Volume 01, Issue 01, Juni 2026 : ISSN: xxx–xxx Received: Mei 27, 2026. Revised: Juni 19, 2026

Journal of Adaptive Intelligent Systems

Vision Transformer-Based Rice Leaf Disease Classification with Imbalanced Data Handling Using Image Augmentation

Nasywa Alya Musyaffa' [1]* Tamam Asrori [2] Dyah Ariyanti [3] Dwi Putri Kartini [4] Ira Aprilia [5]

[1,2,3,4]*Universitas Panca Marga, Probolinggo, Indonesia*
* Corresponding author's Email : musyaffa.a.nasywa@gmail.com

Abstract: Image-based classification of rice leaf diseases helps identify issues quickly and accurately, but data imbalance poses a challenge in deep learning models. This study analyzes the effect of handling imbalanced data using a Vision Transformer-based image augmentation strategy on the Paddy Doctor dataset which contains 10,407 images across 9 disease classes and 1 healthy class. Experiments tested five scenarios including no augmentation, flipping, rotate, cropping, and noise injection, evaluated using accuracy, precision, recall, and F1-Score metrics. Results show the scenario without augmentation provides the best performance with 96.29% accuracy and a 0.9588 macro F1-score on test data. All augmentation techniques decrease model accuracy by 1.41% to 1.94%. Among the tested techniques, cropping gave the best results while rotate produced the lowest performance. This research concludes that generic image augmentation does not always improve Vision Transformer performance because its effectiveness heavily depends on visual characteristics and data distribution.

Keywords: vision transformer, image augmentation, rice leaf disease classification, data imbalance, deep learning.

1. Introduction

Rice is a primary food source that plays a strategic role in supporting national food security, particularly in agrarian countries such as Indonesia [1]. According to the Central Bureau of Statistics (Badan Pusat Statistik) [2], rice production in Indonesia in 2024 reached 53.14 million tons of dry unhulled grain (Gabah Kering Giling/GKG), a decrease of 1.55% compared with the previous year. This condition underlines the importance of efforts to increase rice productivity. One factor that may contribute to declining production is rice leaf disease [3], which can disturb plant growth and therefore requires early and accurate identification to minimize potential production losses.

Early disease identification requires an accurate and efficient diagnostic method. Although disease symptoms in rice plants can be observed visually, farmers' limited knowledge may lead to misdiagnosis and inappropriate treatment [4]. Advances in artificial intelligence, particularly deep learning, offer a solution to this limitation through image classification techniques capable of automatically recognizing disease patterns [3].

Convolutional Neural Networks (CNN) have been the dominant early approach, extracting local features such as texture and shape through convolutional layers; however, this approach has a fundamental limitation in handling large visual variation among objects [5]. To address this limitation, the Vision Transformer (ViT) architecture processes images globally by dividing them into a series of small patches and analyzing the contextual relationships among all patches through a self-attention mechanism [5]. This capability enables ViT to capture long-range spatial dependencies and to understand the overall context of an image [5]. - Nevertheless, transformer-based models are typically very large and computationally expensive [6].

In practice, rice leaf disease image datasets tend to exhibit class imbalance, which causes model learning to become biased toward the majority class and reduces performance on minority classes [7]. An approach that can address this problem without increasing model architectural complexity is therefore needed. Image augmentation is an effective technique for increasing data diversity without physically adding new data, allowing a model to learn from a wider range of image variation without requiring additional computational resources [8].

Previous studies [8] on the ViT architecture and [9] on the CNN architecture employed augmentation as a strategy to increase data variation in different domains. However, these studies focused mainly on overall model performance and did not provide an in-depth analysis of minority-class performance. This indicates that research on the effect of augmentation on the distribution of performance across classes remains limited, particularly for rice leaf disease classification.

Building on this background, this study evaluates the effect of image augmentation as a method for handling imbalanced data on the performance of the

Journal of Adaptive Intelligent Systems

Vision Transformer in rice leaf disease classification. The evaluation is conducted not only on overall performance but also specifically on the model's ability to recognize minority classes under imbalanced data conditions, using four single augmentation techniques flipping, rotation, cropping, and noise injection applied separately to the minority classes, and assessed through accuracy, precision, recall, and F1-score. The results are expected to provide empirical understanding of imbalanced-data handling strategies and to serve as a reference for future research on deep-learning-based plant disease classification systems.

2. Related Works

Several studies have applied deep learning approaches, including Convolutional Neural Network (CNN) and Vision Transformer (ViT) architectures, together with image augmentation techniques, across a range of domains related to image-based classification and detection.

Hairani & Widiyaningtyas [10] investigated rice leaf disease detection from digital images using a CNN architecture, addressing the classification of three rice leaf disease classes and applying data augmentation to increase data variation. Their results showed that the CNN approach combined with data augmentation significantly improved model performance, achieving an accuracy of 99.7% and outperforming several previous methods. However, that study did not discuss class-specific performance.

Febriyanto & Syofian [3] applied a Vision Transformer for image-based rice leaf disease classification across five disease classes. The results showed that the Vision Transformer achieved accuracy, precision, recall, and F1-score of 96%, indicating the model's potential to effectively capture visual patterns of rice leaf disease. Nevertheless, this study did not explicitly examine the effect of augmentation on model performance.

Mirwansyah et al. [9] applied image augmentation to address dataset limitations in CNN-based oil palm tree object detection. Their results showed improved model performance, with a mean Average Precision (mAP) of 97.2%, precision of 94.3%, and recall of 93.3%, confirming the effectiveness of image augmentation in improving deep learning performance under limited-data conditions.

Ardiyansyah et al. [8] applied a Vision Transformer for image-based breast cancer classification using mammography images, employing image augmentation as a data-handling strategy without modifying the model architecture. The results showed that the model achieved high classification performance, with an accuracy of 92.50%, precision of 90.48%, recall of 95%, and F1-score of 92.68%, confirming that image augmentation can effectively improve ViT performance without increasing architectural or training complexity.

The propose work image augmentation has been shown to improve the performance of both CNN and Vision Transformer models across several domains, the focus of augmentation in these studies remains general, targeting overall model performance rather than class-specific performance under imbalanced-data conditions. Moreover, none of the prior studies specifically examine the effect of augmentation on rice leaf disease classification using a Vision Transformer with attention to minority-class performance. Building on this gap, the present study focuses on the effect of applying augmentation to minority classes, which frequently experience imbalanced data conditions due to differences in the number of available samples across disease classes.

3. Research Method

This study adopts an experimental research design to evaluate the effect of image augmentation on the performance of a Vision Transformer (ViT) model in classifying rice leaf disease images under imbalanced data conditions.

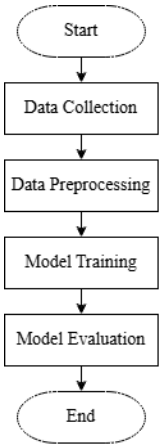

Figure 1. Research Method Diagram

Based on Figure 1, the research process consists of four main stages: data collection, data

Journal of Adaptive Intelligent Systems

preprocessing, model training, and model evaluation. The process begins with the data collection phase, where a collection of rice leaf disease images is gathered under imbalanced conditions. Next, during data preprocessing, these raw images undergo cleaning and image augmentation techniques specifically designed to mitigate the class imbalance and prepare them for optimal input. The enhanced dataset is then utilized in the model training stage, where the Vision Transformer (ViT) architecture learns to identify and recognize patterns associated with different diseases. Finally, the workflow concludes with evaluation, where the trained ViT model's classification performance is rigorously tested using specific metrics to determine the effectiveness of the applied image augmentation methods.

3.1 Data Collection

The first stage involves searching for a dataset matching the required data characteristics, namely rice leaf disease images with an imbalanced class distribution. The dataset used in this study is the Paddy Doctor dataset [11]. This dataset contains a total of 10,407 images divided into 10 classes, which include 9 disease categories and 1 healthy leaf category. The selected dataset is downloaded from Kaggle as the primary data source for this study and subsequently imported into Google Colab for further processing and model training.

3.2 Data Preprocessing

The imported dataset is first read into the programming environment for further processing. It is then divided into three subsets with a 70:15:15 ratio, consisting of a training set (70%) to train the model, a validation set (15%) to validate the model during training, and a test set (15%) for the final evaluation of the model.

Several image augmentation techniques are subsequently applied to the training data only as follows :

1. No augmentation as baseline for comparison to other image augmentation.
2. Flipping generates image variation by horizontally or vertically mirroring the image orientation, increasing pattern diversity without altering the main object information.
3. Rotation generates image variation by rotating the image at a certain angle, improving the model's robustness to changes in object orientation.
4. Cropping generates image variation by extracting a portion of the image area, encouraging the model to focus on relevant local features.
5. Noise Injection generates image variation by adding random pixel-level disturbance, improving the model's robustness to visual distortion or noise.

All images are then resized to a uniform dimension of 224×224 pixels to match the model's input requirements. Preprocessing is completed through two additional processes: label alignment, which ensures that each image is assigned a class label in the form of a consistent numeric index corresponding to the order of classes at the model output; and patch embedding, in which each image is divided into small patches of 16×16 pixels that are subsequently transformed into numerical representations suitable for input into the ViT model. Once these preprocessing steps are completed, the dataset is ready to be used for model training.

3.3 Model Training

Model training begins with the preparation of the previously preprocessed training data. A pretrained Vision Transformer model is then loaded from Hugging Face using the google/vit-base-patch16-224 checkpoint. Training parameters are configured, including a learning rate of 4e-5, a batch size of 32, and a total duration of 8 epochs for each scenario. Training is then carried out using the training dataset and the specified parameter configuration. During training, the process is monitored in terms of training loss, validation loss, and model convergence, as well as the evaluation metrics of accuracy, precision, recall, and F1-score, until a trained model ready for the evaluation stage is obtained.

3.4 Evaluation

Model evaluation begins with preparing the computational environment and supporting tools, including loading the trained model and preparing the validation and test sets. The model is first evaluated on the validation set to monitor its performance during the development process. It is then tested on the test set to objectively measure its generalization

Journal of Adaptive Intelligent Systems

capability on data not used during training. The results of the evaluation on both the validation and test sets are presented through metric comparisons and a confusion matrix, allowing the effect of augmentation on minority-class performance to be assessed consistent with the primary focus of this study.

4. Result and Discussion

This section presents the experimental results of applying single image augmentation techniques to minority classes for rice leaf disease classification using a Vision Transformer (ViT), together with a discussion of their effect on overall and class-specific performance.

4.1 Experimental Scenario Overview

Five training scenarios were designed to evaluate the effect of single, separately applied augmentation techniques on the minority classes of the dataset. Each scenario differs only in the augmentation strategy applied, allowing the specific impact of each image manipulation technique on classification performance to be objectively analyzed. Table 1 summarizes the five training scenarios.

Table 1. Summary of the Five Training Scenarios

Scenario	Augmentation Applied
S1	No Augmentation
S2	Flipping
S3	Rotation
S4	Cropping
S5	Noise Injection

4.2 Data Preprocessing Results

4.2.1 Dataset Distribution and Class Imbalance

The dataset used in this study consists of 10 rice leaf disease classes, including one normal (healthy) class, totaling 10,407 images. The dataset was split into 70% training data (7,284 images), 15% validation data (1,561 images), and 15% test data (1,562 images). The distribution per class is presented in Table 2.

Table 2. Data Split Results

Disease Class	Train (70%)	Validation (15%)	Test (15%)
bacterial_leaf_blight	335	72	72
bacterial_leaf_streak	266	57	57
bacterial_panicle_blight	236	50	51
blast	1216	261	261
brown_spot	675	145	145
dead_heart	1010	216	216
downy_mildew	434	93	93
hispa	1116	239	239
normal	1234	265	265
tungro	762	163	163
Total	7.284	1.561	1.562

The degree of class imbalance was quantified using the imbalance ratio (IR), calculated as the ratio between the number of training samples in the majority class and the number of training samples in each individual class [12]. as defined in Equation (1).

$$IR = \frac{N_{majority}}{N_{class}} \quad (1)$$

where IR denotes the class imbalance ratio, $N_{majority}$ is the number of samples in the largest training class, and N_{class} is the number of samples in the class being evaluated. The normal class, with 1,234 training images, was used as the reference majority class. Table 3 presents the imbalance ratio for each class.

Table 3. Imbalance Ratio of the Training Data

Disease Class	Training Samples	Imbalance Ratio
bacterial_panicle_blight	236	5.229
bacterial_leaf_streak	266	4.639
bacterial_leaf_blight	335	3.684
downy_mildew	434	2.843
brown_spot	675	1.828
tungro	762	1.619
dead_heart	1,010	1.222
hispa	1,116	1.106
blast	1,216	1.015
normal	1,234	1.000

The five classes with the highest imbalance ratio—bacterial_panicle_blight, bacterial_leaf_streak, bacterial_leaf_blight, downy_mildew, and brown_spot—were identified as the minority classes. Image augmentation was therefore applied specifically to these five classes to reduce the class distribution gap during model training. The number of augmented images assigned to each minority class in every scenario was adjusted

Journal of Adaptive Intelligent Systems

based on a predetermined target quota: classes with fewer training samples received a proportionally larger number of augmented images, while classes whose sample count already exceeded the target quota received no additional augmentation.

Overall, the classes with the lowest training sample counts—Bacterial Leaf Blight, Bacterial Leaf Streak, Bacterial Panicle Blight, Brown Spot, and Downy Mildew—received the largest augmentation quotas, while classes with sufficient initial data (Blast, Dead Heart, Hispa, Normal, and Tungro) were not augmented, in order to avoid further increasing the overall class imbalance.

4.2.2 Model Implementation and Training Parameters

The loaded google/vit-base-patch16-224 architecture processes the 224 × 224 pixel inputs into 16 × 16 patches for feature extraction, subsequently mapping them into the ten target disease classes. To isolate the effects of the evaluated augmentation strategies, these training parameters are kept identical across all five experimental scenarios.

4.3 Model Training Results

Training across all five scenarios exhibited a positive learning pattern, with training loss decreasing sharply from the first to the eighth epoch, consistently dropping by more than 99% in every scenario. This indicates that the model was able to effectively learn the training data regardless of the augmentation condition applied, with no indication of convergence failure. Table 4 summarizes the training and validation metrics obtained at the final epoch (epoch 8) for each scenario.

Table 4. Final-Epoch Training and Validation Metrics for All Scenarios

Scenario	Train Loss	Val Loss	Macro F1-Score	Acc
No Augmentation	0.000743	0.122200	0.965585	0.969250
Flipping	0.000473	0.152363	0.958697	0.959641
Rotate	0.000489	0.150312	0.958906	0.961563
Cropping	0.000471	0.155597	0.954874	0.957719
Noise Injection	0.000488	0.160819	0.957998	0.960922

The no-augmentation scenario achieved the lowest final validation loss (0.1222) and the highest macro F1-score (0.9656) and accuracy (0.9693) among all scenarios, followed by rotate (0.9589), flipping (0.9587), noise injection (0.9580), and cropping (0.9549)—although the differences between scenarios remained below 1.2%. While training loss decreased consistently in every scenario, validation loss tended to fluctuate slightly across epochs, which is a normal characteristic of the optimization and weight-adjustment process rather than an indication of instability. The no-augmentation scenario showed the most stable convergence pattern with minimal fluctuation in validation metrics throughout training, whereas the augmented scenarios exhibited larger fluctuations in the early epochs before stabilizing in later epochs—suggesting that the added data variability required additional adaptation time. The gap between training and validation loss remained relatively consistent across all scenarios, indicating that none of the models suffered from severe overfitting or underfitting.

4.4 Model Evaluation Results

4.4.1 Validation Performance Comparison

Because the dataset exhibits class imbalance, evaluation was based not only on accuracy but also on macro precision, macro recall, and macro F1-score, which weight the performance of every class equally and therefore provide a more representative measure of overall classification capability. Table 5 summarizes the validation metrics for all scenarios.

Table 5. Validation Metric Comparison Across Scenarios

Scenario	Accuracy	Macro Precision	Macro Recall	Macro F1-Score
No Augmentation	0.969250	0.967492	0.963883	0.965584
Flipping	0.960922	0.961443	0.961086	0.960903
Rotate	0.962844	0.966678	0.955899	0.960888
Cropping	0.961563	0.956419	0.961588	0.958605
Noise Injection	0.960922	0.961415	0.955554	0.957997

All scenarios achieved a validation accuracy above 96%, with a difference between the best and worst scenarios of less than 1%, indicating that none of the augmentation techniques caused a drastic drop in validation performance. The no-augmentation scenario achieved the highest macro precision (0.9675) and macro recall (0.9639), while the cropping scenario recorded the lowest macro precision (0.9564) and the noise injection scenario the lowest macro recall (0.9556).

Journal of Adaptive Intelligent Systems

4.4.2 Test Comparison and Generalization Comparison

The final test evaluation, conducted on data unseen during both training and validation, provides an unbiased estimate of each model's generalization capability. Table 6 summarizes the test results.

Table 6. Test Metric Comparison Across Scenarios

Scenario	Test Loss	Acc	Macro Precision	Macro Recall	Macro F1-Score
No Augmentation	0.161269	0.962868	0.960390	0.957563	0.958827
Flipping	0.202926	0.947503	0.945623	0.940472	0.942688
Rotate	0.217020	0.944302	0.947636	0.932397	0.939363
Cropping	0.179324	0.950064	0.949862	0.940802	0.944733
Noise Injection	0.202304	0.943021	0.944596	0.938490	0.940404

The no-augmentation scenario again produced the best test performance (accuracy 96.29%, macro F1-score 0.9588) and the lowest test loss (0.1613). Among the augmented scenarios, cropping achieved the highest macro F1-score (0.9447), followed by flipping (0.9427), noise injection (0.9404), and rotate (0.9394)—which recorded the highest test loss (0.2170). The gap between the best and worst scenarios reached nearly 2% in macro F1-score, indicating that certain augmentation techniques were less effective at preserving performance when the model encountered the natural variation present in the test set. This suggests that, although augmentation increases training data variety, it may also cause the model to become somewhat overfitted to the specific characteristics of the applied transformation, reducing robustness to the natural variation of unseen data.

Table 7 compares the validation-to-test performance drop for each scenario and ranks their generalization stability.

Table 7. Validation-to-Test Performance Drop and Stability Ranking

Scenario	Macro F1 (Validation)	Macro F1 (Test)	Val-to-Test Drop	Stability Rank
No Augmentation	0.965584	0.958827	0.006757 (0.68%)	1
Cropping	0.958605	0.944733	0.013872 (1.39%)	2
Noise Injection	0.957997	0.940404	0.017593 (1.76%)	3
Flipping	0.960903	0.942688	0.018215 (1.82%)	4
Rotate	0.960888	0.939363	0.021525 (2.15%)	5

The no-augmentation scenario showed the smallest performance drop from validation to test, confirming that the model trained on the original, unaltered data has the most consistent feature representation. Among the augmented scenarios, cropping showed the smallest drop and the best generalization stability, likely because the operation removes irrelevant background information while preserving the local spatial detail and orientation of the disease symptoms. Rotate showed the largest drop, suggesting that the ViT model relies heavily on the natural spatial orientation of disease symptoms relative to the leaf-vein structure; random rotation disrupts this spatial cue and creates a distribution mismatch that increases misclassification when the model is tested on naturally oriented leaf images.

4.4.3 Per-Class Performance Analysis

At the class level, performance varied considerably depending on the visual complexity and distinctiveness of each disease. Across all five scenarios, the dead_heart class consistently obtained the highest F1-score (ranging from 0.9788 to 0.9860), indicating that its visual characteristics are highly distinct and easily recognized by the model. In contrast, the downy_mildew class consistently obtained the lowest F1-score (ranging from 0.8619 to 0.9005), primarily due to low recall (0.8602–0.9247), suggesting that the model frequently fails to identify this class, likely because of subtle or overlapping visual symptoms with other diseases. Other moderately performing classes included blast (macro F1 ≈ 0.9399) and hispa (macro F1 ≈ 0.9530), both of which showed lower recall relative to precision, indicating that the model was comparatively conservative in predicting these classes.

Table 8 focuses on the five minority classes targeted by augmentation, comparing their test performance across all scenarios.

Table 8. Test Performance of Minority Classes Across Scenarios

Minority Class	Scenario	Precision	Recall	F1-Score
bacterial_panicle_blight	No Augmentation	0.960784	0.960784	0.960784
	Flipping	0.905660	0.941176	0.923077
	Rotate	0.941176	0.941176	0.941176
	Cropping	0.890909	0.960784	0.924528
	Noise Injection	0.890909	0.960784	0.924528
bacterial_leaf_streak	No Augmentation	0.982143	0.964912	0.973451
	Flipping	1.000000	0.929824	0.963636
	Rotate	1.000000	0.964912	0.982143
	Cropping	1.000000	0.947368	0.972973
	Noise Injection	0.982456	0.982456	0.982456
bacterial_leaf_blight	No Augmentation	0.958904	0.972222	0.965517

Journal of Adaptive Intelligent Systems

	Flipping	0.956522	0.916667	0.936170
	Rotate	0.968750	0.861111	0.911765
	Cropping	0.984615	0.888889	0.934307
	Noise Injection	1.000000	0.847222	0.917293
downy_mildew	No Augmentation	0.919540	0.860215	0.888889
	Flipping	0.884211	0.903226	0.893617
	Rotate	0.886364	0.838710	0.861878
	Cropping	0.879121	0.860215	0.869565
	Noise Injection	0.877551	0.924731	0.900524
brown_spot	No Augmentation	0.952381	0.965517	0.958904
	Flipping	0.971223	0.931034	0.950704
	Rotate	0.956835	0.917241	0.936620
	Cropping	0.985507	0.937931	0.961131
	Noise Injection	0.977941	0.917241	0.946619

Two of the five minority classes achieved their best F1-score in an augmented scenario rather than in the no-augmentation baseline: bacterial_leaf_streak reached its highest F1-score (0.9821) under rotate, and brown_spot reached its highest F1-score (0.9611) under cropping. Conversely, bacterial_panicle_blight, bacterial_leaf_blight, and downy_mildew all obtained lower F1-scores in every augmented scenario compared with no-augmentation, indicating that the benefit of augmentation is class-dependent rather than uniform across minority classes.

4.4.4 Confusion Matrix Analysis

Confusion matrices were generated for each scenario to evaluate whether the model produced more false-positive or false-negative errors across the ten disease classes. A detailed evaluation of the prediction error distribution for each augmentation scenario is presented sequentially through Figure 7 to Figure 11 below.

4.4.4.1 Test Results No Augmentation Scenario

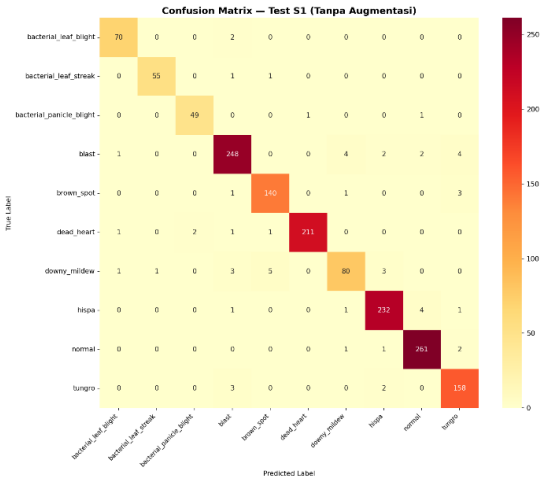

Figure 7. Confusion Matrix No Augmentation Graph

As shown in Figure 7, the baseline model achieves high classification accuracy, evidenced by the high values concentrated along the main diagonal. The highest true-positive counts are recorded in 'normal' (261), 'blast' (248), and 'hispa' (232), indicating strong feature recognition for these dominant classes. Conversely, minor misclassifications are reflected in the low off-diagonal numbers, such as 9 cases where 'bacterial_leaf_streak' is misclassified as 'bacterial_leaf_blight'. Overall, these scattered false-positive and false-negative instances mostly remain in single digits, demonstrating low error rates across the ten disease categories.

4.4.4.2 Test Results Flipping Augmentation Scenario

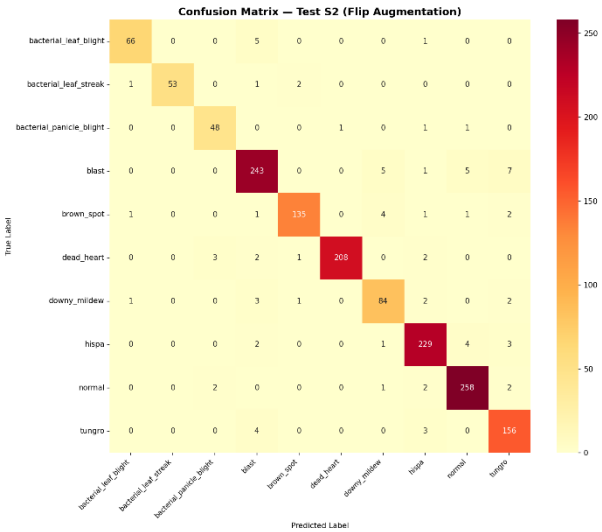

Figure 8. Confusion Matrix Flipping Augmentation Graph

As illustrated in Figure 8, the model maintains high true-positive counts across the primary categories, with 'normal' (258), 'blast' (243), and 'hispa' (229) showing the strongest recognition rates. The introduction of flip augmentation shows stable classification boundaries, although minor visual overlaps persist. For instance, the model still records 9 misclassified cases where 'bacterial_leaf_streak' is predicted as 'bacterial_panicle_blight'. Most off-diagonal false-positive and false-negative values remain consistently low, primarily localized in single digits across the remaining disease classes.

Journal of Adaptive Intelligent Systems

4.4.4.3 Test Results Rotate Augmentation Scenario

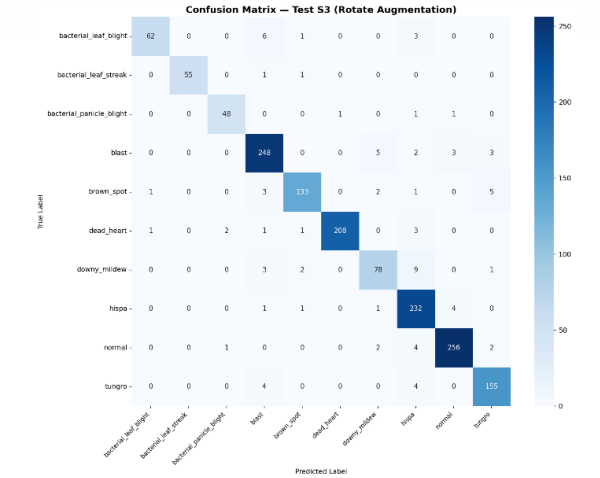

Figure 9. Confusion Matrix Rotate Augmentation Graph

As illustrated in Figure 9, the model trained with rotation techniques maintains steady classification performance, led by 'normal' (256), 'blast' (248), and 'hispa' (232) along the main diagonal. The orientation adjustments yield highly localized errors, though specific classification ambiguities persist. Notably, there are still 9 instances where 'bacterial_leaf_streak' is incorrectly predicted as 'bacterial_panicle_blight'. Across the remaining categories, the off-diagonal false-positive and false-negative instances are minor and remain securely within single digits.

4.4.4.4 Test Results Cropping Augmentation Scenario

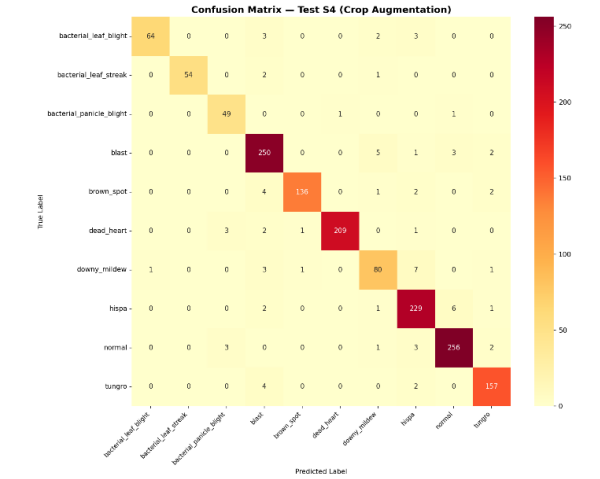

Figure 10. Confusion Matrix Cropping Augmentation Graph

Examination of Figure 10 confirms high classification efficiency, as demonstrated by the prominent distribution of correct predictions along the primary diagonal. The model proves highly capable of recognizing classes like 'normal' (256), 'blast' (250), and 'hispa' (229) despite the localized image clipping. Concurrently, error rates across the off-diagonal cells are tightly constrained. The most noticeable deviation involves 9 samples of 'bacterial_leaf_streak' being mistakenly classified as 'bacterial_panicle_blight'. Beyond this specific visual conflict, other false-positive and false-negative instances remain negligible and stay strictly within single-digit values.

4.4.4.5 Test Results Noise Injection Augmentation Scenario

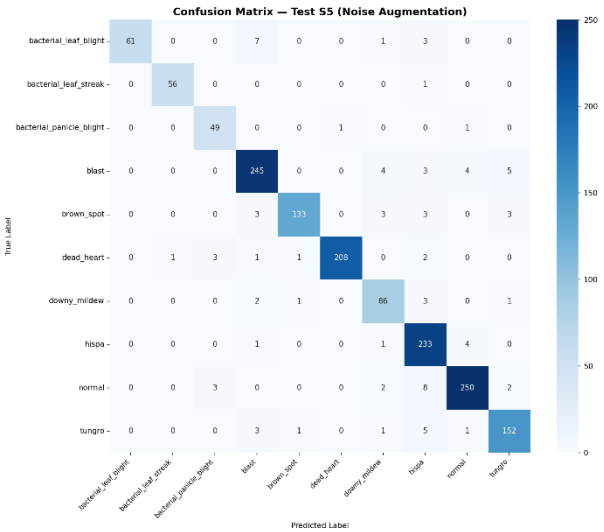

Figure 11. Confusion Matrix Noise Injection Augmentation Graph

Observation of Figure 11 highlights a steadfast recognition performance, demonstrated by the dense distribution of high values tracing the diagonal path. Despite the added pixel-level distortion, the network successfully preserves high classification accuracy in dominant classes, led by 'normal' (250), 'blast' (245), and 'hispa' (233). Concurrently, the off-diagonal cells register low false-positive and false-negative tallies, showing that the model maintains tight control over error propagation. The most prominent visual confusion involves 9 instances of 'bacterial_leaf_streak' being mistakenly assigned to 'bacterial_panicle_blight'. Apart from this persistent ambiguity, all remaining misclassified samples are

Journal of Adaptive Intelligent Systems

sparsely distributed and safely confined to minimal single-digit figures.

Table 10 summarizes the total number of correct and incorrect predictions out of 1,562 test samples for each scenario.

Table 10. Correct and Incorrect Predictions per Scenario

Scenario	Correct Predictions	Incorrect Predictions	Error Rate
No Augmentation	1,504	57	3.65%
Flipping	1,480	82	5.25%
Rotate	1,475	87	5.57%
Cropping	1,484	78	4.99%
Noise Injection	1,473	89	5.70%

Across all scenarios, the diagonal pattern dominates each confusion matrix, confirming that the majority of samples were classified correctly regardless of the augmentation condition. The most frequent misclassification patterns were consistent across scenarios: downy_mildew was frequently confused with hispa, blast, or brown_spot; blast was frequently confused with tungro or downy_mildew; and hispa was occasionally confused with the normal class. These recurring error patterns indicate that the confusion primarily occurs among classes with visually overlapping symptoms rather than being introduced by a specific augmentation technique.

4.5.1 Effect of Augmentation on Model Performance and Minority Classes

During validation, all five scenarios achieved comparable macro F1-scores (differences below 1%), indicating that the model was able to learn the artificially manipulated images effectively when evaluated within a uniform data distribution. However, this equivalence did not persist at the test stage: the no-augmentation scenario retained the highest macro F1-score (0.9588), while all augmented scenarios showed a parallel decline—cropping to 0.9447 (a 1.41% gap), flipping to 0.9427 (1.61%), noise injection to 0.9404 (1.84%), and rotate to 0.9394 (1.94%, the largest gap). This pattern indicates that computer-generated variation does not always align with the natural variation present in real-world data, which can reduce the model's generalization capability. Among the augmentation techniques, cropping consistently showed the best stability, while rotate consistently showed the weakest generalization—most likely because rotation disrupts the spatial orientation cues that the ViT model relies on to recognize disease symptoms relative to the leaf-vein structure.

5. Conclusion

This study demonstrates that image augmentation (flipping, rotation, cropping, noise injection) does not consistently improve Vision Transformer model performance, as the no-augmentation scenario achieved a best accuracy of 96.29% and a best macro F1-score of 0.9588 macro. This performance decline is caused by data distribution mismatch and the loss of critical spatial information, with impacts varying across classes depending on the distinctiveness of each disease's visual features. In conclusion, the effectiveness of augmentation is not universal, its application must be tailored specifically to the visual characteristics of the data and cannot be applied uniformly to address class imbalance.

4.5.2 Effect on Minority-Class Performance

The effect of augmentation on minority classes varied depending on each class's visual characteristics. For bacterial_panicle_blight, none of the augmentation techniques improved upon the no-augmentation F1-score (0.9608); cropping and noise injection in particular reduced precision substantially (to 0.8909) while recall remained stable, suggesting that these techniques made the model overly sensitive and prone to misclassifying other classes as bacterial_panicle_blight. In contrast, bacterial_leaf_streak benefited from rotate and noise injection, both of which produced a higher F1-score than the no-augmentation baseline, supported by high precision and recall—suggesting that this elongated, line-shaped lesion pattern benefits from additional variation in angle and texture, which helps the self-attention mechanism in ViT isolate object edges more precisely.

The bacterial_leaf_blight class was the most negatively affected by augmentation: F1-score dropped from 0.9655 (no augmentation) to as low as 0.9118 under rotate and 0.9173 under noise injection, driven mainly by a drop in recall (to 0.8611 and 0.8472, respectively). This suggests that extreme rotation angles and pixel-level noise obscure the primary spatial features the model relies on for recognition. The downy_mildew class, which had the lowest baseline F1-score (0.8889) among the minority classes, showed the most varied response:

Journal of Adaptive Intelligent Systems

flipping and noise injection increased recall (to 0.9032 and 0.9247, respectively) at the cost of precision, while rotate further reduced its F1-score to 0.8619—indicating high sensitivity to changes in vertical or horizontal orientation. The brown_spot class benefited most clearly from cropping, which increased its F1-score to 0.9611 (above the no-augmentation baseline of 0.9589), supported by the highest precision among all scenarios (0.9855); because brown_spot lesions appear as small, scattered round spots, removing background information through cropping likely helps the model focus more precisely on the relevant lesion area.

Overall, these class-level results confirm that augmentation does not have a uniform, linear effect on ViT performance: each class's visual geometry and pixel structure determines its sensitivity to a given image manipulation technique. While some techniques (e.g., rotate, noise injection) expanded detection coverage for specific classes, this improvement in recall often came at the expense of precision due to blurred spatial feature boundaries.

4.5.3 Overall Effectiveness of Image Augmentation

The no-augmentation scenario achieved the best balance between precision and recall for several classes, including bacterial_panicle_blight (F1 = 0.9608) and bacterial_leaf_blight (F1 = 0.9655), indicating that the visual features learned from the original, unaltered data already closely represent the true object characteristics; introducing less-compatible synthetic variation in these cases risks reducing rather than improving performance.

Nevertheless, when data manipulation is genuinely necessary—for example, to address a shortage of images in classes with substantial data imbalance—cropping stands out as the most viable augmentation choice, producing performance closest to the no-augmentation baseline (a 1.41% gap in macro F1-score) with the best stability among the augmented scenarios (a 1.39% validation-to-test drop). This technique performs well because it removes irrelevant background information, allowing the model to focus directly on the main disease object without distorting its shape. Noise injection also shows some practical value, improving recall for classes such as bacterial_leaf_streak and downy_mildew, although at the cost of reduced precision.

Conflicts of Interest

The author declare that there are no conflicts of interest related to this study

Author Contribution

Conceptualization, Nasywa Alya Musyaffa' and Tamam Asrori; methodology, Nasywa Alya Musyaffa' and Tamam Asrori; validation, Dyah Ariyanti; formal analysis, Nasywa Alya Musyaffa'; data curation, Tamam Asrori and Ira Aprilia; writing-original draft preparation, Nasywa Alya Musyaffa'; writing-review and editing, Nasywa Alya Musyaffa', Tamam Asrori, and Dyah Ariyanti; visualization, Nasywa Alya Musyaffa'; supervision, Tamam Asrori and Dwi Putri Kartini; project administration, Dyah Ariyanti; funding acquisition, Tamam Asrori and Dyah Ariyanti.

References

[1] R. Fuadi and Istiqomah, "Produktivitas Padi sebagai Indikator Modernisasi Pertanian dan Kaitannya dengan Ketahanan Pangan di Jawa Tengah Tahun 2023," *J. Agrica*, vol. 18, no. 2, pp. 308–314, 2025, doi: 10.31289/agrica.v18i2.15779.

[2] BPS 2024, "Luas Panen dan Produksi Padi Di Indonesia 2024 (Angka Tetap)," *Badan Pus. Stat.*, vol. 8, no. 15, p. 52, 2025, [Online]. Available: https://www.bps.go.id

[3] T. Febriyanto and S. Syofian, "Implementasi Deep Learning Menggunakan Vision Transformer Untuk Klasifikasi Penyakit Daun Padi," *J. TIFDA (Technology Inf. Data Anal.*, vol. 1, no. 2, pp. 34–39, 2024, doi: 10.70491/tifda.v1i2.47.

[4] A. T. Adiana, Jumadi, and E. Nurlatifah, "Klasifikasi Penyakit Pada Daun Padi Menggunakan Model Hybrid CNN-SVM," *SMATIKA STIKI Inform. J.*, vol. 6, no. 1, pp. 258–267, 2025, doi: https://doi.org/10.32664/smatika.v15i02.1548.

[5] R. W. Taufiqurrahman, N. Yudistira, and A. W. Widodo, "Penerapan Model Vision Transformer Dalam Klasifikasi Hama Pada

Journal of Adaptive Intelligent Systems

Pertanian," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 10, pp. 2548–964, 2025, [Online]. Available: http://j-ptiik.ub.ac.id

[6] K. Han *et al.*, "A Survey on Vision Transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, 2023, doi: 10.1109/TPAMI.2022.3152247.

[7] M. Imani, "Imbalanced data classification using graph based transformation," *Sci. Rep.*, vol. 15, no. 1, pp. 1–14, 2025, doi: 10.1038/s41598-025-18804-2.

[8] M. S. Ardiyansyah, F. R. Umbara, and M. Melina, "Klasifikasi Kanker Payudara Berbasis Deep Learning Menggunakan Vision Transformer dengan Teknik Augmentasi Data Citra," *JURIKOM (Jurnal Ris. Komputer)*, vol. 12, no. 3, pp. 241–250, 2025, doi: 10.30865/jurikom.v12i3.8619.

[9] D. Mirwansyah, A. Solichin, R. Hardi, N. Wanti Wulan Sari, N. Arista Riski, and D. Aldo, "Augmentasi Citra Pohon Kelapa Sawit untuk Deteksi Objek Berbasis Deep Learning," *Metik J.*, vol. 9, pp. 195–202, 2025, doi: 10.47002/metik.v9i1.1001.

[10] H. Hairani and T. Widiyaningtyas, "Augmented Rice Plant Disease Detection with Convolutional Neural Networks," *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 8, no. 1, pp. 27–39, 2024, doi: 10.29407/intensif.v8i1.21168.

[11] Petchiammal, B. Kiruba, Murugan, and P. Arjunan, "Paddy Doctor: A Visual Image Dataset for Automated Paddy Disease Classification and Benchmarking," *ACM Int. Conf. Proceeding Ser.*, no. January 2023, pp. 203–207, 2023, doi: 10.1145/3570991.3570994.

[12] J. Al Amien, H. A. Ghani, N. I. Saleh, and Y. Fatma, "A comprehensive evaluation of multiclass imbalance techniques with ensemble models in IoT environments," vol. 22, no. 3, pp. 690–701, 2024, doi: 10.12928/TELKOMNIKA.v22i3.25887.