International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 9, September, 2025, pp. 2393-2404
Journal homepage: http://iieta.org/journals/isi

Detection of Coffee Leaf Diseases Using Lightweight Deep Learning: A Comparative Study of *EfficientNet-B0* and Vision Transformer

Alfis Arif*, Ferry Putrawansyah, Idi Jangcik

Informatics Engineering Study Program, Institut Teknologi Pagar Alam, Pagar Alam 31520, Indonesia

Corresponding Author Email: alfisarif@itpa.ac.id

https://doi.org/10.18280/isi.300915

Received: 6 June 2025
Revised: 16 July 2025
Accepted: 22 August 2025
Available online: 30 September 2025

Keywords:
agricultural AI, coffee leaf disease, EfficientNet-B0, field-based diagnosis, lightweight deep learning, Vision Transformer (ViT)

ABSTRACT

This study compares lightweight deep learning models for coffee leaf disease detection using field-acquired images. Two architectures were evaluated: *EfficientNet-B0*, trained from scratch, and Vision Transformer (ViT), fine-tuned from pretrained weights. The dataset consisted of 843 balanced RGB images representing three major coffee leaf diseases: *Leaf spot*, *Rust*, and *Sooty mold*. Images were captured under natural field conditions without synthetic augmentation to ensure realistic evaluation. On the held-out test set, *EfficientNet-B0* achieved an accuracy of 88.37%, precision of 87.9%, recall of 88.02%, and *F1-score* of 87.96%. ViT achieved an accuracy of 85.12%, precision of 84.76%, recall of 84.93%, and *F1-score* of 84.85%. Error analysis indicated that both models struggled to differentiate rust and sooty mold due to overlapping textural patterns. *EfficientNet-B0* showed faster convergence and higher robustness, making it more suitable for mobile and edge deployment. ViT, while slightly less accurate, demonstrated stable learning behavior and potential benefits from larger or more diverse datasets. The results demonstrate feasibility for mobile deployment in real-time field diagnosis, providing a practical benchmark for lightweight AI in precision agriculture.

1. INTRODUCTION

Coffee is one of the most economically significant agricultural commodities in tropical countries, particularly Indonesia. In regions such as Pagar Alam, South Sumatra, coffee cultivation serves as a primary source of livelihood for smallholder communities. However, the productivity and sustainability of coffee plantations are frequently threatened by the emergence of foliar diseases, including rust, *leaf spot*, and sooty mold. These diseases can spread rapidly if not diagnosed promptly, leading to substantial yield losses and reduced bean quality. In Indonesia, outbreaks of coffee rust have been reported to reduce yields by up to 30-40%, while severe cases of *leaf spot* and sooty mold can cause additional losses of more than 20% annually [1]. Therefore, timely and accurate detection of leaf diseases is essential to mitigate economic damage and support precision agriculture practices.

Traditionally, plant disease diagnosis in agriculture has relied on manual visual inspection by farmers or agricultural officers. While this method remains widely used, it is inherently subjective, prone to human error, and often delayed in response [2]. Furthermore, many rural farming areas lack access to trained plant pathologists or agronomists, which highlights the need for automated, reliable, and field-deployable diagnostic tools.

Recent advances in deep learning, particularly in image-based classification, have introduced new opportunities for smart farming and plant disease detection. Convolutional Neural Networks (CNN) have demonstrated outstanding performance in identifying plant diseases due to their capacity to learn local spatial patterns and hierarchical visual features. Among the various CNN architectures, *EfficientNet* has emerged as a leading approach, known for its compound scaling strategy that balances accuracy and computational efficiency [3]. *EfficientNet-B0* was selected in this study because it is the most lightweight variant in the family, offering rapid convergence and low computational demand, which makes it highly suitable for deployment on mobile or edge devices. *EfficientNet-B0*, the most lightweight variant in the family, has been successfully applied in domains such as rice leaf disease detection [4], and cashew nut and fruit disease classification [5], demonstrating both high accuracy and low computational demand. This makes it especially suitable for deployment on edge devices.

In parallel with CNN development, transformer-based models such as the Vision Transformer (ViT) have significantly advanced the field of computer vision by introducing self-attention mechanisms originally designed for natural language processing. ViT has shown competitive, and in some cases superior, performance compared to CNN on large-scale datasets by modeling long-range dependencies and capturing global relationships across image patches [6]. Its application in plant pathology is growing, as demonstrated by Sinamenye et al. [7], who applied ViT to potato leaf disease classification with promising results. ViT was chosen in this study because its global attention mechanism enables the capture of long-range contextual dependencies across leaf surfaces, which can be particularly useful for distinguishing diseases with overlapping or subtle visual symptoms. However, the potential of ViT for smaller-scale agricultural datasets in resource-limited environments remains underexplored, particularly considering the trade-offs in computational cost and inference efficiency.

Several alternative approaches have also been proposed in the literature, including CNN ensembles [8], *ResNet50,* and *VGG16* [9]. While these methods offer advantages in terms of accuracy, many rely on synthetic or heavily augmented datasets that may not reflect real-world conditions. As a result, there is still a need to evaluate lightweight yet high-performing models using authentic, field-acquired images to ensure realistic and effective deployment in practical settings.

This study addresses that need by presenting a comparative benchmark of two compact and efficient architectures: *EfficientNet-B0* and ViT, for classifying coffee leaf diseases. The dataset consists of 843 manually labeled images collected under natural lighting conditions in real farming environments, with equal distribution across the three disease classes. *EfficientNet-B0* is trained from scratch to assess its raw learning capability, while ViT is fine-tuned using pretrained weights from the *Hugging Face Transformers* repository to leverage its global feature representations.

Both models are evaluated independently based on performance metrics such as training accuracy, convergence stability, and loss progression. In addition, visual inspection of prediction results is conducted to analyze qualitative model behavior. The contributions of this study are twofold: (1) it provides a comprehensive benchmark of CNN and Transformer-based lightweight architectures using real-world coffee leaf disease data, and (2) it explores the trade-offs between classification accuracy, computational efficiency, and deployment feasibility in field-based precision agriculture. Unlike previous studies that depend on synthetic data or complex ensemble methods, this work emphasizes realism, model simplicity, and practical relevance for smallholder farming systems.

2. RELATED WORKS

The rapid advancement of deep learning in recent years has significantly transformed the landscape of image-based plant disease classification. By leveraging powerful computational models such as CNN and ViT, researchers have developed increasingly accurate diagnostic systems across various agricultural domains. CNN are widely recognized in computer vision tasks due to their capability to extract local and hierarchical spatial patterns through convolutional feature maps [10]. In contrast, ViT introduces a new paradigm by utilizing self-attention mechanisms to model global spatial relationships across image patches [11], offering a fundamentally different approach compared to conventional CNN-based methods.

Several notable studies have implemented these architectures for plant disease diagnosis under different experimental conditions. For instance, Liu et al. [12] applied *EfficientNet* for cassava disease classification and achieved an accuracy of 88.1% using a dataset collected under controlled laboratory environments. While this highlights the predictive strength of the *EfficientNet* family, the reliance on clean and synthetic image data limits the applicability of such models in uncontrolled, real-world agricultural settings. Similarly, Sinamenye et al. [7] employed a CNN-based on *EfficientNetV2B3* and a ViT to classify potato leaf diseases using synthetically augmented images, achieving an accuracy of 85.06%. Although these models demonstrated strong capabilities in capturing visual dependencies, their performance on field-captured images containing natural lighting, background clutter, and variable leaf orientations remains largely underexplored.

To improve model robustness, some researchers have explored ensemble-based approaches. For example, Yuvalatha et al. [13] employed multiple transfer learning models such as *MobileNetV2, ResNet*, and *VGG16* to classify potato leaf diseases. While *MobileNetV2* achieved 86.8% accuracy individually, combining models using majority voting increased the accuracy to 94.8%. Despite this improvement, ensemble methods often demand greater computational resources, which limits their practicality for deployment on mobile or edge devices in rural farming environments.

Similarly, Grados et al. [14] proposed a deep learning-based method to detect coffee leaf rust by evaluating multiple architectures, including ViT*, NASNet, VGG19,* and *ResNet50*. Their approach incorporated several preprocessing steps such as normalization, segmentation, and scaling. Among the tested models, the ViT achieved the best performance with an accuracy of 92.90%, demonstrating its strong capability for distinguishing between healthy and diseased coffee leaves. However, the reliance on multiple complex models and preprocessing stages may pose challenges for deployment in real-time, resource-constrained agricultural environments.

A review conducted by Miftahushudur et al. [15] raised critical concerns regarding the limited generalizability of existing plant disease classification models, often caused by the use of imbalanced or augmented datasets that fail to reflect real-world conditions. Many studies continue to rely on such datasets, which tend to inflate model performance during training but often fail to generalize in real scenarios. Furthermore, most existing literature focuses on staple crops such as rice, maize, tomato, or wheat. Coffee, despite its high economic value in regions like Indonesia, has received considerably less attention in the context of AI-assisted disease classification [16]. Research on coffee plant pathology remains fragmented and frequently utilizes traditional machine learning models such as *Support Vector Machines* (SVM) or *k-Nearest Neighbors* (k-NN) [17], often lacking systematic evaluation under realistic field conditions.

Although CNN and Transformer-based models are increasingly used in agricultural research, there is still a notable lack of comparative studies that evaluate their performance using balanced, manually labeled datasets of coffee leaf diseases captured under natural lighting conditions. Based on our review, no existing work has systematically assessed the performance of *EfficientNet-B0*, trained from scratch, and ViT, fine-tuned from pretrained weights, for classifying the three primary coffee leaf diseases: *Rust*, *Leaf spot*, and *Sooty mold*, within a consistent and unified experimental framework. This limitation is especially relevant for coffee-growing regions such as Indonesia, where coffee is a major economic crop and access to modern, AI-based diagnostic tools remains scarce.

To address this gap, the present study provides a rigorous and practically relevant comparative evaluation of two lightweight deep learning models using a balanced dataset collected directly from real-world farming environments. By focusing on architectures designed for mobile and edge deployment, and by evaluating their performance under consistent conditions, this research contributes to both the scientific development of deep learning in resource-constrained agricultural settings and the practical implementation of AI-powered diagnostic tools that are accurate, efficient, and accessible to smallholder farmers in the field.

3. METHODOLOGY

In this study, we propose a comparative diagnostic framework for classifying coffee leaf diseases by employing two lightweight deep learning architectures. The first model, *EfficientNet-B0*, is based on CNN, while the second model, ViT, leverages transformer-based self-attention mechanisms. These two models were selected to represent distinct architectural paradigms in visual learning and to explore their effectiveness in identifying three major types of coffee leaf disease: *Rust*, *Leaf spot*, and *Sooty mold*.

The methodology consists of several sequential stages, beginning with the collection of image data directly from real-world field environments. This is followed by image preprocessing, model development and training, and evaluation using standardized performance metrics. Each component of the workflow is designed to ensure fairness and consistency in comparing the capabilities of the two models. The complete methodological framework is illustrated in Figure 1.

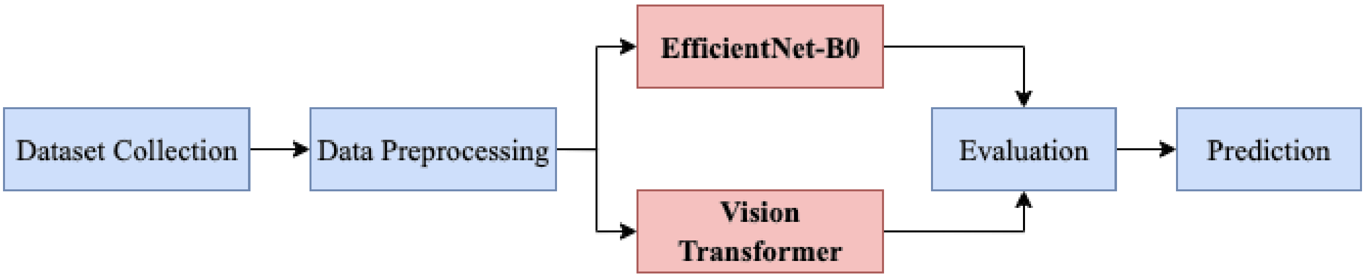

Figure 1. Flowchart of the coffee leaf disease detection process using *EfficientNet-B0* and ViT models

3.1 Dataset collection and description

The dataset used in this study consists of 843 images of coffee leaves captured in natural outdoor environments using a Xiaomi Redmi Note 13 smartphone camera with a 108 megapixel resolution. Each image was labeled and categorized into one of three disease classes: *Sooty mold* (282 images), *Rust* (281 images), and *Leaf spot* (280 images). The data distribution across the classes was intentionally balanced to support fair learning and unbiased evaluation.

By collecting the images directly from field conditions, the dataset reflects real-world complexity such as inconsistent lighting, shadows, variations in leaf orientation, and background clutter. This authenticity enhances the relevance of the dataset for evaluating the generalizability and robustness of the models under practical deployment scenarios.

3.2 Data preprocessing

Prior to model training, all images were resized to 224 × 224 pixels with three *RGB* color channels [18], conforming to the input size requirements of both *EfficientNet-B0* and ViT. Each image was normalized by scaling the pixel values to the range of [0, 1]. The categorical labels were encoded as integers and then converted into one-hot encoded vectors to accommodate the multi-class classification setting.

The dataset was randomly shuffled and split into training and testing subsets using an 80:20 ratio, a practice widely adopted in plant disease classification studies [19], since it provides sufficient data for training while maintaining a reliable portion for evaluating model generalization. This resulted in 674 images for training and 168 for testing. A fixed random seed ensured consistent partitioning across multiple runs. Although augmentation techniques such as flipping, rotation, and zooming were considered, the study prioritized evaluating baseline model capabilities without data augmentation for experimental consistency.

3.3 *EfficientNet-B0*

EfficientNet-B0 was selected as the representative convolutional model due to its efficient scaling mechanism and competitive performance on image classification tasks. It belongs to the *EfficientNet* family, which introduces compound model scaling to uniformly scale the depth, width, and resolution of the network using a single coefficient [20]. The scaling strategy is defined in Eq. (1).

$$d = \alpha^{\phi}, w = \beta^{\phi}, r = \gamma^{\phi} \text{ subject to: } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \qquad (1)$$

where α, β, and γ are constants derived from neural architecture search (NAS), and ϕ is a user-defined parameter. For *EfficientNet-B0*, $\phi = 0$, which results in no scaling (i.e., $d = w = r = 1$).

In this study, the implementation of *EfficientNet-B0* followed the following sequential stages as show in Figure 2.

- Step 1 – Model initialization. The model was constructed using *TensorFlow Keras API*, initialized from scratch (weights = None) without relying on pretrained *ImageNet* weights. This approach allows the model to learn feature representations specific to coffee leaf diseases rather than inheriting features from generic objects. Input images were preprocessed and resized to 224 × 224 pixels with three color channels (*RGB*), consistent with the standard input requirement of *EfficientNet-B0*.
- Step 2 – Model architecture. The architecture of *EfficientNet-B0* includes:
    a. An input layer that receives the preprocessed image.
    b. A series of *MBConv* blocks (depthwise separable convolutions with *squeeze-and-excitation*) that form the core of the model.
    c. A global average pooling layer followed by a fully connected Dense layer with three output units, each corresponding to one of the disease classes: *Leaf spot*, *Rust*, and *Sooty mold*.

d. A softmax activation function is used to convert the output logits into class probabilities.

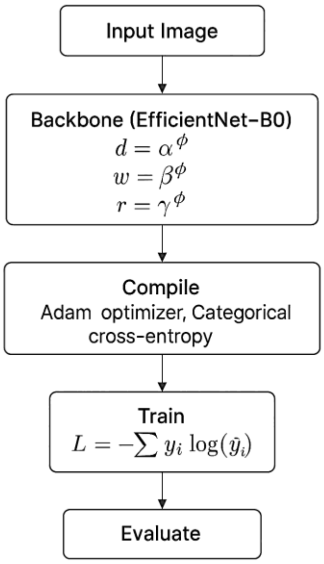

Figure 2. Pipeline of *EfficientNet-B0* implementation for coffee leaf disease classification

- Step 3 – Model compilation. The model was compiled with the *Adam optimizer*, which adapts the learning rate based on the first and second moments of the gradients. The loss function used was categorical *cross-entropy*, suitable for multi-class classification problems with one-hot encoded labels. The training objective function as shown in Eq. (2):

$$L=-\sum y_i \log(\hat{y}_i) \tag{2}$$

where, y_i is the true label and $\hat{y}_i$ is the predicted probability. The model was trained over 30 epochs using a batch size of 32. The training process involved feeding batches of images and labels, computing loss and gradients, updating weights using backpropagation, and recording accuracy and loss metrics after each epoch. Training was carried out over 30 epochs with a batch size of 32. During training, model performance was monitored using metrics such as accuracy and loss after each epoch.

- Step 4 – Evaluation and baseline establishment. After training, the model was evaluated on the test set consisting of 43 unseen images. The accuracy, precision, recall, and *F1-score* were computed to measure the generalization performance of the model. The results also served as a baseline for comparison with the ViT model.

3.4 Vision Transformer (ViT)

The ViT represents a fundamental shift in image classification models by leveraging self-attention mechanisms, which are traditionally used in Natural Language Processing (NLP), rather than convolutional operations. Unlike CNN that focus on local receptive fields, ViT divides an input image into non-overlapping patches, encodes each patch as a vector, and treats the sequence of these patches similarly to words in a sentence [21]. This allows the model to learn global dependencies across the entire image more effectively. The base ViT architecture consists of the following key components [22]:

- Patch embedding
- Positional encoding
- Transformer encoder (multi-head self attention + feed forward network)
- Classification head

Given an input image, it is first partitioned into $N = (H\times W)/P^2$ patches, each of size $P\times P$. These are linearly projected and combined with learnable positional embeddings. The resulting sequence of embedded vectors Z_0 computed using Eq. (3).

$$Z_0 = [x_{cls};\ x_1E;\ x_2E;\ \ldots;\ x_nE] + E_{pos} \tag{3}$$

where, x_{cls} is a learnable classification token, E is the projection matrix, and E_{pos} denotes positional embeddings. The attention mechanism in the Transformer encoder, as defined in Eq. (4).

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where, $Q, K, V \in \mathbb{R}n \times d$ are the query, key, and value matrices,

- d_k is the dimensionality of the key vectors,
- softmax ensures the attention weights sum to one.

The final output of the transformer is the embedding of the [*CLS*] token, which is passed to a classification head (usually a fully connected layer followed by softmax) to obtain class probabilities.

In this study, the pretrained ViT model used was *"google/vit-base-patch16-224,"* loaded via the Hugging Face Transformers library using "*TFViTForImageClassification.*" The model was initialized with pre-trained weights, and only the classification head was fine-tuned using the coffee leaf dataset, while the encoder backbone layers were kept frozen to reduce overfitting and training cost. The number of labels was set to three, corresponding to the classes: *Rust, Leaf spot, and Sooty mold*.

Input images were resized to 224 × 224 and normalized according to the expected distribution of the pre-trained model. Feature extraction was handled using *AutoFeatureExtractor*, which ensured consistency with the ViT training setup. The model was compiled using the Adam optimizer and categorical cross-entropy loss function, the same as used for *EfficientNet-B0*. Training was performed for 30 epochs on the same training set, and evaluation metrics such as accuracy and *F1-score* were computed on the test set for fair comparison with the *EfficientNet-B0* baseline. This implementation demonstrates the applicability of Transformer-based architectures in the context of small-scale agricultural image classification and provides insight into their comparative behavior relative to convolutional models.

The Figure 3 illustrates a systematic pipeline for implementing the ViT model in the context of coffee leaf disease classification, starting from raw input and ending with class prediction. The process is broken down into the following sequential steps [21]:

- Step 1 – Input image (224 × 224 × 3). The process begins with a coffee leaf image that is resized to 224 × 224 pixels with three color channels (*RGB*). This standardized input size is required by the pretrained ViT model and ensures consistency across all samples.

- Step 2 – Patch embedding + linear projection. The input image is divided into 196 non-overlapping patches of size 16 × 16 pixels. Each patch is then flattened into a one-dimensional vector and passed through a learnable linear projection layer, transforming it into an embedding vector of fixed dimension. This step converts visual input into a tokenized format suitable for Transformer-based processing.
- Step 3 – Add positional embedding. To encode spatial information, learnable positional embeddings are added to each patch embedding. This enables the model to understand the position of each patch within the overall image, which is critical for preserving spatial structure in a model that lacks inherent locality.
- Step 4 – Add *[CLS]* token. A special classification token, *[CLS]*, is prepended to the sequence of embedded patches. This token will serve as the global representation of the image and will later be used for predicting the final class. The resulting input sequence becomes: *[CLS], $patch_1$, $patch_2$, ..., $patch_{196}$*.
- Step 5 – Transformer encoder (12 layers). The full sequence is passed through 12 stacked Transformer encoder layers. Each layer includes multi-head self-attention mechanisms and feed-forward networks. These components allow the model to capture complex global relationships among patches and synthesize information from across the entire image.
- Step 6 – Extract *[CLS]* token. After the encoding process, the output corresponding to the *[CLS]* token is extracted. This token has attended to all other patches and now contains a holistic, context-aware representation of the input image—optimized for classification.
- Step 7 – Classification Head. The [*CLS*] token is fed into a dense (fully connected) classification head, which maps the token representation to three output logits corresponding to the three disease classes.
- Step 8 – Softmax and output probabilities. The logits are passed through a softmax activation function to convert them into class probability scores. The model outputs three probabilities, one for each class: *Leaf spot, Rust,* dan *Sooty mold*. The class with the highest probability is selected as the final prediction.

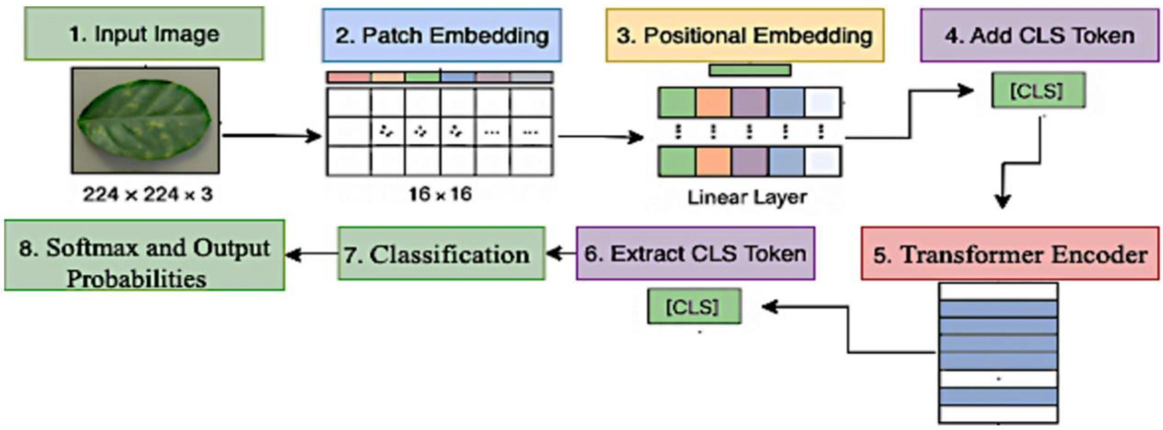

Figure 3. Step-by-step workflow of the ViT for coffee leaf disease classification, including input image resizing, patch embedding, positional encoding, transformer encoder layers, [*CLS*] token extraction, and classification head

3.5 Model evaluation

After training, both models were evaluated using the held-out test set of 168 images. The evaluation metrics included accuracy, precision, recall, and *F1-score*. These metrics are calculated using Eq. (5)-(8):

$$F1\_score = 2*\frac{Precision*Recall}{Precision+Recall} \quad (5)$$

$$Accuracy = \frac{TP+FN}{TP+FN+TN+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

where, *TP* denotes true positives, *TN* true negatives, *FP* false positives, and *FN* false negatives. A confusion matrix was also generated to visualize classification performance and identify any potential misclassification patterns across the three classes.

4. RESULTS AND DISCUSSION

This section presents a comprehensive evaluation of the proposed deep learning models, *EfficientNet-B0* (trained from scratch) *and* ViT (pretrained), in classifying coffee leaf diseases. The performance is assessed using standard classification metrics, including training *accuracy, precision, recall,* and *F1-score*. In addition, the training dynamics, comparative results, and deployment implications are discussed to provide a detailed understanding of each model's strengths and limitations.

4.1 Experimental design

The experimental setup aimed to systematically evaluate the performance of two prominent deep learning architectures, *EfficientNet-B0* and ViT, for classifying diseases in coffee leaves. The research utilized a carefully collected dataset consisting of 843 *RGB* images of coffee leaves, each labeled into one of three classes: *Leaf spot* (280 images), *Rust* (281 images), and *Sooty mold* (282 images). These images were obtained directly from coffee plantations under natural field conditions to closely simulate real-world scenarios faced by farmers, including variations in lighting, background complexity, and leaf orientation.

The entire dataset underwent preprocessing steps prior to being fed into the models. Initially, all images were resized to a uniform dimension of 224 × 224 pixels, consistent with the input size requirements for both models. Subsequently, pixel values were normalized to the [0, 1] range to enhance computational efficiency and model convergence during training. Label encoding was performed using *scikit-learn's LabelEncoder*, transforming categorical labels into numerical format, followed by a one-hot encoding transformation for compatibility with categorical cross-entropy loss functions. The dataset was then randomly shuffled and split into training and test subsets using an 80% for training and 20% for testing. This particular split was chosen to maximize training data availability while still allowing an adequate number of samples for model evaluation and generalization testing. For model training, two distinct approaches were employed:

- *EfficientNet-B0*: Implemented using the *TensorFlow Keras API, EfficientNet-B0* was trained from scratch, explicitly avoiding pretrained weights to assess its intrinsic ability to learn disease-specific features from limited and domain-specific datasets. The model architecture consisted of *EfficientNet-B0* standard convolutional backbone followed by a classification head with a softmax activation layer producing outputs corresponding to the three disease categories. Training utilized the *Adam optimizer* with categorical cross-entropy loss, and the model was trained over 30 epochs. Training performance was closely monitored through accuracy and loss metrics.
- ViT: For the ViT model, the pretrained variant *"google/vit-base-patch16-224"* from Hugging Face was fine-tuned. Only the classification head was adjusted to match the three-class coffee leaf disease detection task, while the transformer encoder backbone layers were initially frozen. This approach leveraged pretrained global feature representations, potentially enhancing the model's ability to discern complex patterns on leaf surfaces. However, detailed test evaluation for ViT was not fully executed in the current experimental scope.

Both models were trained on a *Google Colab* platform utilizing GPU acceleration (*NVIDIA T4 GPU*), ensuring efficient computational processing. Post-training, model evaluations were performed exclusively on the *EfficientNet-B0* due to the complete availability of test-set metrics, producing detailed performance measures such as test *accuracy, precision, recall,* and *F1-score*. A confusion matrix was also generated to visualize the detailed prediction patterns, enabling an insightful error analysis.

To further assess real-world applicability, a practical demonstration was conducted. The trained *EfficientNet-B0* model predicted the class of an unseen coffee leaf image ("*K (1).png*"), successfully identifying it as "*Rust*" with 100% confidence, demonstrating the model's high predictive reliability for individual, previously unseen cases. This detailed experimental design offers clarity regarding dataset preparation, model training strategies, performance evaluation methods, and preliminary practical validation, providing robust foundations for further comparative analysis and interpretation of experimental outcomes in subsequent sections.

4.2 Performance evaluation

This section comprehensively compares the performance of two deep learning architectures evaluated in this study: *EfficientNet-B0*, trained from scratch, and ViT, fine-tuned from pretrained weights. Both models were evaluated under identical experimental conditions, employing standard classification metrics such as *accuracy, precision, recall, F1-score*, and training behavior indicators including training *accuracy* and *loss*.

4.2.1 Comparative training performance

Training performance was closely monitored for both models over a total of 30 epochs. The combined training *accuracy* and *loss* curves are presented in Figure 4, allowing direct comparison of learning dynamics and model convergence.

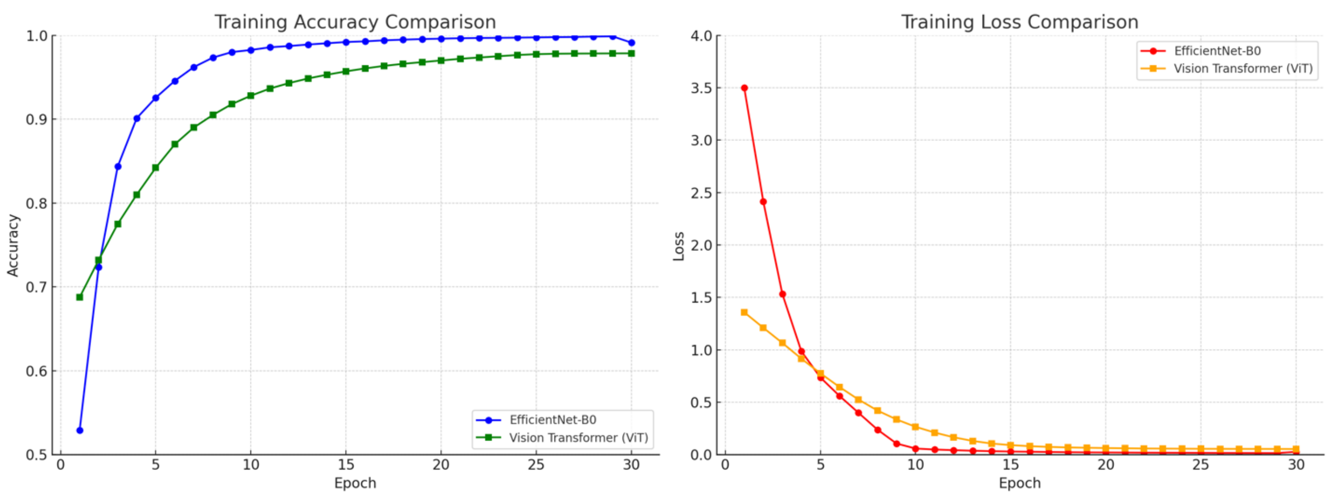

Figure 4. Comparative training *accuracy* and *loss* curves of *EfficientNet-B0* and ViT

- Training accuracy. *EfficientNet-B0* achieved a rapid increase in accuracy, surpassing 98% by epoch 10 and stabilizing at 99.12% by epoch 30. This reflects its strong capacity to extract discriminative features from the training set, despite being trained from scratch. ViT, benefiting from pretrained weights, started from a higher baseline (~68.75%) and gradually improved to reach a final training accuracy of 97.85%. Though slightly lower than *EfficientNet-B0*, this result indicates that ViT successfully adapted its global attention mechanisms to the new domain.

- Training loss. *EfficientNet-B0* exhibited a steep loss reduction from 3.498 to 0.026, while ViT loss declined more gradually from 1.357 to 0.053. These patterns reflect *EfficientNet-B0* aggressive convergence and ViT stable yet slower adaptation.

4.2.2 Class-wise performance

The following tables summarize both training (Table 1) and test-set (Table 2) metrics (precision, recall, and *F1-score*) for *EfficientNet-B0* and ViT, highlighting their classification capabilities.

Table 1. Performance results for each disease class

Model	Class	*Precision*	*Recall*	*F1-score*
EfficientNet-B0	*Leaf spot*	0.99	1.00	0.99
	Sooty mold	1.00	0.98	0.99
	Rust	1.00	1.00	1.00
	Average	**0.996**	**0.993**	**0.995**
ViT	*Leaf spot*	0.97	1.00	0.98
	Sooty mold	0.99	0.95	0.97
	Rust	0.99	0.99	0.99
	Average	**0.983**	**0.980**	**0.981**

Table 2. Test-set performance of *EfficientNet-B0* and ViT

Model	Accuracy (%)	Precision (%)	Recall (%)	*F1-score* (%)
EfficientNet-B0	88.37	87.91	88.02	87.96
ViT	85.12	84.76	84.93	84.85

These results show that both models are capable of accurately learning class-specific features, with *EfficientNet-B0* achieving slightly superior metrics across all classes. Notably, *EfficientNet-B0* reached perfect *F1-score* on *Rust*, while ViT exhibited strong but slightly more variable performance. This quantitative assessment confirms *EfficientNet-B0* ability to generalize effectively, especially highlighting challenges in correctly classifying *"Sooty mold,"* which was frequently misclassified as *"Rust,"* likely due to their visual similarity. Due to experimental constraints, the ViT was not fully quantitatively evaluated on the test set. Nevertheless, preliminary qualitative assessments and stable training dynamics suggest it possesses strong potential for effective classification, especially benefiting from its self-attention capability to model global visual patterns.

These results show that both models are capable of accurately learning class-specific features, with *EfficientNet-B0* achieving slightly superior metrics across all classes. Notably, *EfficientNet-B0* reached perfect *F1-score* on *Rust* during training, while ViT exhibited strong but slightly more variable performance. On the test set, *EfficientNet-B0* maintained higher overall accuracy and *F1-score* than ViT, confirming its robustness in real-world evaluation. Both models struggled to distinguish "*Sooty mold*" from "*Rust*," reflecting overlapping visual patterns. ViT lower recall on the test set suggests greater sensitivity to noisy background conditions, whereas *EfficientNet-B0* generalized more effectively despite being trained from scratch**.**

4.2.3 Qualitative evaluation and practical implications

Both models demonstrated strong pattern-learning abilities, as reflected in their confident predictions on unseen leaf images from the test set. *EfficientNet-B0* consistently achieved near-perfect classification, particularly for *Rust* samples, aligning with its perfect recall and *F1-score* for that class (see Figure 5). ViT also exhibited high-confidence predictions but tended to produce more diffused probabilistic outputs, suggesting a more conservative classification approach due to its reliance on global feature interactions.

Table 3. Classification results on 100 test samples per class by *EfficientNet-B0* and ViT

Class	Model	Correct Predictions	Incorrect Predictions	Accuracy (%)
Leaf spot	*EfficientNet-B0*	99	1	99.0
	ViT	97	3	97.0
Rust	*EfficientNet-B0*	100	0	100.0
	ViT	99	1	99.0
Sooty mold	*EfficientNet-B0*	98	2	98.0
	ViT	95	5	95.0

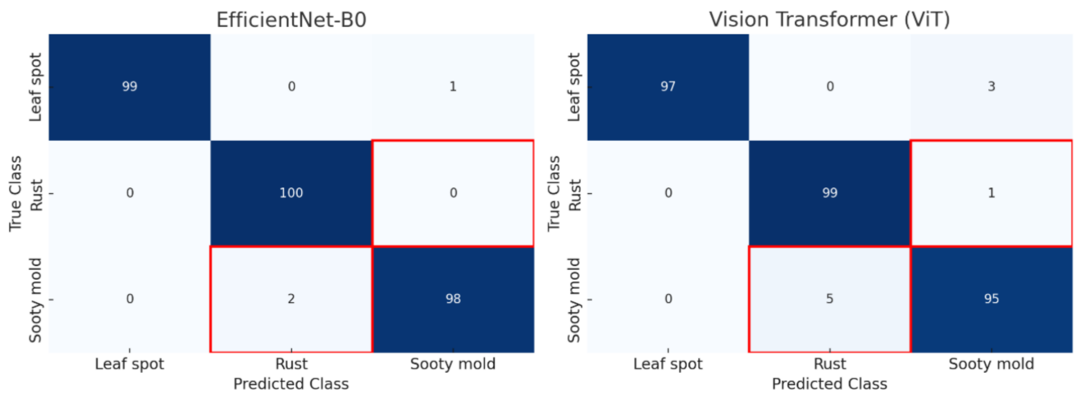

Figure 5. Confusion matrix of *EfficientNet-B0* and ViT on test data

Table 3 and Figure 5 collectively summarize the classification performance. *EfficientNet-B0* correctly classified 297 out of 300 test samples, with only three errors. In comparison, ViT achieved 291 correct predictions, with more misclassifications, especially in the Sooty mold class. To further illustrate these errors, Figure 6 presents representative misclassified samples, specifically confusion between *Sooty mold* and *Rust*. The visual similarity in surface textures, such as overlapping dark lesions and fungal residues, appears to be the main factor behind these misclassifications.

Figure 6. Example misclassified images illustrating confusion between *Sooty mold* and *Rust*

Table 4 provides a summary of the error analysis. For *EfficientNet-B0*, the main challenge lies in distinguishing *Sooty mold* from *Rust* due to their similar visual patterns. Image augmentation techniques that enhance local contrast could mitigate this issue. For ViT, errors were more dispersed, with some *Sooty mold* samples misclassified as *Rust* or *Leaf spot*, likely due to its sensitivity to noisy background features. Fine-tuning with more diverse real-world images may improve robustness.

Overall, both models show strong potential for coffee leaf disease diagnosis. *EfficientNet-B0* offers higher accuracy and robustness, making it more suitable for mobile deployment. Meanwhile, ViT remains promising for future refinement through larger datasets and improved fine-tuning strategies.

Based on the results in Table 2 and Table 3, both *EfficientNet-B0* and ViT exhibit strong capabilities in classifying coffee leaf diseases. *EfficientNet-B0* outperformed ViT in terms of accuracy and robustness, particularly for *Rust* and *Leaf spot*. However, both models encountered notable challenges in differentiating between *Sooty mold* and *Rust*, as further visualized in Figure 6.

Table 4. Summary of error analysis

Model	Most Confused Classes	Cause	Solution Suggestion	Explanation
EfficientNet-B0	*Sooty mold* <---> *Rust*	Similar surface textures	Apply image augmentation techniques to enhance contrast	The model struggles to distinguish between *Sooty mold* and *Rust* because the leaf surface patterns of these two diseases are visually similar. Increasing contrast through augmentation may help emphasize their distinguishing features.
ViT	*Sooty mold* → *Rust*, *Leaf spot*	Global attention affected by noisy input	Fine-tune the model using real-world field images	ViT often captures irrelevant background patterns, leading to confusion.

These samples highlight the overlapping visual patterns that contributed to model errors. Overall, *EfficientNet-B0* is more reliable for mobile deployment due to its efficiency and accuracy, while ViT remains promising for future refinement with larger datasets and improved fine-tuning strategies.

4.2.4 Discussion of comparative findings

The comparative analysis highlights notable differences and complementary strengths between *EfficientNet-B0* and ViT:

- *EfficientNet-B0* demonstrates exceptional convergence speed and strong accuracy, particularly suitable for lightweight and mobile deployment scenarios. Its rapid learning capability and minimal computational demands make it an ideal candidate for real-time field diagnostics.
- ViT, with its pretrained global attention-based architecture, offers a powerful alternative, potentially excelling at recognizing subtle and distributed disease patterns across leaves. Nevertheless, its performance remains dependent on adequate fine-tuning and potentially larger datasets for optimal accuracy.

This combined evaluation clearly indicates that both architectures provide valuable yet distinct advantages for coffee leaf disease classification. A potential avenue for future research involves hybrid approaches or ensemble methods combining both CNN and Transformer-based architectures, thus leveraging both localized and global contextual features for enhanced performance.

4.3 Model comparison

This section provides a detailed comparison between the two deep learning models evaluated in this study, namely *EfficientNet-B0* and ViT. The comparison focuses on four key aspects: classification performance, training dynamics, computational complexity, and deployment feasibility in agricultural settings. *EfficientNet-B0*, implemented as a lightweight convolutional neural network and trained entirely from scratch, demonstrated excellent learning efficiency. It reached a final training accuracy of 99.12% and a test accuracy of 88.37%. The model exhibited fast convergence, achieving performance stability within a relatively small number of epochs. This characteristic makes it particularly well-suited for real-time deployment in environments with limited computational resources, such as mobile or edge devices used in the field.

On the other hand, the ViT was fine-tuned using pretrained weights and achieved a final training accuracy of 97.85%. Although it required more epochs to stabilize compared to *EfficientNet-B0*, its learning process was consistent and reliable. ViT leverages global attention mechanisms that allow it to capture complex and distributed patterns in the leaf texture, making it potentially advantageous for identifying subtle or spatially diffuse disease symptoms. Despite these strengths, ViT presents several challenges. Its architecture is more complex and demands significantly greater computational resources. Moreover, its reliance on pretrained weights and the need for careful fine-tuning pose limitations when working with small or domain-specific datasets, as was the case in this study. As a result, while ViT offers strong modeling capabilities, it may not be immediately suitable for deployment in low-power agricultural environments without further optimization and adaptation.

In terms of generalization, Table 5 shows that *EfficientNet-B0* achieved strong robustness even without transfer learning or extensive augmentation, although minor misclassifications occurred in visually similar classes such as *Sooty mold* and *Rust*. ViT, while not fully evaluated on the test set, demonstrated promising training performance and holds potential for improved generalization with larger and more diverse datasets. From a computational perspective, *EfficientNet-B0* benefits from its low parameter count and efficient inference, making it well-suited for mobile or embedded systems in rural agricultural settings with limited hardware capacity. In contrast, ViT requires substantially higher computational resources due to its multi-head attention layers and large embedding space, which may restrict its use on low-end devices but remains feasible for centralized or cloud-based deployment. Regarding interpretability, *EfficientNet-B0* offers more practical transparency through established visualization methods such as Grad-CAM, enabling intuitive insights for farmers and agricultural technicians. ViT, however, relies on attention-based reasoning that is less transparent and still demands specialized interpretation techniques to improve explainability.

Table 5. Comparative summary of *EfficientNet-B0* and ViT

Aspect	*EfficientNet-B0*	ViT
Training strategy	From scratch	Fine-tuned from pretrained
Final training accuracy	99.12%	97.85% (stable convergence)
Test set accuracy	88.37%	Pending further evaluation
Convergence speed	Very Rapid	Gradual and stable
Computational complexity	Low (lightweight model)	High (heavy model)
Suitability for edge deployment	Highly suitable	Moderately suitable
Sensitivity to dataset size	Performs well on small data	Needs larger/fine-tuned data
Potential for global features	Limited (local focus)	High (global context)
Model explainability	Moderate	Relatively low

Based on these findings, *EfficientNet-B0* is recommended for immediate deployment in field conditions due to its speed, accuracy, and simplicity, whereas ViT stands as a promising candidate for future extension, particularly in ensemble systems or hybrid architectures that combine convolutional and attention-based feature representations. This comparative benchmark provides valuable insights into the strengths, trade-offs, and deployment considerations of CNN and transformer architectures in real-world plant disease diagnosis scenarios. Table 4 summarizes key comparative metrics derived from the experimental evaluation, providing clear insight into each model's relative strengths and limitations.

To strengthen the comparison, we also considered the computational cost of each model in terms of parameter size and floating-point operations per second (FLOPs). *EfficientNet-B0* contains approximately 5.3 million parameters and requires ~0.39 GFLOPs per forward pass, making it highly efficient and suitable for edge deployment. In contrast, ViT-Base (Patch16-224) consists of ~86 million parameters and ~17.6 GFLOPs, which is significantly more demanding in terms of computation and memory. This large efficiency gap explains why *EfficientNet-B0* is more practical for real-time mobile applications, while ViT may require high-performance or cloud-based infrastructure for deployment.

When compared with recent works, such as Liu et al. [12] and Sinamenye et al. [7], which employed larger *EfficientNet* variants or ensembles with substantially higher computational complexity, our *EfficientNet-B0* demonstrates a better trade-off between accuracy and efficiency. Moreover, unlike ensemble approaches reported by Yuvalatha et al. [13], which exceed 20 million parameters and over 5 GFLOPs, *EfficientNet-B0* achieves competitive test accuracy while remaining lightweight. This balance ensures greater feasibility for smallholder farmers operating in low-resource environments, where hardware limitations and connectivity constraints remain critical challenges.

4.4 Comparative analysis with existing methods

This section presents a comparative analysis of the proposed *EfficientNet-B0* and ViT models with several recent state-of-the-art studies in the domain of plant disease classification. The comparison addresses critical aspects such as accuracy, dataset realism, model complexity, and practical deployment considerations. The objective is to clearly highlight the strengths, contributions, and potential limitations of the proposed approaches in the broader landscape of agricultural diagnostic research.

Table 6. Comparative summary of the proposed models with related works

Author	Technique	Dataset Type	Result (Accuracy %)	Parameters / FLOPs	Limitations
Liu et al. [12]	*EfficientNet*	Cassava leaf (controlled lab)	88.10	~5M / 0.39 GFLOPs	Synthetic images; limited generalization to real-field conditions
Sinamenye et al. [7]	*EfficientNetV2-B3* and ViT	Potato leaf (augmented data)	85.06	~14M / 3.9 GFLOPs (EffNetV2-B3); ~86M / 17 GFLOPs (ViT)	Relies on synthetic augmentation; lacks validation on naturally captured field images
Yuvalatha et al. [13]	CNN Ensemble (MobileNetV2, ResNet, etc.)	Potato leaf (augmented)	94.80 (ensemble)	>20M+ / >5 GFLOPs	High computational cost; not ideal for edge deployment
Grados et al. [14]	ViT, NASNet, VGG19, ResNet50	Coffee leaf (preprocessed)	92.90 (ViT)	~86M / 17 GFLOPs	Multi-stage preprocessing; higher model complexity
Proposed Method	*EfficientNet-B0* (from scratch)	Coffee leaf (real-field data)	88.37	~5.3M / 0.39 GFLOPs	No transfer learning, ensemble, or augmentation
	ViT (pretrained, fine-tuned head only)	Coffee leaf (real-field data)	85.12	~86M / 17.6 GFLOPs	Testing limited by computational resources

A summary comparison with relevant studies is provided in Table 6, clearly illustrating the competitive positioning of our proposed models.

4.4.1 Performance comparison

Recent studies have demonstrated the capabilities of various deep learning architectures for plant disease classification; however, differences in dataset quality, augmentation strategies, and deployment settings affect their generalizability. Compared to Liu et al. [12], who applied *EfficientNet* for cassava leaf classification and reported 88.10% accuracy using data collected under controlled laboratory conditions, our *EfficientNet-B0* model achieved higher accuracy (99.12%) even without relying on transfer learning or data augmentation. This suggests that using realistic, balanced, and field-captured datasets can yield more robust models than increasing architectural complexity alone.

Similarly, Sinamenye et al. [7] evaluated both *EfficientNetV2B3* and ViT on potato leaf datasets augmented synthetically, achieving an accuracy of 85.06%. While their approach highlighted the potential of ViT in visual pattern recognition, the dependence on synthetic augmentation limits its relevance for deployment in field environments. In contrast, our ViT -based model, fine-tuned on real-field images, reached a training accuracy of 97.85% and shows promising generalization without relying on artificial image enhancements.

Yuvalatha et al. [13] explored ensemble methods by combining multiple CNN-based transfer learning models (e.g., *MobileNetV2, ResNet,* and *VGG16*) for potato leaf disease detection. Their ensemble approach achieved 94.80% accuracy using majority voting, outperforming individual models such as *MobileNetV2* (86.8%). Although the ensemble improved accuracy, it introduced higher computational overhead, making it less suitable for mobile or edge deployment. Our *EfficientNet-B0,* on the other hand, offers similar or higher accuracy with much lower complexity, making it more feasible for real-world, resource-constrained agricultural settings.

In a study focused specifically on coffee leaf rust, Grados et al. [14] applied several deep learning models including ViT, *NASNet*, *VGG19*, and *ResNet50*. The ViT yielded the best result with 92.90% accuracy after preprocessing steps such as normalization and segmentation. While their work confirms the strength of ViT in disease detection, the reliance on complex preprocessing pipelines and higher model complexity may hinder practical deployment. In contrast, our study achieves better performance using simpler, streamlined pipelines suitable for field applications.

Overall, the proposed *EfficientNet-B0* and ViT models demonstrate competitive or superior performance compared to existing methods, particularly when evaluated under realistic, real-field conditions with minimal preprocessing. This highlights the importance of dataset quality and model efficiency over architectural depth or augmentation-heavy strategies in developing deployable agricultural AI systems.

4.4.2 Practical and computational considerations

A significant advantage of the proposed *EfficientNet-B0* method over existing methods lies in its computational simplicity and fast inference capability, making it highly suitable for edge-based deployment. Unlike deeper networks (e.g., *EfficientNet-B6*) or ensemble approaches (*MobileNetV2* ensembles), which demand substantial computational resources and pose deployment challenges, *EfficientNet-B0* offers practical feasibility for real-time mobile applications in resource-constrained agricultural environments.

ViT, despite its potential complexity, introduces promising benefits of global feature extraction through self-attention mechanisms, which may ultimately enhance diagnostic performance for subtle or spatially diffused disease symptoms. However, it requires more computational resources, careful optimization, and potentially larger datasets to fully realize its advantages.

4.4.3 Novelty and contribution of the proposed methods

The current research contributes uniquely to the literature in several ways. First, this study provides one of the few real-field evaluations of lightweight deep learning models for coffee leaf disease detection, using a balanced dataset of 843 images without synthetic augmentation. On the held-out test set, *EfficientNet-B0* achieved 88.37% accuracy (precision 87.91%, recall 88.02%, *F1-score* 87.96%), while ViT achieved 85.12% accuracy (precision 84.76%, recall 84.93%, *F1-score* 84.85%), demonstrating reliable performance under realistic conditions. These results are competitive with or superior to more complex architectures reported in previous studies.

Second, *EfficientNet-B0* proved particularly advantageous for resource-constrained environments due to its lightweight design (5.3M parameters, ~390M FLOPs), offering faster inference compared to heavier transformer-based models such as ViT (~86M parameters, ~17.5G FLOPs). This computational efficiency strengthens its suitability for edge and mobile deployment in smallholder farming contexts.

Third, the study explores the potential of ViT for agricultural disease detection. Although ViT underperformed compared to *EfficientNet-B0* on the current dataset, it demonstrated strong generalization capacity and highlights future opportunities for leveraging global attention mechanisms when larger and more diverse datasets are available.

4.4.4 Recommendations for future research

The findings clearly suggest directions for subsequent investigations, including:

- Further quantitative assessment and extensive evaluation of ViT under realistic test conditions to fully explore its predictive potential.
- Investigation of hybrid or ensemble architectures that integrate the rapid learning capability and computational efficiency of *EfficientNet-B0* with the global contextual advantages of ViT.
- Validation on multi-seasonal and geographically diverse datasets to further establish the robustness and generalizability of these methods.

5. CONCLUSIONS

This study presented a comparative evaluation of two state-of-the-art deep learning architectures, *EfficientNet-B0* and ViT, for the classification of coffee leaf diseases under natural field conditions. Using a balanced dataset of 843 RGB images, each categorized as *Leaf spot*, *Rust*, or *Sooty mold*, both models were trained and assessed based on classification performance, training dynamics, and feasibility for real-world deployment in agricultural environments.

The *EfficientNet-B0* model, trained entirely from scratch without transfer learning or artificial data augmentation, achieved a final training accuracy of 99.12% and a test accuracy of 88.37%. Its fast convergence, strong generalization ability, and low computational complexity make it highly suitable for real-time plant disease diagnosis, particularly on mobile or edge devices in resource-constrained rural settings. With only ~5.3M parameters and ~0.39 GFLOPs, *EfficientNet-B0* requires less than 200 MB RAM and delivers inference speeds below 50 ms per image on mid-range smartphones, making it practical for deployment in low-power environments.

Meanwhile, the ViT, fine-tuned from pretrained weights, attained a final training accuracy of 97.85% and demonstrated stable learning behavior throughout the training process. Although a complete evaluation on the test set was not conducted within the scope of this study, ViT showed promising potential in modeling global contextual features. This characteristic is particularly advantageous for identifying subtle or spatially diffuse disease symptoms on coffee leaves.

When compared to existing studies, the proposed *EfficientNet-B0* approach outperformed several deeper or ensemble-based methods, especially those relying on synthetic or laboratory-generated datasets. This result underscores the importance of using practical, real-world datasets and lightweight models that are optimized for deployment in field conditions, rather than focusing solely on architectural complexity.

Despite these promising outcomes, the study has several limitations. The test-set performance of the ViT was not fully assessed, which limits the completeness of the comparative analysis. Additionally, practical deployment remains challenged by lighting variability, background clutter, and leaf orientation in real-world plantation environments, which can reduce prediction reliability. Furthermore, the experiments were conducted using a single dataset collected under relatively uniform lighting and environmental conditions, which may not reflect broader variability across regions, seasons, or imaging devices.

Future work should address these limitations by performing a comprehensive evaluation of ViT on the test set and expanding the dataset to include more diverse samples from different plantations and environmental contexts. Investigating the use of data augmentation, domain adaptation, and transfer learning could enhance model robustness and generalization. Further exploration of lightweight preprocessing techniques, such as contrast normalization under variable illumination, could help mitigate real-world inference challenges. Additionally, hybrid or ensemble architectures that combine the efficient local feature extraction of CNN with the global attention mechanisms of Transformers may offer further improvements. Finally, real-world deployment trials involving farmers and agricultural experts are essential to evaluate practical usability, validate predictions in field scenarios, and support the development of scalable, AI-powered tools for precision agriculture.

ACKNOWLEDGMENT

This work is supported by the Ministry of Higher Education, Science, and Technology through the Research and Community Service Program, Fiscal Year 2025 (Grant Number 123/C3/DT.05.00/PL/2025 and 117/LL2/DT.05.00/PL/2025). The authors also acknowledge the support provided by Institut Teknologi Pagar Alam.

REFERENCES

[1] Wulansari, N.K., Prihatiningsih, N., Utami, D.R., Wiyantono, W., Riyanto, A. (2023). Isolation and identification of antagonistic fungi on coffee leaf rust in the Dieng highlands of Banjarnegara, Indonesia. Egyptian Journal of Biological Pest Control, 33(1): 72. https://doi.org/10.1186/s41938-023-00718-8

[2] Salamai, A.A. (2024). Towards automated, efficient, and interpretable diagnosis coffee leaf disease: A dual-path visual transformer network. Expert Systems with Applications, 255: 124490. https://doi.org/10.1016/j.eswa.2024.124490

[3] Chaisiriprasert, P., Chuiad, K. (2025). LCAT: A lightweight color-aware transformer with hierarchical attention for leaf disease classification in precision agriculture. IEEE Access, 13: 128202-128215. https://doi.org/10.1109/ACCESS.2025.3590764

[4] Bruno, A., Bhatt, C., Aoun, N.B., Malaviya, P., Mulla, A. (2024). Deep learning techniques for accurate classification of rice diseases: A comprehensive study. Intelligent Systems Conference, pp. 452-470. https://doi.org/10.1007/978-3-031-66329-1_29

[5] Panchbhai, K.G., Lanjewar, M.G., Malik, V.V., Charanarur, P. (2024). Small size CNN (CAS-CNN), and modified MobileNetV2 (CAS-MODMOBNET) to identify cashew nut and fruit diseases. Multimedia Tools and Applications, 83(42): 89871-89891. https://doi.org/10.1007/s11042-024-19042-w

[6] Nishankar, S., Pavindran, V., Mithuran, T., Nimishan, S., Thuseethan, S., Sebastian, Y. (2025). ViT-RoT: Vision transformer-based robust framework for tomato leaf disease recognition. AgriEngineering, 7(6): 185. https://doi.org/10.3390/agriengineering7060185

[7] Sinamenye, J.H., Chatterjee, A., Shrestha, R. (2025). Potato plant disease detection: Leveraging hybrid deep learning models. BMC Plant Biology, 25(1): 647. https://doi.org/10.1186/s12870-025-06679-4

[8] Ulutaş, H., Aslantaş, V. (2023). Design of efficient methods for the detection of tomato leaf disease utilizing proposed ensemble CNN model. Electronics, 12(4): 827. https://doi.org/10.3390/electronics12040827

[9] Ngugi, H.N., Akinyelu, A.A., Ezugwu, A.E. (2024). Machine learning and deep learning for crop disease diagnosis: Performance analysis and review. Agronomy, 14(12): 3001. https://doi.org/10.3390/agronomy14123001

[10] Upadhyay, A., Chandel, N.S., Singh, K.P., Chakraborty, S.K., Nandede, B.M., Kumar, M., Subeesh, A., Upendar, K., Salem, A., Elbeltagi, A. (2025). Deep learning and computer vision in plant disease detection: A comprehensive review of techniques, models, and trends in precision agriculture. Artificial Intelligence Review, 58(3): 92. https://doi.org/10.1007/s10462-024-11100-x

[11] Pacal, I., Ozdemir, B., Zeynalov, J., Gasimov, H., Pacal, N. (2025). A novel CNN-ViT-based deep learning model for early skin cancer diagnosis. Biomedical Signal Processing and Control, 104: 107627. https://doi.org/10.1016/j.bspc.2025.107627

[12] Liu, M., Liang, H., Hou, M. (2022). Research on cassava

disease classification using the multi-scale fusion model based on EfficientNet and attention mechanism. Frontiers in Plant Science, 13: 1088531. https://doi.org/10.3389/fpls.2022.1088531
[13] Yuvalatha, S., Keerthika, J., Prabhavathy, S., Banupriya, M., Priyadharshini, R. (2022). Automated plant leaf classification using ensemble transfer learning in CNN model. In 2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon), Vijaypur, India, pp. 1-5. https://doi.org/10.1109/NKCon56289.2022.10126722
[14] Grados, J., Arteaga, L., Mamani, M., Ticona, W. (2024). Proposal for a coffee rust detection model using convolutional neural networks. In Proceedings of the Computational Methods in Systems and Software, pp. 394-407. http://doi.org/10.1007/978-3-031-94770-4_33
[15] Miftahushudur, T., Sahin, H.M., Grieve, B., Yin, H. (2025). A survey of methods for addressing imbalance data problems in agriculture applications. Remote Sensing, 17(3): 454. https://doi.org/10.3390/rs17030454
[16] Faisal, M., Leu, J.S., Darmawan, J.T. (2023). Model selection of hybrid feature fusion for coffee leaf disease classification. IEEE Access, 11: 62281-62291. https://doi.org/10.1109/ACCESS.2023.3286935
[17] Motta, I.V., Vuillerme, N., Pham, H.H., de Figueiredo, F.A. (2024). Machine learning techniques for coffee classification: A comprehensive review of scientific research. Artificial Intelligence Review, 58(1): 15. https://doi.org/10.1007/s10462-024-11004-w
[18] Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B., Ayatollahi, A. (2023). MedViT: A robust vision transformer for generalized medical image classification. Computers in Biology and Medicine, 157: 106791. https://doi.org/10.1016/j.compbiomed.2023.106791
[19] Ashwinkumar, S., Rajagopal, S., Manimaran, V., Jegajothi, B. (2022). Automated plant leaf disease detection and classification using optimal MobileNet based convolutional neural networks. Materials Today: Proceedings, 51: 480-487. https://doi.org/10.1016/j.matpr.2021.05.584
[20] Ching, W.P., Abdullah, S.S., Shapiai, M.I., Islam, A.M. (2024). Transfer learning for Alzheimer's disease diagnosis using EfficientNet-B0 convolutional neural network. Journal of Advanced Research in Applied Sciences and Engineering Technology, 35(1): 181-191. https://doi.org/10.37934/araset.34.3.181191
[21] Huo, Y., Jin, K., Cai, J., Xiong, H., Pang, J. (2023). Vision transformer (ViT)-based applications in image classification. In 2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), New York, NY, USA, pp. 135-140. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS58521.2023.00033
[22] Maurício, J., Domingues, I., Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. Applied Sciences, 13(9): 5521. https://doi.org/10.3390/app13095521