNetLarge model enhanced with attention mechanisms for fine-grained plant diseases classification, achieving high accuracy but requiring substantial computational resources. Roy et al. [9] employed a DenseNet-integrated YOLOv4 model for fine-grained object detection in tomato plant disease classification, showcasing improved accuracy but facing challenges with inter-class variability. Araújo et al. [10] explored Siamese Convolutional Network for fine-grained plant species classification, improving inter-class recognition but struggling with computational efficiency.

Recent studies have started exploring the application of Vision Transformers architectures for plant classification. Nhut et al. [11] demonstrated the potential of ViT and BEiT models in plant species classification task. In particular, Nhut et al. successfully trained a ViT model on the VNPlant-200 dataset, achieving an accuracy of 98.24 percent using a training protocol with a learning rate of 10^{-6}. However, their research did not explore the impact of alternative hyperparameters, such as variations in learning rates, neuron configurations, or activation functions, which could further improve performance or reduce computational costs.

This study aims to address these gaps by investigating the performance of two Vision Transformer models, ViT-B/16 (base model) and ViT-L/16 (large model), by tuning the hyperparameter in the MLP head on the VNPlant-200 dataset. We tuned the learning rate, number of neurons, and type of activation functions to unveil both model's capabilities. In addition, we investigated the average running time to analyze latency between the models' predictions. The dataset contains 200 plant species with 20,000 images captured under diverse natural conditions, posing significant challenges due to high intra-class similarity and visual variability [12]. This study evaluates the impact of key hyperparameters on model performance. Computational efficiency is also assessed on both CPU and GPU devices to determine practical applicability.

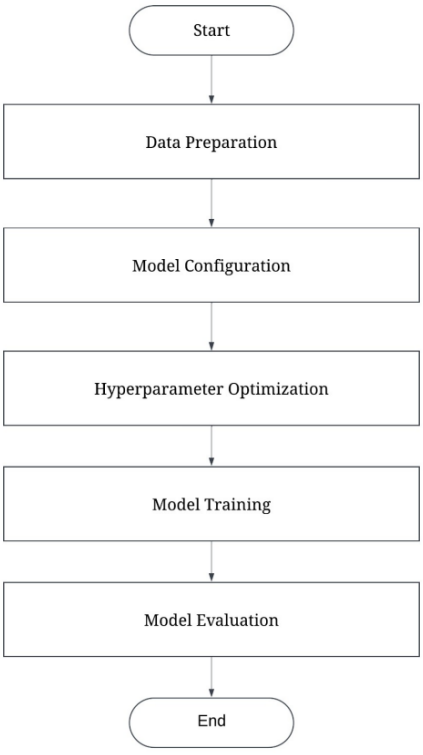

Figure 1. Research Stages Flowchart

The main contribution of this study lies in advancing the performance and efficiency of Vision Transformer–based plant species classification. First, we present an optimized learning rate configuration for the ViT/16 model on the VNPlant-200 dataset, which demonstrates faster convergence compared to previous work [11]. Second, we introduce a baseline design for the MLP head—covering the number of neurons and activation functions—that achieves improved performance over prior studies and produces results competitive with BEiT architectures [11]. Finally, we provide a detailed computational efficiency analysis by reporting average inference times on both CPU and GPU, highlighting the strong dependency of ViT models on GPU resources for real-time plant species classification.

II. RESEARCH METHOD

This study investigates the impact of hyperparameter optimization on the performance of Vision Transformer (ViT) models for fine-grained plant classification. The research stages to be undertaken are illustrated in Figure 1 through a flowchart diagram that presents the experimental procedure. The process consists of four stages: dataset preparation, model configuration, hyperparameter optimization, and evaluation metrics.

A. Dataset and Preprocessing

The VNPlant-200 dataset, consisting of 20,000 images from 200 medicinal plant species, was used in this study. The images were captured under natural conditions, exhibiting variability in lighting, angles, and backgrounds [12]. Various CNN models have been evaluated on the VNPlant-200 images [13], including VGG16, ResNet50, InceptionV3, DenseNet121, and Xception. Among these, the Xception model achieved the highest accuracy, exceeding 88%. Due to this strong performance, the VNPlant-200 dataset is widely regarded as a benchmark for fine-grained plant classification research [13].

1) Data Division

The dataset was divided into training data (60%) and testing data (40%) subsets using stratified sampling to maintain class distribution in both subsets [13]. The training data was used for model training, while the testing data was reserved for performance evaluation.

2) Data Augmentation

To improve generalization and address overfitting, RandAugment was applied to the training set with a magnitude of 30. Augmentation techniques included random rotations, scaling, cropping, and noise addition to simulate diverse conditions [14].

3) Data Normalization

Images were resized to 384x384 pixels for ViT-B/16 model and 512x512 pixels for the ViT-L/16 model. Pixel values were normalized using the mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225], aligned with preprocessing for ImageNet pre-trained weights.

B. Vision Transformer

The Vision Transformer (ViT) is a deep learning architecture inspired by transformer networks originally developed for machine translation. Similar to transformers, ViT learns embeddings through positional encoding followed by a multi-head self-attention mechanism. An image is split into fixed-size patches that undergo linear projection with positional embeddings. Multi-head self-attention is applied to the processed patches, followed by traditional feed-forward networks or a Multi-Layer Perceptron (MLP).

Concretely, the model starts by splitting the image into flattened 2D patches $x_p \epsilon \mathbb{R}^{N \times (P^2.C)}$ where $N = HW/P^2$. H and W denote the height and width of the image, while P refers to the resolution of each image patch. In ViT-B/16, P is 16, which represents the patch resolution (P^2). Meanwhile, C is the number of channels, typically the RGB channels at the input layer.

The flattened patches x_p are projected through trainable parameters (E), followed by positional embedding (E_{pos}), as shown in (1). The multi-head self-attention mechanism is then applied as shown in (2), using layer normalization (LN) from the previous layer, followed by the MLP as expressed in (3). Final classification is performed in the MLP head represented in (4).

$$z_0 = \left[x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{pos}, \quad E \epsilon \mathbb{R}^{(P^2.C)\times D}, E_{pos} \epsilon \mathbb{R}^{(N+1)\times D} \tag{1}$$

$$z_l' = MSA\left(LN(z_{l-1})\right) + z_{l-1}, \quad l = 1, 2, \ldots, L \tag{2}$$

$$z_l = MLP\left(LN(z_l')\right) + z_l' \tag{3}$$

$$y = LN(z_l^0) \tag{4}$$

$$[q, k, v] = zU_{qkv} \tag{5}$$

$$A = softmax(\frac{qk^T}{\sqrt{D_h}}) \tag{6}$$

$$SA(z) = Av \tag{7}$$

$$MSA(z) = [SA_1(z); SA_2(z); \ldots; SA_k(z)]U_{msa} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

The Multi-Head Self-Attention (MSA) block is constructed by concatenating multiple Self-Attention (SA) mechanisms. The embedding z is projected into query (q), key (k), value (v) vectors using matrix U_{qkv} as shown in (5). Attention scores are computed using the SoftMax normalization of the dot product between q and k as expressed in (6). The SA output is obtained by multiplying the attention matrix with the value (v), as shown in (7). MSA is produced by concatenating the outputs of multiple SA heads and applying a projection matrix, as formulated in (8).

Two ViT models, ViT-B/16 and ViT-L/16, were used for feature extraction and classification tasks. These models differ in architecture complexity, with ViT-B/16 serving as the base model and ViT-L/16 as the larger, more complex model. Both models were initialized with pre-trained weights from the IMAGENET1K_SWAG_E2E_V1 dataset, fine-tuned using supervised weakly annotated tags (SWAG) to provide rich visual feature representations. The final classification head of each ViT model was replaced with a custom Multi-Layer Perceptron (MLP) consisting of one hidden layer. The number of neurons and activation function in this layer varied depending on the experimental configuration.

C. Model Training and Experimental Procedure

Both ViT models were trained using the AdamW optimizer with betas set to (0.55, 0.9). The training process ran for 100 epochs to test balance computational efficiency and convergence, with mini-batches of size 16. The study systematically evaluated the impact of three hyperparameters (learning rate, number of neurons in the Multi-Layer Perceptron or MLP, and activation function) to determine their influence on the performance of Vision Transformer models. The learning rate was evaluated first to identify the optimal configuration for stable and efficient training. Four learning rates (10^{-6}, 10^{-5}, 10^{-4}, and 10^{-3}) were tested to assess their effect on accuracy and F1-score. The number of neurons in the MLP was optimized next using the best performing learning rate identified from the previous experiment. Configurations of 64, 128, 256, 512, and 1024 neurons were tested to examine their influence on the model's capacity to learn and represent patterns in the dataset. Finally, activation functions were

TABLE 1
LEARNING RATE PERFORMANCE COMPARISON

Model	Learning Rates	Accuracy	F1 Score
ViT-B/16	10^{-6}	0.8842	0.8743
ViT-B/16	10^{-5}	0.9555	0.9525
ViT-B/16	**10^{-4}**	**0.9597**	**0.9580**
ViT-B/16	10^{-3}	0.9529	0.9512
ViT-L/16	10^{-6}	0.9267	0.9219
ViT-L/16	10^{-5}	0.9809	0.9806
ViT-L/16	**10^{-4}**	**0.9811**	**0.9810**
ViT-L/16	10^{-3}	0.9752	0.9751

evaluated using the best learning rate and the optimal neuron configuration. Thirteen commonly used activation functions, including ReLU, Leaky ReLU, Tanh, Hard Tanh, and Sigmoid, were analyzed to identify their impact on model performance.

D. Evaluation Metrics

The evaluation of the models in this study was conducted to assess their performance in fine-grained plant species classification tasks. Key evaluation metrics, such as the confusion matrix, accuracy, and F1-score, were used to comprehensively analyze the models' performance. The confusion matrix was used to analyze performance at the individual class level. It details the number of correct predictions (true positives), incorrect predictions classified as other classes (false positives), and instances that were not recognized as the correct class (false negatives) [15]. This matrix provides valuable insights into the strengths and weaknesses of the model for specific plant species, aiding in understanding misclassification patterns [16].

Accuracy measures the proportion of correctly classified samples to the total number of samples, as defined in (9). Precision evaluates the proportion of true positive predictions among all samples predicted as positive, calculated using (10). Recall measures the proportion of true positive predictions out of all actual positive samples, providing insight into the model's ability to detect all relevant instances of a class, calculated using (11). The F1-score is the harmonic mean of precision and recall, balancing the trade-off between these two metrics. It is especially valuable when dealing with imbalanced datasets, with the calculation shown in (12).

III. RESULT AND DISCUSSION

A. Result

This subsection presents the results of the experiments conducted on the ViT-B/16 and ViT-L/16 models using the VNPlant-200 dataset. The evaluation metrics include accuracy, F1 Score, and computation time.

1) Learning Rate Evaluation

The effect of varying the learning rate on the ViT-B/16 and ViT-L/16 models was evaluated using the VNPlant-200 dataset. Learning rates ranging from 10^{-6} to 10^{-4} were systematically tested. The results, summarized in Table 1, show that a learning rate of 10^{-4} consistently yielded the highest accuracy and F1-score for both models. Specifically, ViT-B/16 achieved a maximum accuracy of 0.9597 and an F1-score of 0.9580, while ViT-L/16 achieved a maximum accuracy of 0.9811 and an F1-score of 0.9810 at the same learning rate. This configuration facilitated faster convergence compared to lower learning rates, which often led to underfitting.

Significant because it demonstrates that a learning rate of 10^-4 not only provides better stability and faster convergence compared to other learning rates but also outperforms previous research that used a learning rate of 10^{-6}. Prior studies that trained ViT models on the VNPlant-200 dataset with a 10^{-6} learning rate reported a maximum accuracy of 0.9824. By comparison, this study achieves higher accuracy with a more efficient learning rate configuration, highlighting the importance of hyperparameter optimization in fine-grained classification tasks.

Figures 2 and 3 illustrate the training progress of all learning rates across both ViT-B/16 and ViT-L/16. In both figures, the x-axis represents the number of epochs and the y-axis represents accuracy, where 0 corresponds to 0 percent and 1.0 corresponds to 100 percent accuracy. The figures show that a

TABLE 2
NEURON PERFORMANCE COMPARISON

Model	Number of Neurons	Accuracy	F1 Score
ViT-B/16	64	0.9470	0.9455
ViT-B/16	128	0.9564	0.9546
ViT-B/16	256	0.9574	0.9553
ViT-B/16	512	0.9567	0.9550
ViT-B/16	**1024**	**0.9586**	**0.9568**
ViT-L/16	64	0.9708	0.9704
ViT-L/16	128	0.9776	0.9774
ViT-L/16	**256**	**0.9800**	**0.9799**
ViT-L/16	512	0.9795	0.9794
ViT-L/16	1024	0.9796	0.9795

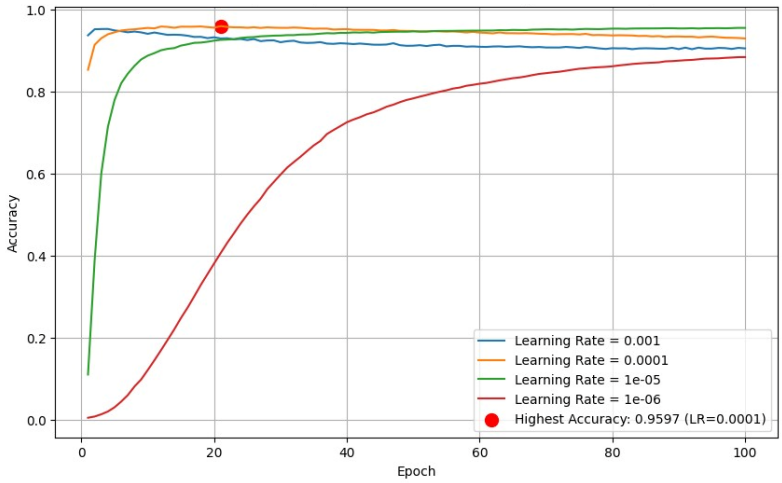

Figure 2. Val Accuracy Over epochs for Different Learning Rates on ViT-B/16

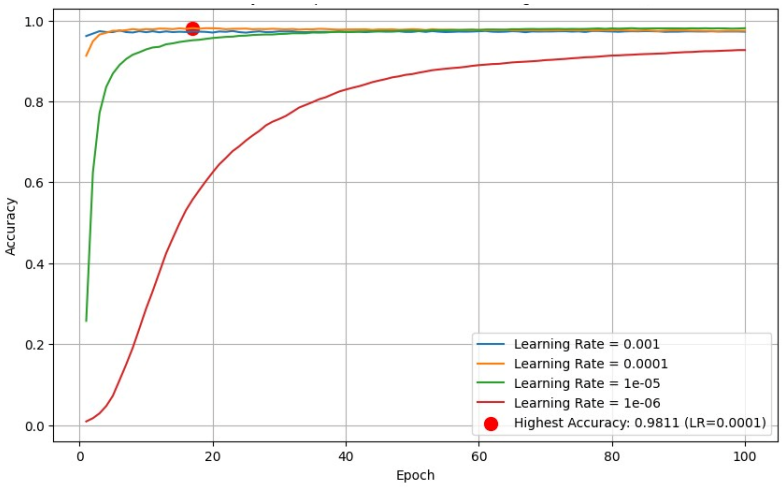

Figure 3. Val Accuracy Over epochs for Different Learning Rates on ViT-L/16

learning rate of 10^{-4} enables the models to achieve optimal performance in significantly fewer epochs compared to other configurations. Models trained with 10^{-4} converge rapidly, reaching peak accuracy within the first 10-20 epochs, while lower learning rates such as 10^{-6} exhibit a slower learning curve and require longer training durations to approach similar performance levels.

2) Neuron Configuration Evaluation

The model performance was further evaluated by varying the number of neurons in the Multi-Layer Perceptron (MLP) layers using a learning rate of 10^{-4}. The results, summarized in Table 2, indicate that ViT-B/16 performed best with 1024 neurons, achieving a maximum accuracy of 0.9586 and an F1-score of 0.9568. In contrast, ViT-L/16 achieved its best performance with 256 neurons, reaching a maximum accuracy of 0.9800 and an F1-score of 0.9799. These findings suggest that increasing the number of neurons beyond these configurations often results in overfitting, particularly on the VNPlant-200 dataset.

TABLE 3
ACTIVATION FUNCTION PERFORMANCE COMPARISON

Model	Activation Function	Accuracy	F1 Score
ViT-B/16	ELU	0.9627	0.9608
ViT-B/16	Hard Sigmoid	0.9585	0.9562
ViT-B/16	Hard Tanh	0.9691	0.9675
ViT-B/16	Leaky ReLU	0.9586	0.9571
ViT-B/16	PReLU	0.9566	0.9547
ViT-B/16	ReLU	0.9595	0.9578
ViT-B/16	RReLU	0.9601	0.9580
ViT-B/16	SeLU	0.9652	0.9637
ViT-B/16	Sigmoid	0.9603	0.9583
ViT-B/16	SiLU	0.9567	0.9556
ViT-B/16	Softplus	0.9570	0.9554
ViT-B/16	Softsign	0.9688	0.9675
ViT-B/16	**Tanh**	**0.9692**	**0.9684**
ViT-L/16	ELU	0.9793	0.9791
ViT-L/16	Hard Sigmoid	0.9821	0.9819
ViT-L/16	**Hard Tanh**	**0.9855**	**0.9854**
ViT-L/16	Leaky ReLU	0.9780	0.9778
ViT-L/16	PReLU	0.9807	0.9805
ViT-L/16	ReLU	0.9800	0.9798
ViT-L/16	RReLU	0.9797	0.9795
ViT-L/16	SeLU	0.9811	0.9809
ViT-L/16	Sigmoid	0.9817	0.9815
ViT-L/16	SiLU	0.9787	0.9785
ViT-L/16	Softplus	0.9782	0.9780
ViT-L/16	Softsign	0.9847	0.9846
ViT-L/16	Tanh	0.9852	0.9851

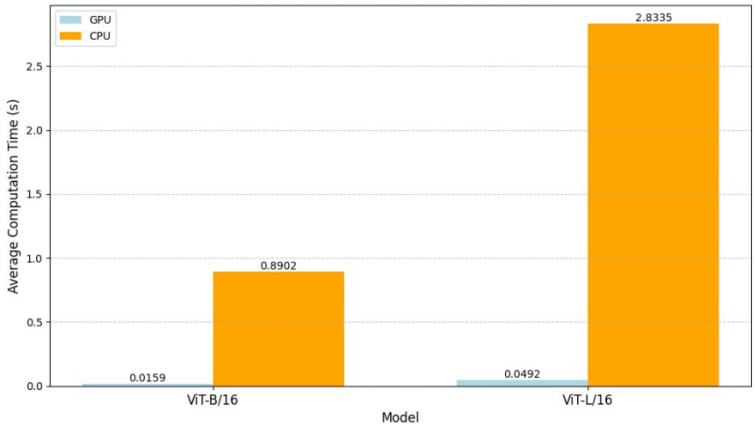

Figure 4. Average Computation Time Comparison

3) Activation Function Experimentation

Various activation functions commonly used in deep learning, including Sigmoid, Tanh, ReLU, and Hard Tanh, were tested in conjunction with each model's optimal number of neurons from the previous experiment. The results in Table 3 show that ViT-B/16 achieved its best performance using the Tanh activation function, with a maximum accuracy of 0.9692 and an F1-score of 0.9684. Meanwhile, ViT-L/16 performed best with the Hard Tanh activation function, achieving a maximum accuracy of 0.9855 and an F1-score of 0.9854. These findings highlight the importance of selecting activation functions that align with the complexity of the dataset and the architecture of the Vision Transformer models.

4) Computation Time

The computational efficiency of the models was compared across CPU and GPU environments, visually represented in Figure 4. In the figure, the bars represent models running on either the CPU (orange) or GPU (blue), while the y-axis indicates the latency or running time in seconds. Computation time was measured by processing 100 random images from the VNPlant-200 dataset. These images were randomly selected from different classes to ensure representative results. The evaluation was conducted using the best performing configuration for each architecture: ViT-B/16 using a learning rate of 10^{-4}, 1024 neurons, and the Tanh activation function, and ViT-L/16 using a learning rate of 10^{-4}, 256 neurons, and the Hard Tanh activation function.

To ensure reproducibility and consistency, all experiments conducted using Google Colaboratory. The hardware environment consisted of an NVIDIA T4 GPU with 16 GB VRAM, an Intel Xeon processor running at 2.20 GHz, and 12.7 GB of system RAM. The software environment included Ubuntu 22.04 as the operating system and Python 3.10 as the programming language. The models were implemented using the PyTorch framework, version 2.5.1 with CUDA 12.1 support (cu212), along with supporting libraries such as NumPy, Matplotlib, scikit-learn, and Torchvision. This configuration reflects a realistic cloud-based deployment setting and enables reliable benchmarking of computational performance across different hardware types.

B. Discussion

The results demonstrate that a learning rate of 10^{-4} provides the best balance between model stability and convergence speed. Lower learning rates, such as 10^{-6}, led to slower generalization, while a higher learning rate of 10^{-3} caused oscillations and reduced generalization performance. This finding aligns with previous studies indicating the importance of appropriately tuning learning rates for transformer-based models. This study also shows that using 10^{-4} as the learning rate results in superior accuracy compared to previous research that utilized 10^{-6}.

The number of neurons in the Multi-Layer Perceptron (MLP) head significantly influence model performance. ViT-B/16, with its simpler architecture, required a larger hidden layer (1024 neurons) to effectively capture complex patterns. In contrast, ViT-L/16, which already possesses a more complex architecture, performed best with fewer neurons (256 neurons). Increasing the number of neurons beyond the optimal configuration led to overfitting, especially for ViT-L/16. These findings suggest that the optimal neuron configuration depends on the complexity of the model architecture. Simpler architectures benefit from a larger hidden layer, while more complex models require fewer neurons to maintain generalization and efficiency.

The selection of activation functions played a crucial role in model performance. Tanh and Hard Tanh outperformed widely used functions such as ReLU and Leaky ReLU. This indicates that activation function selection must align with dataset complexity and model architecture. Tanh and Hard Tanh effectively manage non-linear relationships, contributing to superior performance in fine-grained classification tasks. Unlike ReLU based functions, which can suffer from dead neurons, Tanh and Hard Tanh maintain smooth gradients that help the models capture subtle visual differences more effectively.

The computational efficiency analysis shows a clear trade-off between model accuracy and processing time. ViT-L/16 demonstrated higher accuracy but required 3.1 times more computation time than ViT-B/16 on GPU devices. This trade-off is critical for real-world applications, where resource availability and latency requirements must be considered.

The ViT models in this study surpassed the performance of previous CNN-based approaches, which struggled with fine-grained classification tasks due to their limitations in capturing global dependencies. While this study optimizes learning rates, neuron configurations, and activation function, further research could explore additional hyperparameters such as weight decay and optimizer setting or the number of hidden layers. Testing these configurations on other fine-grained classification datasets could also provide a broader understanding of generalization capabilities.

1) Comparison with Previous Study

In this section, we highlight a comparison between our experiments and the previous study conducted by Nhut et al. [11]. In terms of learning rate, our results showed that a learning rate of 10^{-6} results in slower learning. Based on our empirical findings, a learning rate of 10^{-4} improved the performance while providing faster convergence for both ViT-B/16 and ViT-L/16. Therefore, we recommend 10^{-4} as the primary choice for training ViT models due to their computational complexity.

Regarding performance, the modifications applied to the MLP head improved both accuracy and F1-score. The accuracy achieved in this study reached approximately 0.9855 [11], which is higher than the 0.9824 reported in previous experimentation, an improvement of roughly 0.31%. Additionally, we evaluated the latency of both models, which was not examined in the prior study.

2) Limitations

Despite the promising results achieved in this study, several limitations should be acknowledged. First, the hyperparameter tuning process was performed sequentially, where each hyperparameter was

optimized independently based on the best result from the previous stage. This approach, while efficient, may have overlooked potential interactions between hyperparameters that could further improve performance if tuned jointly.

Second, the experiments were conducted using the VNPlant-200 dataset, which, although diverse and representative, reflects a specific domain of medicinal plants from Vietnam. Therefore, the generalizability of the findings to other plant classification datasets or broader fine-grained classification tasks remains to be validated.

Third, the evaluation of computational performance was limited to a single environment using Google Colaboratory with a T4 GPU and standard CPU settings. Different hardware setups, particularly those with lower specifications or edge devices, might yield different performance trade-offs.

Lastly, this study did not explore the effects of advanced optimization techniques such as learning rate schedulers, weight decay regularization, or different ViT architectures, which could potentially enhance the efficiency and robustness of ViT models in real-world scenarios.

IV. CONCLUSION

This study successfully identified the optimal configurations of Vision Transformer (ViT) models for fine-grained plant species classification using the VNPlant-200 dataset. A learning rate of 10^{-4} provided the best performance for both ViT-B/16 and ViT-L/16. ViT-B/16 achieved peak performance with 1024 neurons and the Tanh activation function, while ViT-L/16 performed optimally with 256 neurons and the Hard Tanh activation function, achieving an accuracy of 0.9855 and an F1-score of 0.9854. Despite ViT-L/16 outperforming ViT-B/16 in accuracy, it required significantly more computation time, highlighting a trade-off between performance and efficiency.

Future research can explore advanced ViT architectures such as ViT-H/16, open-set recognition for unseen species, and smaller patch sizes to improve feature extraction. Additionally, lightweight models like Mobile ViT and Neural Architecture Search (NAS) can optimize performance and efficiency for real-time applications. Expanding the model's evaluation to diverse datasets with higher inter-class variability will further validate robustness for real-world biodiversity monitoring and agricultural tasks.

REFERENCES

[1] M. Itani, M. Al Zein, N. Nasralla, and S. N. Talhouk, 'Biodiversity conservation in cities: Defining habitat analogues for plant species of conservation interest', *PLoS ONE*, vol. 15, no. 6, p. e0220355, Jun. 2020, doi: 10.1371/journal.pone.0220355.

[2] Y. Wu, X. Feng, and G. Chen, 'Plant Leaf Diseases Fine-Grained Categorization Using Convolutional Neural Networks', *IEEE Access*, vol. 10, pp. 41087–41096, 2022, doi: 10.1109/ACCESS.2022.3167513.

[3] S. Ghosh, A. Singh, Kavita, N. Z. Jhanjhi, M. Masud, and S. Aljahdali, 'SVM and KNN Based CNN Architectures for Plant Classification', *Computers, Materials & Continua*, vol. 71, no. 3, pp. 4257–4274, 2022, doi: 10.32604/cmc.2022.023414.

[4] F. Khalid and A. A. Romle, 'Herbal Plant Image Classification using Transfer Learning and Fine-Tuning Deep Learning Model', vol. 35, no. 1, 2024.

[5] Z. Liu *et al.*, 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[6] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, 'Training data-efficient image transformers & distillation through attention'.

[7] S. Mehta and M. Rastegari, 'MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer', Mar. 04, 2022, *arXiv*: arXiv:2110.02178. doi: 10.48550/arXiv.2110.02178.

[8] G. Yang, Y. He, Y. Yang, and B. Xu, 'Fine-Grained Image Classification for Crop Disease Based on Attention Mechanism', *Front. Plant Sci.*, vol. 11, p. 600854, Dec. 2020, doi: 10.3389/fpls.2020.600854.

[9] A. M. Roy, R. Bose, and J. Bhaduri, 'A fast accurate fine-grain object detection model based on YOLOv4 deep neural network', Oct. 30, 2021, *arXiv*: arXiv:2111.00298. Accessed: Oct. 11, 2024. [Online]. Available: http://arxiv.org/abs/2111.00298

[10] V. M. Araujo, A. S. Britto Jr., L. E. S. Oliveira, and A. L. Koerich, 'Two-View Fine-grained Classification of Plant Species', Oct. 04, 2021, *arXiv*: arXiv:2005.09110. Accessed: Oct. 09, 2024. [Online]. Available: http://arxiv.org/abs/2005.09110

[11] D. T. N. Nhut, T. D. Tan, T. N. Quoc, and V. T. Hoang, 'Medicinal plant recognition based on Vision Transformer and BEiT', *Procedia Computer Science*, vol. 234, pp. 188–195, 2024, doi: 10.1016/j.procs.2024.02.165.

[12] T. N. Quoc and V. T. Hoang, 'VNPlant-200 – A Public and Large-Scale of Vietnamese Medicinal Plant Images Dataset', in *Integrated Science in Digital Age 2020*, vol. 136, T. Antipova, Ed., in Lecture Notes in Networks and Systems, vol. 136, Cham: Springer International Publishing, 2021, pp. 406–411. doi: 10.1007/978-3-030-49264-9_37.

[13] T. Nguyen Quoc and V. Truong Hoang, 'Medicinal Plant identification in the wild by using CNN', in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, Korea (South): IEEE, Oct. 2020, pp. 25–29. doi: 10.1109/ICTC49870.2020.9289480.

[14] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, 'RandAugment: Practical automated data augmentation with a reduced search space', 2019, *arXiv*. doi: 10.48550/ARXIV.1909.13719.

[15] D. Chicco, N. Tötsch, and G. Jurman, 'The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation', *BioData Mining*, vol. 14, no. 1, p. 13, Feb. 2021, doi:

10.1186/s13040-021-00244-z.

[16] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, 'Multi-label Classifier Performance Evaluation with Confusion Matrix', in *Computer Science & Information Technology*, AIRCC Publishing Corporation, Jun. 2020, pp. 01–14. doi: 10.5121/csit.2020.100801.