JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Enhanced Vision Transformer and Transfer Learning Approach to Improve Rice Disease Recognition

Rahadian Kristiyanto Rachman[1], De Rosal Ignatius Moses Setiadi[1,*], Ajib Susanto[1], Kristiawan Nugroho[2], and Hussain Md Mehedul Islam[3]

[1] Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia; e-mail : rahadiankristiyanto03@gmail.com; moses@dsn.dinus.ac.id; ajib.susanto@dsn.dinus.ac.id
[2] Department of Information Technology and Industry, StikubankUniversity, Semarang, Indonesia; e-mail : kristiawan@edu.unisbank.ac.id
[3] Software Engineer, The Mathworks, Inc., United States; e-mail: mehadi.cuet@gmail.com
* Corresponding Author : De Rosal Ignatius Moses Setiadi

Abstract: In the evolving landscape of agricultural technology, recognizing rice diseases through computational models is a critical challenge, predominantly addressed through Convolutional Neural Networks (CNN). However, the localized feature extraction of CNNs often falls short in complex scenarios, necessitating a shift towards models capable of global contextual understanding. Enter the Vision Transformer (ViT), a paradigm-shifting deep learning model that leverages a self-attention mechanism to transcend the limitations of CNNs by capturing image features in a comprehensive global context. This research embarks on an ambitious journey to refine and adapt the ViT Base(B) transfer learning model for the nuanced task of rice disease recognition. Through meticulous reconfiguration, layer augmentation, and hyperparameter tuning, the study tests the model's prowess across both balanced and imbalanced datasets, revealing its remarkable ability to outperform traditional CNN models, including VGG, MobileNet, and EfficientNet. The proposed ViT model not only achieved superior recall (0.9792), precision (0.9815), specificity (0.9938), f1-score (0.9791), and accuracy (0.9792) on challenging datasets but also established a new benchmark in rice disease recognition, underscoring its potential as a transformative tool in the agricultural domain. This work not only showcases the ViT model's superior performance and stability across diverse tasks and datasets but also illuminates its potential to revolutionize rice disease recognition, setting the stage for future explorations in agricultural AI applications.

Keywords: Multi-head Attention; Paddy Disease Classification; Rice Leaves Disease Recognition; Self-Attention; Transfer Learning; Vision Transformer.

Received: April, 1st 2024
Revised: April, 22nd 2024
Accepted: April, 25th 2024
Published: April, 26th 2024

1. Introduction

Rice is a crucial food source for nearly half the global population, fulfilling over 21% of the worldwide calorie intake, crucial for food security and societal well-being[1]. The year 2021 saw global rice production soar to 787 million tons, highlighting the urgent need for a swift shift to sustainable agricultural and food systems to safeguard global food security, especially concerning rice. However, efforts to maintain and increase rice production cannot be separated from the threat of disease [2]–[5].

Computer vision has experienced significant development within the agricultural sector, with numerous studies focusing on plant disease recognition through machine learning (ML) approaches. Various methods, including logistic regression (LR), support vector machine (SVM), k-nearest neighbors (KNN), and Random Forest (RF), have been employed[6]–[8]. These ML methods usually require manual feature extraction and selection for image recognition tasks, which may prove suboptimal for analyzing complex images. They struggle with high-dimensional data and are not inherently designed for capturing spatial hierarchies in images. ML methods also tend to perform relatively slowly when processing large datasets and struggle with imbalanced datasets[9]. In contrast, deep learning (DL) approaches can

automatically learn hierarchical features from raw images. They offers superior performance in recognizing patterns and objects in images due to their deep architecture that can capture intricate details and spatial relationships[10], [11]. One widely used DL approach is the Convolutional Neural Network (CNN), which excel at image recognition by learning feature hierarchies directly from images and automatically identifying important features through specialized layers. This architecture mimics biological visual processing, making CNNs more effective for spatial data tasks, offering superior accuracy and efficiency compared to traditional ML methods[2], [10], [12], [13]. However, CNN has disadvantages, especially in the training process, which can consume a lot of resources and require large amounts of computing power and data, thus necessitating a longer training period than traditional ML methods. In contrast, transfer learning leverages models pre-trained on large data sets, thereby significantly reducing computational resource and data requirements. It accelerates training and enhances performance on tasks with limited data, making it a more efficient approach in certain scenarios[13], [14]. Significant advancements have been made in the development of CNN-based transfer learning models for image recognition tasks. Notable examples include VGG[15], MobileNet [16], EfficientNet [17], [18], Xception [19], Residual Network (ResNet) [20], dan AlexNet[21]. However, CNN methods or CNN-based transfer learning still face limitations due to their reliance on local convolution operations for feature detection, where each pixel interacts only with its immediate neighbors in a small environment.

The Vision Transformer (ViT) model emerges as an alternative to overcome CNN's shortcomings by modeling long-range dependencies between image parts[22]–[24]. ViTs employ self-attention mechanisms to capture global context, thereby improving performance in complex visual tasks [25]–[27]. The ViT architecture, adapted from a scalable transformation mechanism, can be adjusted for various image sizes and types, offering new opportunities for detailed image analysis and object detection [28]. Several studies have tested ViT for image classification tasks, such as in research [29] obtained a result of 98.49%, and research [30] produced an accuracy value of up to 99%, wherein both studies, it was also compared that ViT showed better performance than prior-art. However, further exploration is required to ascertain ViT's specific advantages in disease classification tasks in rice plants.

Based on the existing literature, there is a significant potential for the Vision Transformer (ViT) model to outperform Convolutional Neural Networks (CNNs) in image recognition tasks. This research aims to delve deeper into the application of ViT for classifying rice diseases. Our contributions include:

1. Tailoring the ViT model for the specific task of rice disease recognition.
2. Evaluating the ViT model's performance on both balanced and imbalanced datasets.
3. Comparing ViT and CNN models to assess their effectiveness in rice plant disease classification, with a focus on ViT's global pattern recognition capabilities.

The remainder of this paper will delve into the development and evaluation of the Vision Transformer (ViT) model for disease recognition in rice plants. Section 2 will cover a comprehensive review of ViT literature. Section 3 will discuss our research methodology, encompassing data collection, the model training process, and the metrics we used for evaluation. Following this, Section 4 will present our experimental findings and provide a detailed analysis of these results and their broader significance. The paper will conclude with a summary of our key findings and offer recommendations for future investigations in the realm of plant disease detection through image processing technologies.

2. Preliminaries

In the evolving landscape of machine learning and its application to agricultural technology, the Vision Transformer (ViT) model emerges as a groundbreaking approach, offering a novel perspective on image-based classification tasks. Unlike conventional convolutional neural networks (CNNs) that process images through localized convolution operations, ViT leverages the power of the Transformer architecture, originally designed for natural language processing, to interpret visual data. This section delves into the foundational elements of ViT, including its innovative use of self-attention mechanisms to analyze images as sequences of patches, thereby capturing intricate patterns and relationships within the data. By dissecting the mechanics of ViT and its components, such as self-attention, we aim to illuminate its potential to surpass traditional methods in accurately classifying rice diseases. Through a comprehensive review of related literature and theoretical frameworks, we will explore the current

state of ViT development, identify gaps in existing research, and highlight the unique contributions of our study in optimizing and applying ViT for enhanced disease recognition in rice plants.

2.1. Vision Transformer (ViT)

Vision Transformer (ViT) represents a neural network model architecture that adopts a Transformer approach for processing visual information, especially images. In contrast to traditional methods, which often use convolutional layers for image processing, ViT works by changing the input images into small patches, where each patch will be represented in vector form[31]. These patch vectors are then flattened into a one-dimensional sequence, serving as the input for the Transformer model.Each patch vector is infused with positional information (position embedding) to preserve the spatial context of the image. Subsequently, this sequence is processed by the Transformer encoder, which comprises several self-attention layers and a Multi-Layer Perceptron (MLP). This configuration enables the model to discern the intricate relationships among different parts of the image. The output from the Transformer encoder is then directed to the classification layer, which generates class predictions[25]. The architecture of Vision Transformer is illustrated in Figure 1.

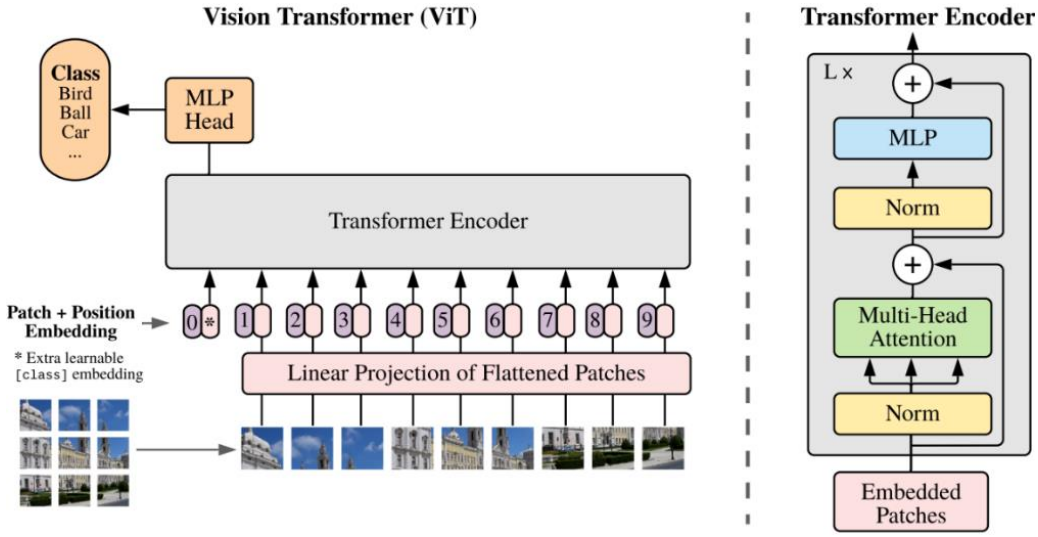

Figure 1. Vision Transformer (ViT) Architecture[25]

2.2 Self-Attention

The self-attention mechanism is a key component of the Transformer architecture in artificial neural networks, enables the model to focus on and evaluate the significance of various relationships between elements within a data sequence, such as words in text or patches in an image [28]. Figure 2 visualizes the self-attention mechanism, illustrating how the model assesses the importance of each element within a data sequence in relation to others, thereby focusing on the weighted significance of these elements.

Following are the basic steps of the self-attention mechanism:

1. **Image Division**: Images are segmented into small chunks, analogous to "words" or tokens in natural language models. Each patch in the image is considered an entity treated by self-attention.
2. **Query, Key, and Value Representation**: Each image patch has a Query, Key, and Value representation. This is obtained through a linear transformation of each image patch.
3. **Conformity Score Calculation** (Dot Product): Each patch is used to query the conformity score (dot product) with every other patch in the key image. This score indicates how relevant the patch-query is to the patch-key.
4. **Normalization and Weighting:** The suitability scores are normalized using a softmax function, generating weights that emphasize the most relevant patches in relation to the query patch. These weights are then used to multiply by the values of the image patches, producing a new weighted representation for each patch based on its relationship to other patches in the image.

5. **Integration and Output:** The weighted representations of each patch are combined to produce an output of self-attention at the level of the whole image. This output is then used for subsequent steps in the ViT model.

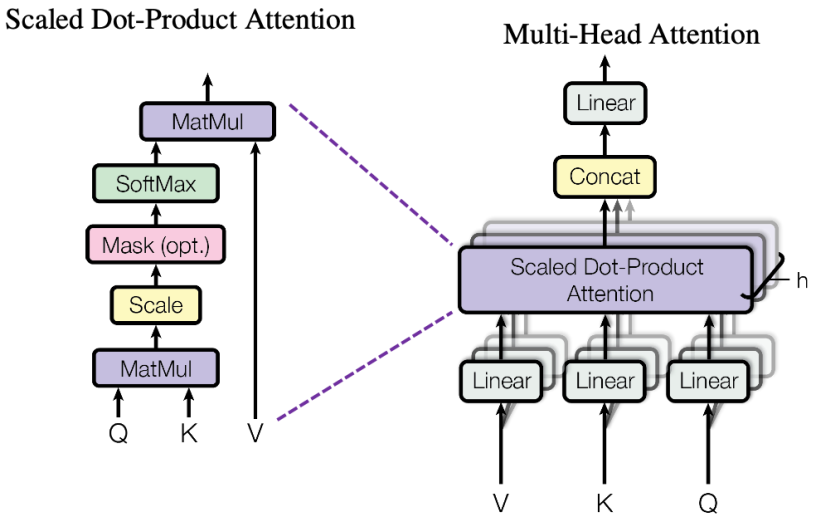

Figure 2. Self-Attention Mechanism[28]

Through this self-attention mechanism, the model can understand the relationship between each element in the data sequence and determine its importance in the context of its relationship to other elements. This allows the model to focus on the most important parts of the input data sequence. This arrangement evolves into multi-head attention, comprising a stack of self-attention layers, enhancing the model's capacity to capture diverse relationships within the data.

2.3 Review of the State-of-the-Art

As the landscape of image classification evolves, Vision Transformers (ViT) have emerged at the forefront of research, demonstrating significant potential in various domains. This section reviews the current state of the art, focusing on the application of ViT across different fields and comparing its performance with traditional Convolutional Neural Networks (CNNs). Through a detailed examination of recent studies, we aim to highlight the advancements brought about by ViT, particularly its use of self-attention mechanisms for global pattern recognition and its impact on improving classification accuracy.

Vision Transformers (ViT) have undergone extensive evaluation in the field of image classification, with numerous studies showcasing their remarkable efficacy. Specifically, research [32], the focus was on diabetic retinopathy in the retina. The self-attention mechanism in ViT was utilized to recognize the level of diabetic retinopathy, employing a collection of retinal images for training and performance testing. ViT achieved an accuracy performance of 91.4% with specificity, precision, and recall of 97.7%, 92.8%, 92.6%, respectively. The comparative evaluation results of ViT against the CNN model are very comparative. The self-attention mechanism in ViT is very promising for recognizing the level of diabetic retinopathy. Furthermore, research [33] explored the performance of ViT in detecting fractures using a manually annotated dataset of fracture images according to the AO/OTA system. The ViT model was deployed for classification and subsequent evaluation, which was then compared to the CNN approach and manual classifications by medical professionals. ViT accurately predicted 85% of the images, achieving precision, recall, and f1-score values of 77%, 76%, 77% respectively. Notably, the accuracy of doctors' diagnoses improved by 29% when supplemented by ViT predictions, reaching a 97% accuracy rate.

In research [29] ViT was used for remote sensing image classification tasks with 3 datasets, namely Merced land, AID, and Optimal-31. The multi-head attention mechanism was a focal point in this research for global pattern recognition. The evaluation results on each dataset tested on ViT obtained accuracy results of 98.49% on the Merced dataset, 95.86% on the AID dataset, and 95.56% on the Optimal-31 dataset.

The synthesis of the state-of-the-art reveals that the Vision Transformer (ViT) method not only shows superiority but also holds promise for future advancements in image classification tasks. Evaluation metrics from prior research illustrate ViT's competitive edge over traditional Convolutional Neural Network (CNN) models. The distinct operational mechanisms of self-attention and multi-head attention within ViT are pivotal, diverging significantly from those of CNN models. Specifically, the self-attention mechanism empowers the model with the capability to discern and prioritize different segments of an image by assigning variable weights to specific features, thereby understanding its contextual importance. Concurrently, the multi-head attention architecture facilitates the model in processing these segments through multiple self-attention mechanisms in parallel. This approach enables the model to capture a comprehensive array of aspects and relationships within the image, viewing them from diverse perspectives. Such a multifaceted view substantially augments the model's proficiency in pattern and object recognition, by considering the minutiae of the image from various angles.

Building upon this advanced understanding, the proposed research endeavors to harness the multi-head attention capability of ViT specifically for the nuanced task of rice plant disease classification. By optimizing and tailoring the ViT model to this context, this study aims to explore the uncharted potentials of ViT in recognizing and classifying the countless diseases affecting rice plants. The forthcoming sections will delve into the methodology employed in developing and evaluating the ViT model for this purpose.3. Proposed Method

This research was carried out in several stages, namely data collecting, data preprocessing, training, and evaluation. Each stage is discussed in more detail in the subsections below.

3.1 Data Collection

This research uses two datasets, each with distinct characteristics and varying record counts, as illustrated in Figure 3. In addition to varying classes, these datasets also differ in the number of classes and the distribution of records. The first dataset [34] is a balanced dataset, 2628 color images with the .jpg extension, which has six classes. For additional details, refer to Table 1. The dataset is segmented into training and validation data, as depicted in Figure 3. This dataset has six classes, and the data distribution in each class is presented in Table 2. Images for each class are presented in Figure 4.

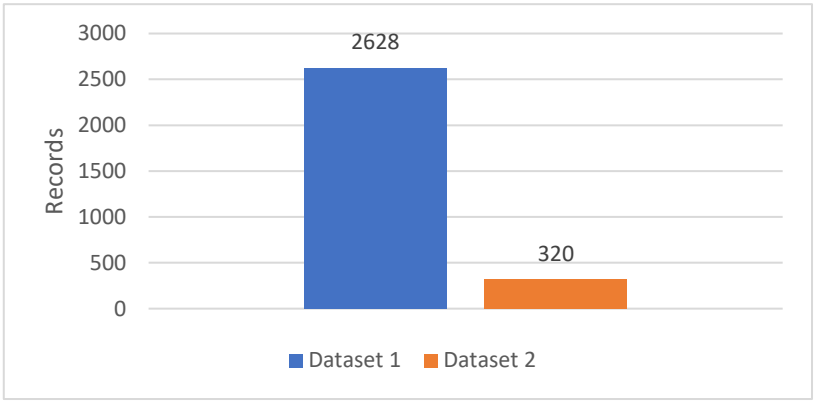

Figure 3. Datasets Record

Table 1. First Dataset Description.

Class	Records
Bacterial Leaf Blight	438
Brown Spot	438
Healthy	438
Leaf Blast	438
Leaf Scald	438
Narrow Brown Spot	438
Total	2628

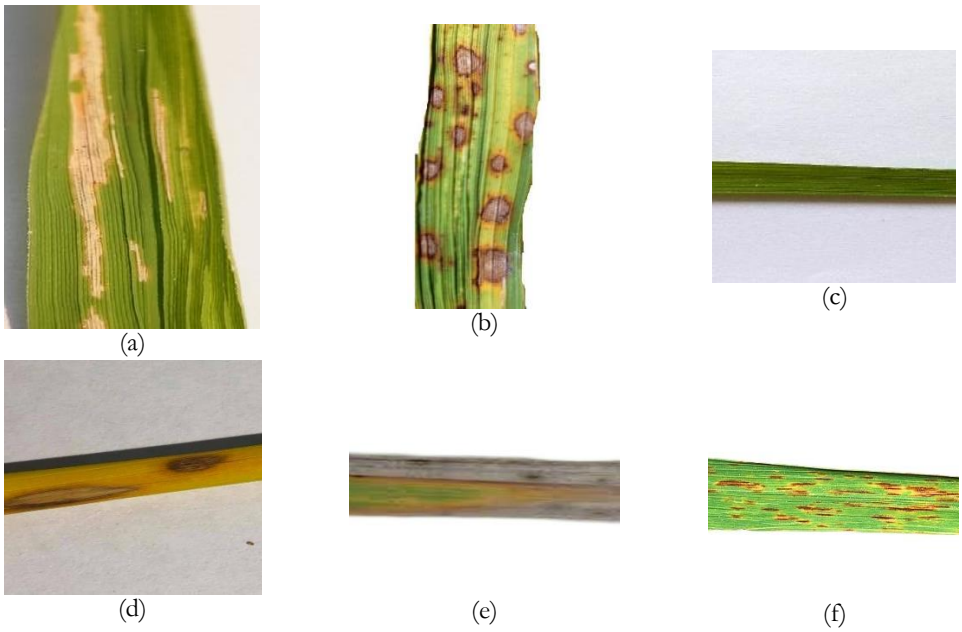

Figure 4. Image samples for each class in the first dataset (a) Bacterial Leaf Blight, (b) Brown Spot, (c) Healthy, (d) Leaf Blast, (e) Leaf Scald, (f) Narrow Brown Spot

Table 2. Second Dataset Details

Class	Records
Brown spot	40
Leaf smut	40
Blast	80
Blight	80
Tungro	80
Total	320

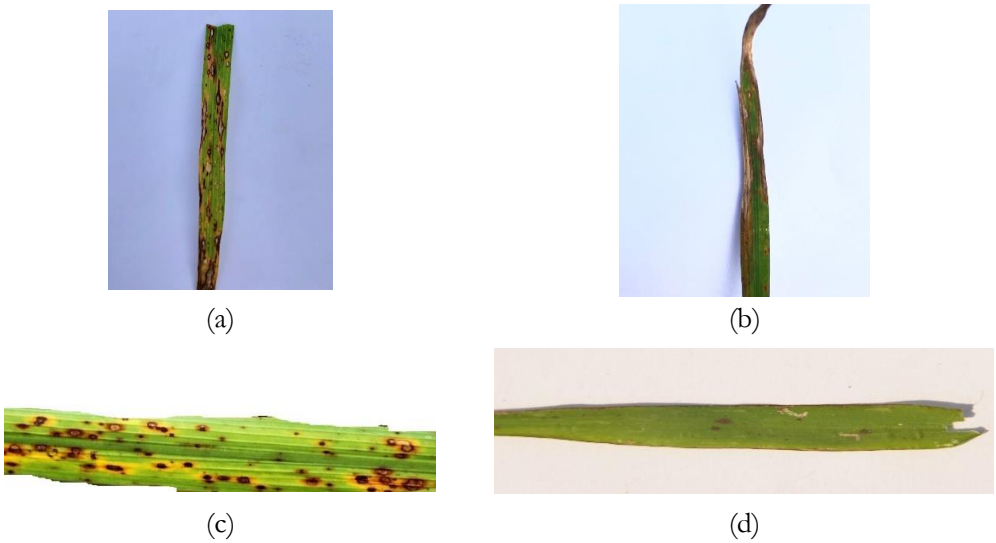

Figure 5. Image samples for each class in the second dataset (a) Blast, (b) Blight, (c) Brown Spot, (d) Leaf Smut

Meanwhile, the second dataset [35] comprises 320 color images in .jpg format, categoriezed into five distinct classes with varying data volumes, as presented in Table 3. In this study, the tungro class will be used because it refers to research [1], and this research compares

it with that research. So, only four classes are used in this dataset. For an in-depth analysis of the class distribution and to understand the extent of data imbalance, refer to Table 3. Figure 5 illustrates sample images from the selected classes within the second dataset, providing a visual representation of the data utilized in this study.

3.2 Preprocesing Dataset

All images across the datasets were resized to 224×224×3 to ensure uniform dimensions. Subsequently, the datasets were partitioned into training, validation, and testing subsets, adhering to an 80:10:10 ratio. Figure 7 presents the data distribution following the splitting process for the first dataset. Meanwhile, due to limited data volume the dataset is only divided into training and testing data subsets. The partitioning outcomes for this dataset are depicted in Figure 8.

Data augmentation was performed on the training subset to enhance image diversity, aiming to facilitate an optimal training process[13], [36], [37]. Some augmentation techniques applied include rotation, flip, and image enlargement. The goal is to vary the training data, helping the model adapt to object orientation, viewpoint, and shape differences. Visualization of the argumentation technique can be seen in Figure 6.

Figure 6. Example of Augmentation Results in Second Dataset

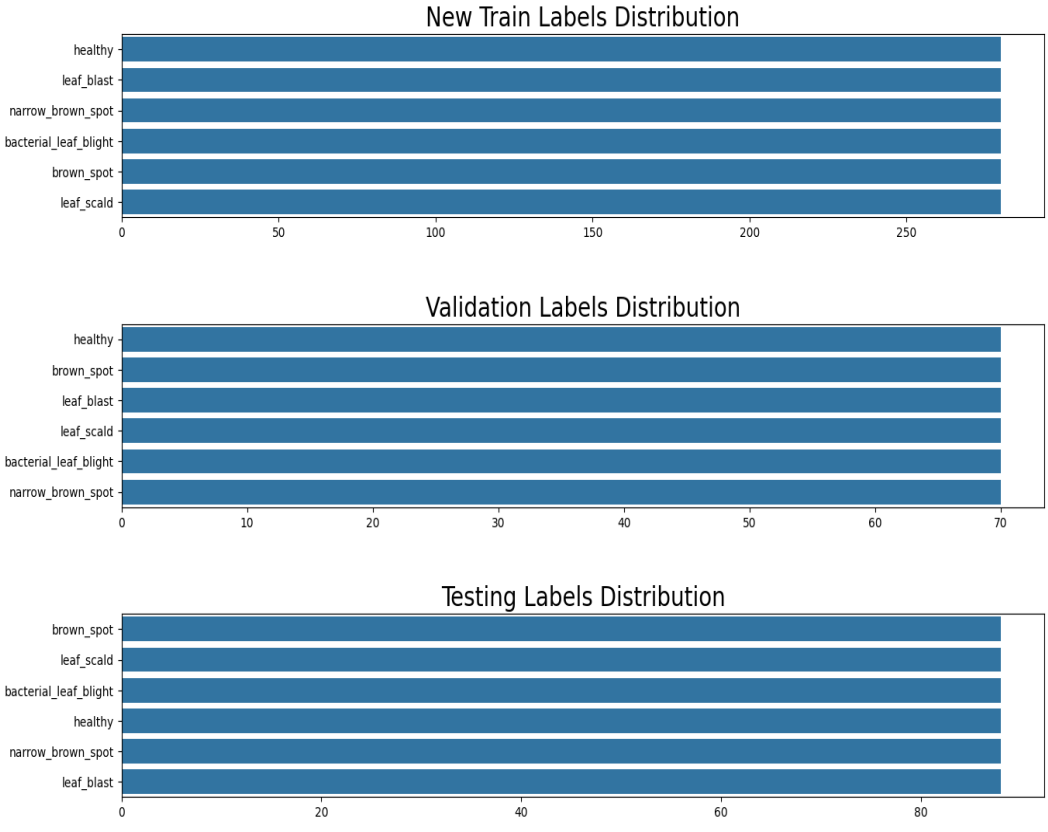

Figure 7. Splitting First Dataset

Figure 8. Splitting Second Dataset

3.3 Proposed Model Vision Transformer (ViT)

The Vision Transformer (ViT) is available in two principal variants: Base (B) and Large (L). ViT Base (B) has fewer layers and dimensions than ViT Large (L). The base is lighter and faster in learning and inference, suitable for limited resources or where inference speed is important. Meanwhile, ViT Large offers more layers and dimensions, making it more immersive and capable of capturing more complex information, but it requires more computing resources. For this study, the ViT Base (B) model, specifically the ViT-B16, was selected due to its adequacy for the relatively modest dataset size, constrained research resources, and its sufficient computational efficiency for the task at hand. Additionally, the selection of the ViT-B16 model was influenced by its adaptability for deployment in mobile environments, a critical consideration given the model's inherent flexibility and the future research aspirations. ViT-B16 was previously pre-trained on the ImageNet-21K (14 million images, 21,843 classes) and imagenet2012 (1 million images, on classes) datasets from Keras. This model has several important configurations, namely patch size, hidden dimension, Multi-Layer Perceptron (MLP) dimension, hidden dimension, attention heads, and encoder depth, and the configurations, detailed in Table 3. Patch Size is the size of a small piece of image or patch. The hidden dimension and MLP dimension define the size of the embedding layer and the number of hidden units in the MLP layers of the Transformer encoder, respectively. The number of attention heads in the Multi-Headed Self-Attention (MSA) enables the model to discern diverse patterns or relationships within the input. Lastly, the encoder depth specifies the count of encoder transformer blocks within the model.

Table 3. Proposed ViT Model Configuration

Configuration	Value
Patch Size	16
Hidden dimension	758
MLP dimension	3072
Attention Head	12
Encoder depth	12

Next, enhanced ViT B16 was carried out by adding several layers and changing the top layer according to the number of classes classified. Details are provided in Table 4.

Table 4. Detail of Proposed ViT Model

Layers (type)	Output Shape	Param#
vit-b16 (Functional)	(None, 768)	85798656
flatten_1 (Flatten)	(None, 768)	0
dense_3 (Dense) activation='relu'	(None, 64)	49216
dense_4 (Dense) activation='relu'	(None, 32)	2080
dense_5 (Dense) activation=' softmax'	(None, 4), (None, 6)	132
Total params: 85850084 (327.49 MB)		
Trainable params: 85850084 (327.49 MB)		
Non-trainable params: 0 (0.00 Byte)		

The proposed ViT model is compiled using several hyperparameters and tuning strategies, including the loss function and optimizer metrics, to evaluate the model during the training and validation process. The chosen loss function is 'categorical_crossentropy', with 'Adam' as the optimizer, and 'Accuracy' as the metric for model evaluation. Early stopping is also activated. The aim is to stop the training process based on the `val_loss` value, specifically when the model start experiencing an increase in this value, to prevent overfitting. The primary goal of employing a reduced learning rate is to dynamically adjust it during training, enhancing the efficiency of the learning process. The epoch size in this training is 25, with a batch size of 32. For more details, see Table 5.

Table 5. Hyperparameter dan Tunning Parameter ViT_B16

Parameter	Value
Optimizer	Adam (learning_rate = 0.0001)
Loss funtion	CategoricalCrossentropyc(label_smoothing = 0.0001)
Metrics	Accuracy
Callback	EarlyStopping (monitor=' val_loss ', patience=5, restore_best_weights=True)
	ReduceLROnPlateau (monitor=' val_loss ', patience=3, factor=0.001, verbose=1)
Batch size	32
Epochs	25

3.4 Model Evaluation

The evaluation of the model's performance is structured in two distinct parts: The first involves assessing the training and validation datasets using accuracy and loss metrics. The second part focuses on evaluating the test dataset through the use of a confusion matrix and accuracy metrics. The accuracy metrics employed in this research include recall, f1-score, precision, and overall accuracy. Recall quantifies the model's capability to correctly identify all true positive cases. Precision determines the proportion of positively predicted instances that are indeed positive. The f1-score, the harmonic mean of recall and precision, offering a balance between them. Specificity assessing the model's effectiveness in identifying true negative instances, thereby minimizing false positive errors. Accuracy measuring the model's overall performance in classifying the dataset accurately[38], [39]. This comprehensive evaluation aims to elucidate the Vision Transformer (ViT) model's efficacy in classification tasks prior to its practical application.

4. Results and Discussion

This research was carried out using Python language with a Jupyter Notebook editor, while the hardware specifications were an Intel I7 gen 11 processor and 16GB memory. As a note, the computing was carried out without using a GPU. Model testing uses two test scenarios, which include testing on the first dataset, namely the balance dataset, and the second on an imbalanced and relatively small dataset. The evaluation results are evaluated with the four matrices described in section 3.4 and compared with popular CNN models such as MobileNet [16], EfficientNetV2 [9], VGG16 [15], Xception [19], as well as models that use the

same dataset (specifically the second dataset). Apart from that, training time is also measured using the library time in Python.

4.1 First Scenario Test

In the first dataset, all images were used with a total of 2628, divided into training data of 1680, validation data of 420, and test data of 528. Table 6 illustrates the results of the ViT proposed model achieving a 99.35%, which shows the superiority of the proposed model compared to other CNN models. However, ViT's training time is relatively longer than other CNN models. The ViT model is relatively more complex compared to CNN models, but it also has satisfactory performance. Moreover, it requires significantly fewer epochs than other models to achieve stable values. Figure 9 presents detailed graphs of accuracy and loss for the training and validation stages. The analysis of the results depicted in the graph highlights several significant advantages of the proposed model. At the beginning of training, the model demonstrated exceptional learning capability, with a sharp decrease in loss and a rapid increase in accuracy, indicating that the used architecture could efficiently capture important features of the data. High accuracy on validation data indicates not only proficient learning of the training dataset but also effective generalization to previously unseen data. Despite initial indications of overfitting, characterized by a gap between training and validation loss, the model managed to maintain high accuracy without showing a significant increase in validation loss. This indicates that the model can maintain its generalization ability as training continues. Stabilization of loss and accuracy on validation data shows that the model has achieved consistent and reliable performance.

Table 6. Measurement Results on the First Dataset

Model	Epochs	Accuracy		Loss		Train Time (seconds)	Model Size (MB)
		Train	Val	Train	Val		
MobileNet	17	95.54%	89.29%	11.72%	27.75%	754	24.58
EfficientNetV2 B0	25	97.38%	93.10%	8.93%	18.72%	1,520	22.9
VGG 16	25	87.56%	77.86%	34.19%	53.85%	6,889	62.27
Xception	25	96.96%	91.90%	9.17%	26.21%	3,282	104.09
Proposed	16	99.35%	97.62%	1.94%	7.63%	20,457	327.49

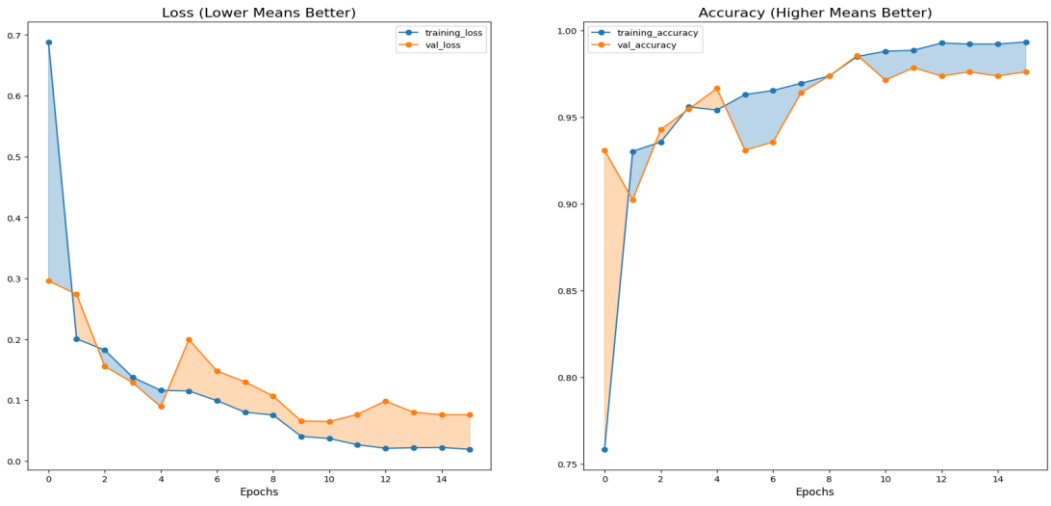

Figure 9. ViT_B16 Loss and Accuracy Graph on the First Dataset

The results of the test data evaluation in Table 7 show that the proposed model has excellent performance with recall, precision, specificity, f1-score and overall accuracy values of 0.971591, 0.972956, 0.995076 0.971709, and 0.971591, respectively. The highest results on other models are EfficientNet V2 B0 with recall, precision, specificity, f1-score, and overall accuracy values of 0.933712, 0.937222, 0.98645, 0.932942, and 0.933712, respectively.

Meanwhile, the VGG16 model obtained the lowest results. See the confusion matrix visualization in Figure 10 for clearer information on the evaluation results.

Table7. Test Data Evaluation Results on the First Dataset

Model	Recall	Precision	Specificity	F1-Score	Accuracy
Proposed	0.9716	0.9730	0.9951	0.9717	0.9716
MobileNet	0.8883	0.8921	0.9762	0.8869	0.8883
EfficientNetV2 B0	0.9337	0.9372	0.9865	0.9329	0.9337
VGG16	0.7822	0.8006	0.9510	0.7763	0.7822
Xception	0.9242	0.9256	0.9822	0.9242	0.9096

It can be concluded that the model shows excellent reliability in correct and consistent classification. Correlating these results with previous training and validation graphs, the high values are consistent with high validation accuracy and low loss, indicating that the model fits well with the training data and has also been generalized successfully to previously unseen data.

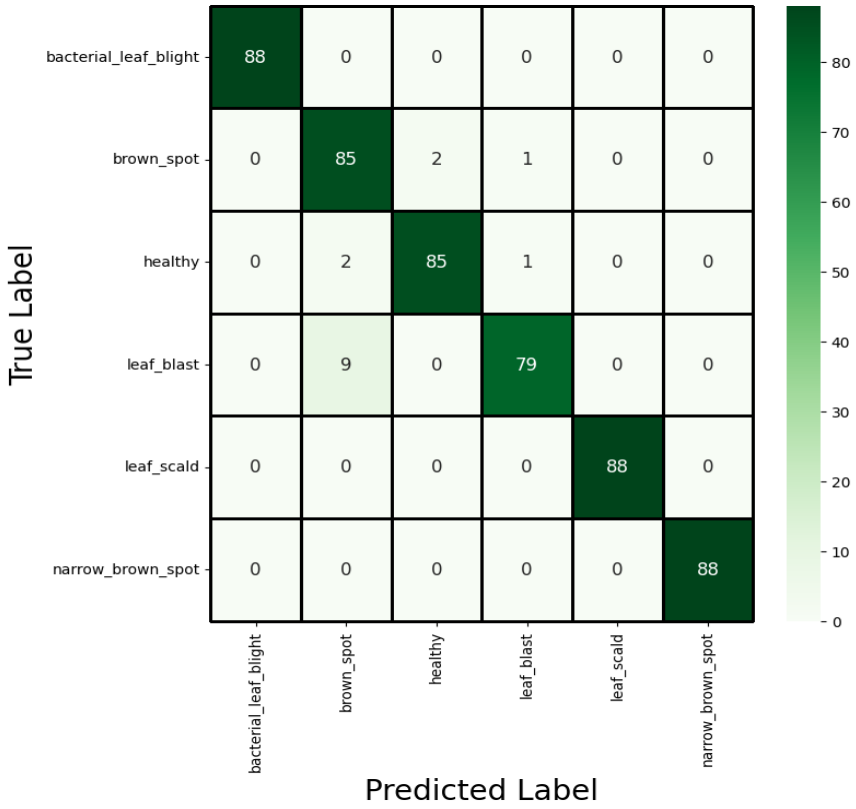

Figure 10. Confusion Matrix of First Dataset

4.2 Second Scenario Test

In this test, the dataset is relatively small and imbalanced, namely, comprising a total of 320 records. However, excluding the tungro class reduces this number to 240 records. The dataset is divided into 192 training data and 48 test data. Therefore, it appears that all models exhibit a tendency towards overfitting to some degree. However, the proposed model demonstrates superiority, achieving a training accuracy of 100% and a validation accuracy of 91.67%, as shown in Table 8. Similar to the first scenario, the training time for the proposed model is the longest, attributable to the greater complexity of ViT compared to other CNN models. Figure 11 presents more detailed training and validation graphs for the proposed model applied to the second dataset.

Table 9 demonstrates the superiority of ViT, showcasing stable results across recall, precision, specificity, f1-score, and overall accuracy, with respective values of 0.979167, 0.981481,

0.99375, 0.979085, and 0.979167. Despite the occurrence of overfitting during training and validation, testing revealed stable and superior results compared to all models, including those in research [2]. Particularly in the context of disease classification, recall is often deemed a more critical metric due to the serious consequences of failing to identify actual disease cases, especially in imbalanced datasets [38], [40]. The results achieved by ViT affirm its stability in handling diverse datasets, whether large, small, or imbalanced. See the confusion matrix visualization in Figure 12 for clearer information on the evaluation results.

Table 8. Measurement Results on the Second Dataset

Model	Epochs	Accuracy		Loss		Train Time (seconds)	Model Size (MB)
		Train	Val	Train	Val		
MobileNet	13	98.44%	87.50%	4.78%	36.07%	129.6	24.58
EfficientNetV2 B0	25	98.44%	91.67%	16.22%	21.97%	346.2	22.9
VGG 16	25	94.79%	93.75%	22.75%	29.89%	706.2	62.27
Xception	9	98.96%	87.50%	3.41%	35.55%	210	104.09
Proposed	20	100%	91.67%	0.44%	18.98%	5,580	327.49

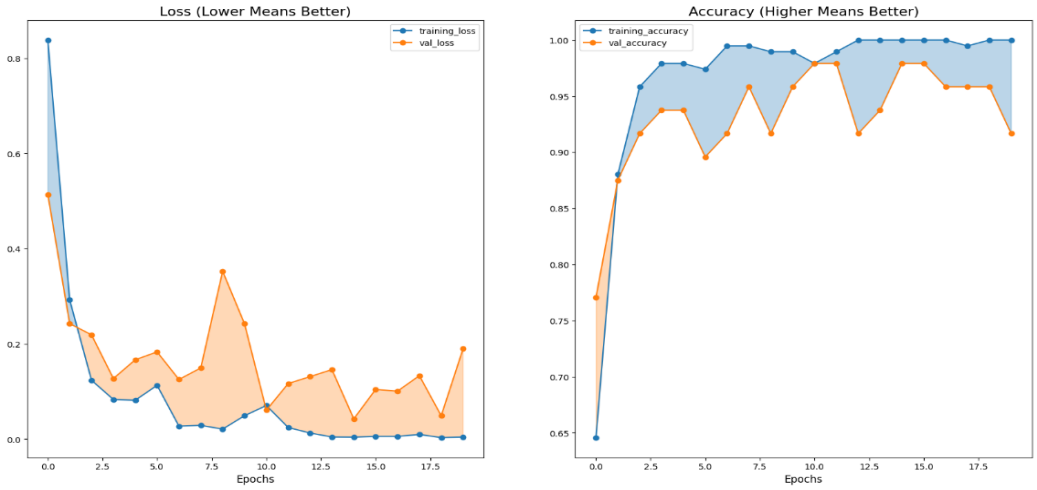

Figure 11. ViT_B16 Loss and Accuracy Graph on the Second Dataset

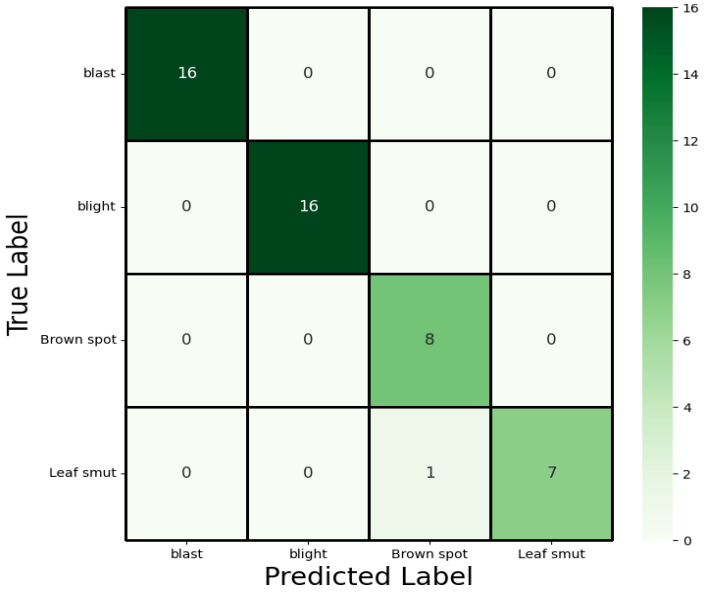

Figure 12. Confusion Matrix on the Second Dataset

Table 9. Test Data Evaluation Results on the Second Dataset

Model	Recall	Precision	Specificity	F1-Score	Accuracy
MobileNet	0.8542	0.8554	0.9547	0.8541	0.8542
EfficientNetV2 B0	0.9375	0.9392	0.9813	0.9373	0.9375
VGG16	0.8958	0.8971	0.9672	0.8958	0.8958
Xception	0.8333	0.8380	0.9469	0.8332	0.8333
Model [2]	0.9300	0.9300	0.9800	0.9300	0.9700
Proposed	0.9792	0.9815	0.9938	0.9791	0.9792

5. Conclusions

The results from two test scenarios demonstrate that the proposed model can outperform others. The first test revealed promising results for handling balanced datasets with larger amounts of data, where ViT achieved an overall accuracy of 97%, surpassing other CNN models such as EfficientNetV2_B0, which achieved an overall accuracy of 93%. It is important to note that the proposed model requires a longer training time compared to other CNN models, attributable to its extensive number of parameters, size, and complexity. The second scenario, which utilized an imbalanced dataset with significantly less data, also showed satisfactory results, indicating that the proposed model is more adaptable to various dataset characteristics. The model's accuracy, assessed using metrics such as recall, precision, specificity, and F1-score, also surpassed the state-of-the-art for the same dataset. The self-attention mechanisms for capturing global context have notably enhanced performance in complex visual tasks. In practical scenarios, classifying rice diseases based on leaf images presents specific challenges, including variations in leaf appearance due to lighting conditions, shooting angles, and the stage of disease development. These challenges necessitate a model capable of understanding complex visual contexts and the subtle differences between disease categories. ViT has shown considerable promise in this regard, owing to its capability to process global information from images. This enables a more accurate identification of disease features by understanding the context of the entire image, rather than relying solely on local features. Consequently, ViT emerges as a promising candidate for further development in the recognition or classification of rice diseases based on leaf images.

The exploration of the Vision Transformer (ViT) model for rice disease recognition presents a promising advancement beyond traditional CNN approaches, offering enhanced performance through global context capture and self-attention mechanisms. The research demonstrated ViT's superiority in handling complex visual tasks, outperforming established CNN models across balanced and imbalanced datasets. Future endeavors could focus on refining ViT's efficiency for broader application, including optimizing its computational demands for real-time use and enhancing its adaptability across diverse agricultural environments. This streamlined approach would not only advance disease classification techniques but also pave the way for integrating cutting-edge AI into sustainable agricultural practices, ultimately contributing to improved crop management and yield.

Author Contributions: Conceptualization: R.K.R. and D.R.I.M.; methodology: R.K.R. and D.R.I.M.; software: R.K.R.; validation: D.R.I.M., A.S. and K.N.; formal analysis: D.R.I.M., A.S. and K.N.; investigation: D.R.I.M., A.S. and K.N.; resources: R.K.R. data curation: X.X.; writing—original draft preparation: R.K.R.; writing—review and editing: H.M.M.I. and D.R.I.M.; visualization: A.S. and K.N.; supervision: D.R.I.M.; project administration: H.M.M.I.; funding acquisition: All.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] E. B. Wijayanti, D. R. I. M. Setiadi, and B. H. Setyoko, "Dataset Analysis and Feature Characteristics to Predict Rice Production based on eXtreme Gradient Boosting," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 299–310, Feb. 2024, doi: 10.62411/jcta.10057.

[2] D. J. Chaudhari and K. Malathi, "Detection and Prediction of Rice Leaf Disease Using a Hybrid CNN-SVM Model," *Opt. Mem. Neural Networks*, vol. 32, no. 1, pp. 39–57, Mar. 2023, doi: 10.3103/S1060992X2301006X.
[3] P. I. Ritharson, K. Raimond, X. A. Mary, J. E. Robert, and A. J, "DeepRice: A deep learning and deep feature based classification of Rice leaf disease subtypes," *Artif. Intell. Agric.*, vol. 11, pp. 34–49, Mar. 2024, doi: 10.1016/j.aiia.2023.11.001.
[4] M. T. Ahad, Y. Li, B. Song, and T. Bhuiyan, "Comparison of CNN-based deep learning architectures for rice diseases classification," *Artif. Intell. Agric.*, vol. 9, pp. 22–35, Jul. 2023, doi: 10.1016/j.aiia.2023.07.001.
[5] S. P. Singh, K. Pritamdas, K. J. Devi, and S. D. Devi, "Custom Convolutional Neural Network for Detection and Classification of Rice Plant Diseases," *Procedia Comput. Sci.*, vol. 218, pp. 2026–2040, 2023, doi: 10.1016/j.procs.2023.01.179.
[6] R. R. Kovvuri, A. Kaushik, and S. Yadav, "Disruptive technologies for smart farming in developing countries: Tomato leaf disease recognition systems based on machine learning," *Electron. J. Inf. Syst. Dev. Ctries.*, vol. 89, no. 6, Nov. 2023, doi: 10.1002/isd2.12276.
[7] S. Saha and S. M. M. Ahsan, "Rice Disease Detection using Intensity Moments and Random Forest," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, Feb. 2021, pp. 166–170. doi: 10.1109/ICICT4SD50815.2021.9396986.
[8] B. Chakraborty *et al.*, "Detection of Rice Blast Disease (Magnaporthe grisea) Using Different Machine Learning Techniques," *Int. J. Environ. Clim. Chang.*, vol. 13, no. 8, pp. 2256–2264, Jun. 2023, doi: 10.9734/ijecc/2023/v13i82190.
[9] M. A. Araaf, K. Nugroho, and D. R. I. M. Setiadi, "Comprehensive Analysis and Classification of Skin Diseases based on Image Texture Features using K-Nearest Neighbors Algorithm," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 31–40, Sep. 2023, doi: 10.33633/jcta.v1i1.9185.
[10] C. Zuo *et al.*, "Deep learning in optical metrology: a review," *Light Sci. Appl.*, vol. 11, no. 1, p. 39, Feb. 2022, doi: 10.1038/s41377-022-00714-x.
[11] S. B. Imanulloh, A. R. Muslikh, and D. R. I. M. Setiadi, "Plant Diseases Classification based Leaves Image using Convolutional Neural Network," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 1–10, Aug. 2023, doi: 10.33633/jcta.v1i1.8877.
[12] M. S. Sunarjo, H. Gan, and D. R. I. M. Setiadi, "High-Performance Convolutional Neural Network Model to Identify COVID-19 in Medical Images," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 19–30, Aug. 2023, doi: 10.33633/jcta.v1i1.8936.
[13] H. T. Adityawan, O. Farroq, S. Santosa, H. M. M. Islam, M. K. Sarker, and D. R. I. M. Setiadi, "Butterflies Recognition using Enhanced Transfer Learning and Data Augmentation," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 115–128, Nov. 2023, doi: 10.33633/jcta.v1i2.9443.
[14] S. Ghosal and K. Sarkar, "Rice Leaf Diseases Classification Using CNN With Transfer Learning," in *2020 IEEE Calcutta Conference (CALCON)*, Feb. 2020, pp. 230–236. doi: 10.1109/CALCON49167.2020.9106423.
[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*, 2015, pp. 1–14.
[16] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," Apr. 2017, [Online]. Available: http://arxiv.org/abs/1704.04861
[17] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," May 2019, [Online]. Available: http://arxiv.org/abs/1905.11946
[18] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.00298
[19] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, vol. 7, no. 3, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[22] A. Kesarwani, S. Das, D. R. Kisku, and M. Dalui, "Non-invasive anaemia detection based on palm pallor video using tree-structured 3D CNN and vision transformer models," *J. Exp. Theor. Artif. Intell.*, pp. 1–29, Jan. 2024, doi: 10.1080/0952813X.2023.2301401.
[23] N. Perwaiz, M. Shahzad, and M. Moazam Fraz, "TransPose Re-ID: transformers for pose invariant person Re-identification," *J. Exp. Theor. Artif. Intell.*, pp. 1–14, May 2023, doi: 10.1080/0952813X.2023.2214570.
[24] X. Gao, Z. Xiao, and Z. Deng, "High accuracy food image classification via vision transformer with data augmentation and feature augmentation," *J. Food Eng.*, vol. 365, p. 111833, Mar. 2024, doi: 10.1016/j.jfoodeng.2023.111833.
[25] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929
[26] F. Jerbi, N. Aboudi, and N. Khlifa, "Automatic classification of ultrasound thyroids images using vision transformers and generative adversarial networks," *Sci. African*, vol. 20, p. e01679, Jul. 2023, doi: 10.1016/j.sciaf.2023.e01679.
[27] G. I. Okolo, S. Katsigiannis, and N. Ramzan, "IEViT: An enhanced vision transformer architecture for chest X-ray image classification," *Comput. Methods Programs Biomed.*, vol. 226, p. 107141, 2022, doi: 10.1016/j.cmpb.2022.107141.
[28] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[29] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Al Dayil, and N. Al Ajlan, "Vision Transformers for Remote Sensing Image Classification," *Remote Sens.*, vol. 13, no. 3, p. 516, Feb. 2021, doi: 10.3390/rs13030516.
[30] D. Shome *et al.*, "COVID-Transformer: Interpretable COVID-19 Detection Using Vision Transformer for Healthcare," *Int. J. Environ. Res. Public Health*, vol. 18, no. 21, p. 11086, Oct. 2021, doi: 10.3390/ijerph182111086.
[31] E. Goceri, "Vision transformer based classification of gliomas from histopathological images," *Expert Syst. Appl.*, vol. 241, no. November 2023, p. 122672, May 2024, doi: 10.1016/j.eswa.2023.122672.
[32] J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, "Vision Transformer-based recognition of diabetic retinopathy grade," *Med. Phys.*, vol. 48, no. 12, pp. 7850–7863, Dec. 2021, doi: 10.1002/mp.15312.

[33] L. Tanzi, A. Audisio, G. Cirrincione, A. Aprato, and E. Vezzetti, "Vision Transformer for femur fracture classification," *Injury*, vol. 53, no. 7, pp. 2625–2634, Jul. 2022, doi: 10.1016/j.injury.2022.04.013.
[34] D. I. D. Saputra, "Rice Leafs Disease Dataset," *Kaggle.com*, 2022. https://www.kaggle.com/datasets/dedeikhsandwisaputra/rice-leafs-disease-dataset (accessed Aug. 16, 2023).
[35] C. G, "Rice-leaf-disease," *Kaggle.com*, 2022. https://www.kaggle.com/datasets/chandrug/riceleafdisease (accessed Oct. 20, 2023).
[36] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 60, 2019, doi: 10.1186/s40537-019-0197-0.
[37] J. J. J. Chen, J. J. J. Chen, D. Zhang, Y. Sun, and Y. A. A. Nanehkaran, "Using deep transfer learning for image-based plant disease identification," *Comput. Electron. Agric.*, vol. 173, no. November 2019, p. 105393, Jun. 2020, doi: 10.1016/j.compag.2020.105393.
[38] F. S. Gomiasti, W. Warto, E. Kartikadarma, J. Gondohanindijo, and D. R. I. M. Setiadi, "Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 396–406, Mar. 2024, doi: 10.62411/jcta.10106.
[39] T. R. Noviandy, K. Nisa, G. M. Idroes, I. Hardi, and N. R. Sasmita, "Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer's Drug Discovery with LightGBM," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 358–367, Mar. 2024, doi: 10.62411/jcta.10129.
[40] F. Mustofa, A. N. Safriandono, A. R. Muslikh, and D. R. I. M. Setiadi, "Dataset and Feature Analysis for Diabetes Mellitus Classification using Random Forest," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 41–48, Jan. 2023, doi: 10.33633/jcta.v1i1.9190.