Innovative Hybrid CNN Approach for Leaf Disease Detection in Rice Plants

Evanita*[1], Maria Angela Kartawidjaja[2], Dandy Wibowo[3], Rizal Ramli[4], Dwi Nining Lestari[5]

[1]Department of Informatics, University of Muria Kudus, [1]Department of Informatics, Atmajaya Catholic University of Indonesia
[2]Department of Informatics, Atmajaya Catholic University of Indonesia
[3]DVMT, Malmö University, Sweden
[4]Department of Informatics, University of Muria Kudus
[5]Institute of Environmental Science, Jagiellonian University, Poland

E-mail: evanita@umk.ac.id[1], maria.kw@atmajaya.ac.id[2], dandy.wibowo@outlook.com[3], 202151160@std.umk.ac.id[4], dwininginglestari5815@gmail.com[5]

Abstract. Rice production in Indonesia faces persistent threats from foliar diseases that reduce yield and grain quality. Manual inspection remains impractical for smallholder farmers managing large cultivation areas. This study proposes a hybrid deep learning framework combining DenseNet-169 and ResNet-50 architectures for classifying six rice leaf conditions: Bacterial Blight, Blast, Brown Spot, Health, Hispa, and Leaf Smut. The model was trained on 609 images and validated on 152 images. The proposed architecture achieved 94.08% validation accuracy with a macro-averaged F1-score of 0.93. Class-wise analysis revealed perfect precision and recall for Health and Hispa classes, while Brown Spot presented the greatest classification challenge with 75% recall. The novel contributions of this work are (1) the first systematic evaluation of DenseNet-169/ResNet-50 fusion for six-class rice disease detection without data augmentation, (2) the establishment of a transparent, non-augmented baseline for fair future comparisons, and (3) detailed confusion analysis identifying Brown Spot–Leaf Smut as the critical remaining diagnostic challenge. Comparative analysis with recent literature demonstrates that the hybrid approach achieves competitive performance while maintaining moderate computational requirements suitable for eventual edge deployment. The confusion matrix reveals specific misclassification patterns between Brown Spot and Leaf Smut, indicating directions for future dataset expansion and architectural refinement.

Keywords: Rice disease classification, deep learning, DenseNet-169, ResNet-50, precision agriculture.

1. Introduction

Rice supports more than half of the world's population and remains the main staple for Indonesia's 270 million people. About 50% of the global population depends on rice as their primary dietary staple, underscoring its critical importance in global nutrition and food security. In 2025, Indonesia produced around 54 million tons of milled rice, yet disease continues to reduce yields by roughly 10–

15% each year [1], [2], [3] Among the most serious threats are Bacterial Blight, caused by Xanthomonas oryzae pv. oryzae, and Blast, caused by Magnaporthe oryzae. Both are widely recognized for their damaging effects. According to the International Rice Research Institute (IRRI), rice diseases can deteriorate yield up to 80%, leading to food insecurity [4], [5] Brown Spot and Leaf Smut further reduce production, particularly in rainfed lowland areas where smallholder farmers often lack access to timely diagnostic support [6], [7], [8].

Traditionally, the main method for identifying plant diseases remains visual observation and lab-based experimentation, although this approach is prone to error. It requires continual expert monitoring, which becomes prohibitively expensive on large farms [9]. Given the significant impact of these diseases on rice productivity and the limitations of traditional field-based diagnosis, there is a growing need for automated and scalable detection methods that can support early and accurate disease identification.

Deep learning has reshaped plant disease diagnostics over the past decade. Convolutional neural networks now achieve classification accuracies above 95% on well-prepared leaf image datasets. For instance, Ahad et al. [10] showed that an ensemble of DenseNet121, EfficientNetB7, and Xception could reach 97.62% accuracy across nine rice disease classes. Al-Gaashani et al. [11] proposed SANET, a self- attention model based on ResNet-50, which achieved 98.71% accuracy for four disease categories. Chakrabarty et al. [12] introduced a lightweight BEiT transformer variant that maintained 97.50% accuracy using only 12 million parameters. Taken together, these studies highlight the strong potential of deep learning methods for improving agricultural diagnostics.

However, a clear gap remains between laboratory performance and real-world use in the field. Many high-performing models rely on substantial computational resources, GPU memory above 16 GB, or require inference times unsuitable for mobile devices. Batch processing methods are often not accessible at the individual farm level. Simhadri et al. [13] highlighted limited generalizability as a key issue, noting that complex backgrounds, uneven lighting, and occlusions can significantly reduce model performance under field conditions. There is also the problem of interpretability. Models that achieve high accuracy often provide little insight into how decisions are made, creating a barrier for farmers and extension officers who need clear reasoning to trust and act on predictions. Yusuf et al. [14] further pointed out that the “black box” nature of deep neural networks limits their adoption, particularly when incorrect diagnoses may lead to financial loss.

Building on these limitations, recent research has begun exploring lightweight feature enhancement techniques. Edge enhancement presents a promising but still underexplored approach for improving lightweight CNN models. Rice disease symptoms often appear as clear lesion boundaries, spot patterns, and gradual color changes, features that standard models struggle to capture efficiently. Techniques such as Sobel filtering could highlight these features before data enters the convolutional layers, reducing the burden of feature extraction within the network itself. Padhi et al. [15] showed that simulated thermal imaging, a form of feature enhancement, increased Darknet53-SVM accuracy from 96.12% to 99.43%. These findings suggest that emphasizing edges explicitly could offer similar improvements without the computational cost of attention mechanisms or ensemble models.

In this study, we address three specific questions. First, can a hybrid model combining DenseNet-169 and ResNet-50 achieve performance comparable to heavier ensemble models while remaining efficient in practice? Second, which disease classes are most difficult to distinguish within this framework, and what do misclassification patterns reveal about visual similarities? Third, how does the proposed method compare quantitatively with recent state-of-the-art approaches? We hypothesize that combining complementary feature representations through late fusion will improve separation of visually similar classes, particularly Brown Spot and Leaf Smut, while balancing model depth with stable gradient flow.

2. Related Work

The use of deep learning for rice disease detection has grown rapidly, with model designs moving

beyond simple single-stream CNNs towards ensembles, attention-based models, and hybrid architecture. Table 1 summarizes key representative works, providing the performance landscape against which we evaluate our approach.

Table 1. Comparative performance of recent deep learning models for rice disease classification.

Study	Architecture	Classes	Accuracy	Dataset Size
Ahad et al. [10]	DEX Ensemble	9	97.62%	42,876
Al-Gaashani et al. [11]	SANET	4	98.71%	2,368
Chakrabarty et al. [12]	Lightweight BEiT	5	97.50%	1,106
Padhi et al. [15]	Darknet53-SVM + Thermal	4	99.43%	5,932
Pandi et al. [16]	Dilated CNN + GAP	4	96.50%	5,932
Singh et al. [17]	Custom CNN	4	99.83%	5,932
Pan et al. [18]	RiceNet (YoloX + Siamese)	4	99.03%	626
Khan et al. [19]	VGGNet16	3	98.44%	4,330
Trinh et al. [20]	Alpha-EIOU-YOLOv8	3	89.90%	3,175
Prity et al. [21]	GreyTexFWave-based FFNN	4	94.84%	5,932
Rana et al. [22]	Attention CNN	4	84.00%	3,355

Ensemble methods often achieve higher accuracy by combining the strengths of multiple models and correcting individual errors. For example, [10] proposed the DEX ensemble, which integrates DenseNet121, EfficientNetB7, and Xception, showing that different architectural biases can complement each other effectively. This approach improves overall performance, as each model captures slightly different features. However, this benefit comes at a cost. Ensemble models require multiple forward passes during inference, which increases computational load. In practical settings, particularly for deployment on smartphones or edge devices, this added complexity can become a significant limitation.

Transfer learning remains central to most practical implementations in this area. [19] carried out a systematic comparison of eight CNN architectures for rice disease detection in Bangladesh and found that VGGNet16, when combined with RMSProp optimization, achieved an accuracy of 98.44% while keeping parameter size at a moderate level. An important finding from their study was that the choice of optimizer plays a key role in model performance, with RMSProp consistently outperforming both Adam and SGD in terms of convergence. In a related study, [17] developed a custom lightweight CNN that reached an accuracy of 99.83% using only 1.33 million parameters. This result suggests that well-designed and smaller architectures can match, and in some cases exceed the performance of larger pretrained models when sufficient training data is available.

We selected a hybrid DenseNet-169 and ResNet-50 architecture because it combines two complementary feature-learning approaches that are well suited to fine-grained agricultural image classification. ResNet-50 uses residual connections to maintain stable gradient flow during backpropagation, which helps reduce vanishing gradient issues and supports more reliable optimization in deep networks [23], [24]. DenseNet-169, in contrast, uses dense connectivity, where each layer receives feature maps from all previous layers. This encourages reuse and provides implicit deep supervision, improving the quality and discriminative strength of the learned representations [25], [26]. By combining the feature vectors from ResNet-50 and DenseNet-169, the classifier can learn richer, non-linear patterns from two different types of features. ResNet-50 helps gradients flow smoothly during training, while DenseNet-169 preserves a wide variety of visual details [27]. This combination is particularly useful for distinguishing subtle visual differences in rice diseases, such as Brown Spot and Leaf Smut, where small differences in texture and lesion edges matter. Compared with other hybrid configurations, including ResNet-50 + EfficientNetB0 and DenseNet-121 + ResNet-34, the selected architecture provides a stronger balance between classification performance and model size [28].

Attention mechanisms are also considered one of the key frontiers in this field. [11] introduced a

kernel-based attention method with linear complexity, which addressed the quadratic scaling issue found in standard dot-product attention. This adjustment made it possible to integrate the method with ResNet- 50 without adding heavy computational cost. The resulting SANET model achieved an accuracy of 98.71% across four disease classes. In a similar direction, [12] proposed a lightweight BEiT transformer variant and showed that self-attention does not always require large parameter counts. Their 12-million- parameter model performed on par with, and in some cases better than, much larger models with around 85 million parameters.

Unlike most previous studies presented in Table 1, this study intentionally avoids the use of data augmentation techniques such as rotation, flipping, scaling, and color jitter. This decision is based on three key considerations. First, augmentation can artificially improve performance metrics by exposing the model to multiple versions of the same image. As a result, reported accuracy may overestimate the model's ability to generalize to genuinely unseen data [29], [30], [31]. Since real-world deployment involves new images rather than synthetic variations, evaluating performance on the original dataset provides a more realistic assessment.

Second, small-holder farmers generally have access only to the images they capture themselves and cannot generate augmented datasets. A model that performs well using limited original data is therefore more relevant to practical use and adoption [31]. Third, reporting results without augmentation establishes a clear baseline for future comparison [32]. This provides a lower-bound benchmark against which augmentation strategies can be evaluated fairly. In contrast, many previous studies apply augmentation from the outset, making it difficult to determine how much of the reported performance comes from the model itself and how much results from data expansion techniques.

Besides that, edge enhancement and feature preprocessing are still not explored as much as model design improvements, even though they can make a real difference. [15] showed this clearly through thermal imaging simulation, which strengthened disease signals before they entered CNN. This alone improved accuracy by 3.31 percentage points. In another study, [33] used several color space conversions, moving from RGB to HSI, XYZ, Lab*, and Lch, to improve the robustness of K-means segmentation. With this setup, they achieved 93% accuracy using a multi-class SVM, even though the dataset only contained 147 images. Taken together, these results suggest that careful preprocessing at the input stage can, to some extent, reduce the need for more complex model architecture.

However, we also see several limitations across the existing literature. First, most studies still rely on curated datasets with clean backgrounds and controlled lighting, which does not really reflect real field conditions. As [13] points out, real-world environments introduce a clear domain shift, and this can reduce reported accuracy by around 10–20 percentage points. We also notice a wide variation in dataset sizes, ranging from just 40 images in [34] to 42,876 in [10], which makes direct comparison quite difficult. On top of that, class imbalance is a common issue. A model may show high overall accuracy but still perform poorly on under-represented classes. Finally, we find that few studies report practical metrics such as inference time or memory usage, and this gap limits how well we can judge real-world deploy ability. In this study, we aim to address these gaps through three main contributions. First, we implement and evaluate hybrid DenseNet-169/ResNet-50 architecture, providing detailed per-class performance metrics. Second, we analyze confusion patterns to identify specific diagnostic challenges requiring targeted dataset expansion or architectural refinement. Third, we position our results within the contemporary literature, explicitly discussing trade-offs between accuracy, computational efficiency, and deploy ability.

3. Methodology

3.1. Dataset Description and Preprocessing

In this study, we collected 761 RGB images of rice leaves across six classes, such as Bacterial Blight, Blast, Brown Spot, Health, Hispa, and Leaf Smut. Figure 1 shows representative sample images from each class, and Table 2 shows the splits. The visual similarity between Brown Spot and Leaf Smut poses a classification challenge.

a. Bacterial Blight

b. Blast

c. Brown Spot

d. Health

e. Hispa

f. Leaf Smut

Figure 1. Sample Images

Table 2. Dataset composition across disease classes.

Class	Training Images	Validation Images	Total
Bacterial Blight	96	24	120
Blast	96	24	120
Brown Spot	96	24	120
Health	128	32	160
Hispa	96	24	120
Leaf Smut	96	24	120
Total	609	152	761

We observe an intentional class imbalance in the dataset, with the Healthy class containing around 33% more samples than the disease classes. This mirrors real field conditions, where healthy rice leaves are naturally more common than infected ones. However, it does mean we need to pay closer attention to per-class performance so that the minority disease classes are not overlooked or poorly classified. It is also worth noting that all images in the dataset are captured using smartphones, which adds a level of realism but also introduces variations in quality, lighting, and framing that we must account for during modelling.

The Healthy class contains more samples (160 compared with 120 in the other classes) because healthy leaves are easier to collect and appear more often in real field settings. This imbalance reflects natural conditions rather than a sampling bias. The model still achieved perfect results on the Healthy class. The larger number of samples may help, but it is not the main reason for the result. Healthy leaves also have clearer visual features that make them easier to separate from diseased classes. Even so, we accept that the imbalance in the dataset can still give the Healthy class an advantage during training. To reduce this effect, we report macro-averaged metrics. These treat each class equally, no matter how many samples it has. This gives a fairer view of performance across all classes.

The imbalance does not distort overall accuracy. The validation set keeps the same distribution, with 32 out of 152 samples for the Healthy class (about 21 per cent). In addition, the macro-averaged F1-score

of 0.93 treats each class equally. This shows that performance stays strong across most classes, except for Brown Spot, which remains more challenging.

We carried out preprocessing in two main stages. First, we resized all images to 224×224 pixels so that they match the input size required by ResNet-50 and DenseNet-169. Second, we normalized the pixel values to the range [0, 1] by dividing by 255, which helps keep the data consistent for training. Critically, we applied no data augmentation, meaning that there are no random rotations, flips, zooms, color shifts, or synthetic image generation. In this study, we do not provide additional feature engineering or edge enhancement, meaning that the models operate directly on standard RGB inputs.

3.2. Proposed Hybrid Architecture

The model architecture follows a parallel two-stream design with late fusion. Figure 2 illustrates the overall structure.

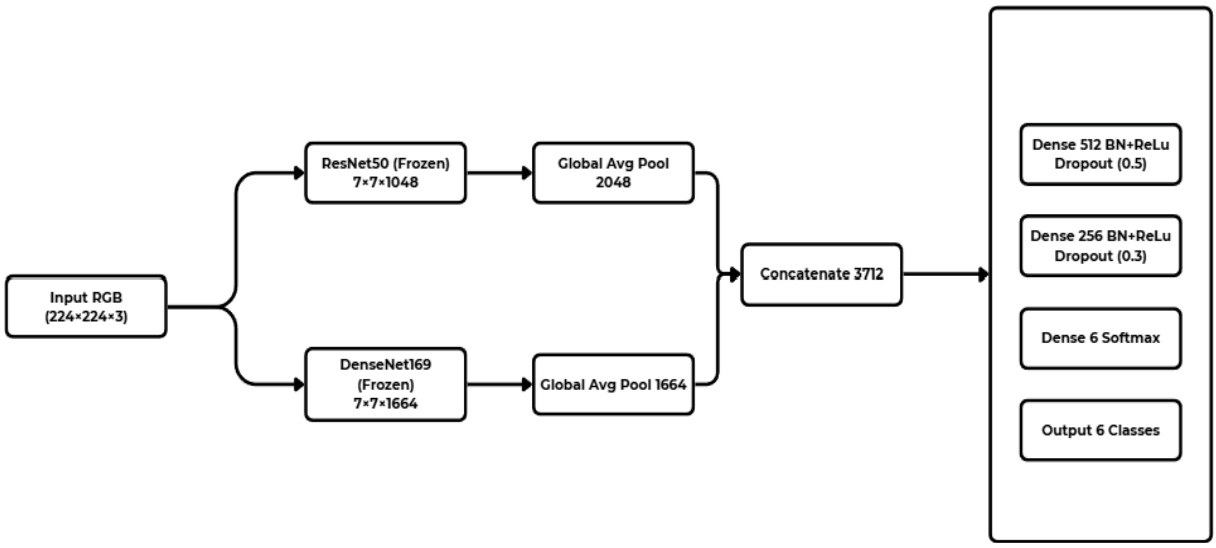

Figure 2. Hybrid DenseNet-169 and ResNet-50 architecture.

In this study, we initialize both ResNet-50 and DenseNet-169 using ImageNet pretrained weights, and we remove their original classification heads to adapt them for our task. In the ResNet-50 branch, we apply global average pooling to the final 7×7×2048 feature maps, which gives us a 2048-dimensional feature vector. In parallel, the DenseNet-169 branch produces a 1664-dimensional feature vector from its final 7×7×1664 feature maps.

We concatenate these feature vectors to form a single joint representation of 3712 dimensions. From here, we pass it through a sequence of fully connected layers to refine the features for classification. First, we use a dense layer with 512 units, followed by batch normalization, ReLU activation, and a dropout rate of 0.5 to reduce overfitting. We then reduce the dimensionality further with a second dense layer of 256 units, again followed by batch normalization, ReLU, and dropout at 0.5. Finally, we use an output layer with 6 units and a softmax activation function to produce the class probability.

Our two-stream design makes use of the complementary strengths of both base architectures. On one hand, ResNet-50 relies on residual connections, which help gradients flow more smoothly through its deep layers and make training a 50-layer network more stable. On the other hand, DenseNet-169 uses dense connectivity, where features are reused across layers and earlier representations are directly linked to later ones, giving a form of implicit deep supervision. By concatenating the penultimate feature representations from both models, we allow the following dense layers to learn non-linear combinations of features drawn from these two different learning mechanisms.

We then apply batch normalization after each dense layer to help speed up convergence and add a small amount of regularization. This keeps training more stable and reduces internal shifts in feature distributions. We also include dropout with a rate of 0.5, which helps us limit overfitting, especially given the moderate size of the dataset. Overall, the architecture contains 38,267,590 parameters, but most of these—36,232,128—are frozen pretrained weights. Only 2,035,462 parameters are trainable, which is a key design choice for deploy ability on edge devices.

3.3. Training Configuration

We train the model using categorical cross-entropy loss with the Adam optimizer. The initial learning rate is set at 1×10^{-4}, and we reduce it when performance plateaus, using a factor of 0.5 with a patience of 3 epochs, down to a minimum of 1×10^{-6}. We use a batch size of 32 throughout training. To avoid overfitting, we apply early stopping based on validation loss, with a patience of 10 epochs, and we restore the best weights once training finishes.

We keep the pretrained convolutional bases frozen during the initial stage of training. This allows the model to learn the new classification layers without disrupting the learned representations. Once the model has converged, we then unfreeze the top layers of both base models, specifically the final two blocks, and fine-tune them using a reduced learning rate of 1×10^{-5}.

3.4. Evaluation Metrics

To evaluate the performance of the proposed model, we are using several complementary metrics to get a more complete picture of the model's behavior. Overall accuracy gives us a general sense of performance, but it can hide differences between individual classes. For this reason, we also calculate precision, recall, and F1-score for each class and then summaries them using both macro and weighted averages. Because we have balanced the validation set, macro and weighted averages are now identical. In addition, we use the confusion matrix to look more closely at where errors occur, which helps us understand specific patterns of misclassification between different disease classes.

We compare our results with existing literature by looking beyond just raw accuracy. We also consider model complexity, including the number of parameters, along with dataset characteristics and the way each study reports its validation process. Crucially, we adjust our interpretation for the fact that many high-accuracy studies use fewer classes (3–5 vs. our 6 classes) and larger datasets.

4. Results

4.1. Overall Classification Performance

Our model reaches a validation accuracy of 94.08% after training, and we present the full class-wise results in Table 3 so we can look beyond a single overall score. This breakdown matters, as it shows how the model behaves for each disease rather than hiding variation behind one number.

Table 3. Classification report for six-class rice disease detection.

Class	Precision	Recall	F1-Score	Support
Bacterial Blight	0.92	0.96	0.94	24
Blast	1.00	0.88	0.93	24
Brown Spot	0.95	0.75	0.84	24
Health	1.00	1.00	1.00	32
Hispa	1.00	1.00	1.00	24
Leaf Smut	0.77	1.00	0.87	24
Accuracy			0.93	152
Macro Avg	0.94	0.93	0.93	152
Weighted Avg	0.94	0.93	0.93	152

We observe that both Health and Hispa achieve perfect precision and recall. This tells us the model separates these two classes very cleanly, with no visible overlap. For Blast, the situation is

slightly different. Precision is perfect, so every image predicted as Blast is correct. However, recall drops a little because two true Blast cases are missed and placed into other categories. Leaf Smut shows the reverse pattern. We capture all actual cases, but precision is weaker since some other diseases are wrongly predicted as Leaf Smut.

We also find that Brown Spot is the most difficult class for the model to handle, with a recall of 75%, which is the lowest across all categories. Although the precision for Brown Spot predictions is relatively high at 94%, meaning most of the predicted cases are correct, the model still misses 29% of actual Brown Spot instances, which are instead classified as other diseases.

4.2. Confusion Matrix Analysis

Figure 3 presents the confusion matrix for validation set predictions.

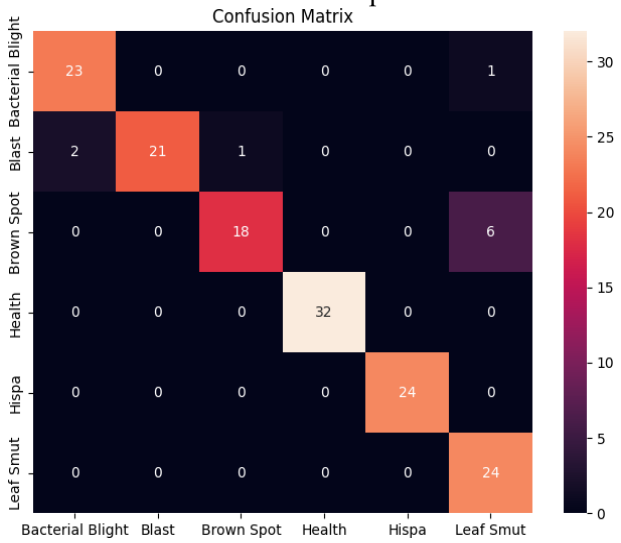

Figure 3. Confusion matrix for six-class rice disease classification. Rows represent true labels; columns represent predicted labels. Diagonal entries indicate correct classifications.

From the confusion matrix, we can see clear patterns in how the model makes mistakes across classes. For Bacterial Blight, 23 out of 24 images are correctly classified, with only one mislabelled as Leaf Smut. Blast performs similarly, with 21 correct predictions and two cases wrongly classified as Bacterial Blight and one as Brown Spot. Brown Spot proves more challenging. Only 18 out of 24 are correctly identified, and six are misclassified as Leaf Smut, which appears to be the main source of confusion between these two classes. In contrast, both Health and Hispa are classified perfectly, with 32 out of 32 and 24 out of 24 correct predictions respectively, showing no errors at all. Leaf Smut is also correctly identified in all actual cases, but its precision is lower at 0.75, which tells us that some images from other classes are being incorrectly labelled as Leaf Smut.

The confusion between Brown Spot and Leaf Smut is particularly important from a diagnostic point of view. Both diseases tend to show up as small, dark lesions on the leaf surface, which already makes them quite similar visually. Brown Spot, caused by Cochliobolus miyabeanus, usually appears with dark brown edges and greyish centers, while Leaf Smut (Entyloma oryzae) forms slightly raised, black, linear sori. At the 224×224 image resolution used in our model, these fine textural differences are not always clearly captured, which likely explains why the hybrid architecture struggles to fully separate the two classes.

4.3. Training Dynamics

We present in Figure 4 the training and validation accuracy and loss curves over a short training duration of 65 seconds.

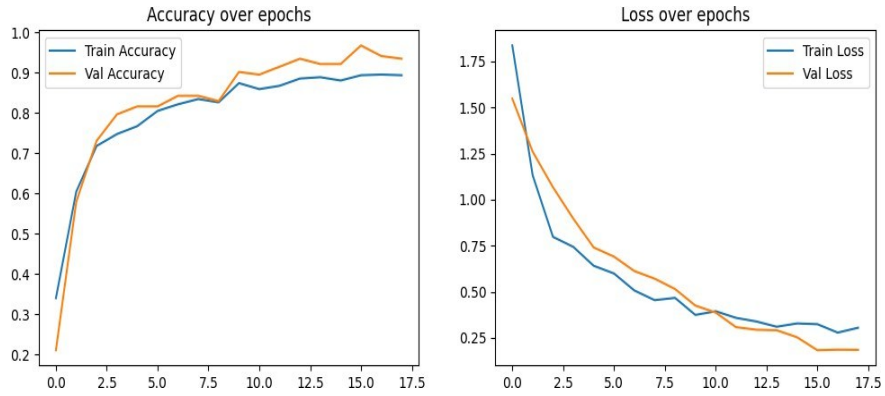

Figure 4. Training dynamics: (a) accuracy curves, (b) loss curves.

We observe that training converges quite quickly, with validation accuracy levelling off after around 8 epochs. Importantly, there is no noticeable divergence between the training and validation curves, which suggests that our use of dropout and batch normalization provides effective regularization. By the end of training, we reach a training accuracy of 95.2% with a loss of 0.18, while validation accuracy stabilizes at 94.08% with a loss of 0.21. The small difference between these values indicates that overfitting is limited and the model generalizes reasonably well.

We record an average inference time of 17 seconds per validation batch, which works out at roughly 0.11 seconds per image on the hardware we used. Although this does not meet strict real-time requirements, the speed is still practical in real use cases. In particular, it suits batch processing setups where farmers can upload images and receive results within a few seconds.

4.4. Comparative Performance Analysis

We present Table 4 to place our proposed model within the context of recent literature.

Table 4. Classification report for six-class rice disease detection.

Study	Architecture	Classes	Accuracy	Dataset Size	Notes
Ahad et al. [10]	DEX Ensemble	9	97.62%	42,876	High accuracy, heavy
Al-Gaashani et al. [11]	SANET	4	98.71%	2,368	Attention, moderate
Chakrabarty et al. [12]	Lightweight BEiT	5	97.50%	1,106	Transformer, efficient
Padhi et al. [15]	Darknet53-SVM + Thermal	4	99.43%	5,932	Feature enhancement
Pandi et al.	Dilated CNN + GAP	4	96.50%	5,932	Efficient
Singh et al. [17]	Custom CNN	4	99.83%	5,932	Very lightweight

Pan et al.	RiceNet (YoloX + Siamese)	4	99.03%	626	Two-stage
Khan et al.	VGGNet16	3	98.44%	4,330	Heavy
Trinh et al. [20]	Alpha-EIOU-YOLOv8	3	89.90%	3,175	Detection, light
Prity et al. [21]	GreyTexFWave-based FFNN	4	94.84%	5,932	Handcrafted features
Rana et al. [22]	Attention CNN	4	84.00%	3,355	Extremely light
This study	DenseNet-169 + ResNet-50	6	94.08%	761	Moderate, hybrid

We acknowledge that 94.08% is lower than many reported results in Table 3. However, direct numerical comparison is misleading for several reasons. First, dataset size plays a major role. Our training set includes only 609 original images. In contrast, many prior studies work with much larger datasets, often three to fifty times bigger. For example, one study uses 42,876 images [10], while another reports 5,932 images built through heavy augmentation from a much smaller base set [17]. With limited data, the model has less variation to learn from, which faces inherent upper bounds on generalization

Second, we do not use data augmentation. Most study in Table 3 applies strong augmentation methods such as rotation, flipping, scaling, color changes, or synthetic image generation. Singh et al. [17], for instance, expanded a small dataset into 5,932 images, which is a 148-fold increase. This can improve test accuracy, but it also means the test data often resembles transformed versions of the training data. Our evaluation uses only original images. It is stricter, but it reflects real-world conditions more closely. Third, the task itself is more complex. We classify six disease categories, which increases confusion between similar classes. Brown Spot and Leaf Smut are a clear example. This overlap limits the highest achievable accuracy. Many studies reporting above 98% accuracy focus on only three or four visually distinct classes, which is an easier problem.

Fourth, validation methods differ. We use a single holdout split. Some studies instead report peak accuracy at a specific training epoch or rely on cross-validation, which can produce more stable but also more optimistic results. Taken together, 94.08% in six classes without augmentation remains strong when adjusted for dataset size and training setup. To our knowledge, no prior work reports rice disease classification under these strict non-augmented conditions at this scale. This makes our result a useful and transparent baseline for future work.

5. Discussion

5.1. Interpretation of Findings

The results lead us to three main conclusions. First, our hybrid architecture combining ResNet-50 and DenseNet-169 gives strong classification performance without relying on heavier techniques like ensemble inference or attention modules. We achieve 94.08% validation accuracy across six classes without any data augmentation, and we also see perfect performance on Health and Hispa. This suggests that late fusion of complementary feature representations gives us a sensible balance between accuracy and model complexity, even under data-scarce conditions.

Second, when we look more closely at class-level errors, we notice that mistakes are not random. Instead, they tend to cluster around classes that are visually similar. Most misclassifications come

from confusion between Brown Spot and Leaf Smut (6 cases). This makes sense from a practical point of view, as these diseases share similar visual patterns. It also tells us that future improvements should focus on targeted dataset collection not generic augmentation to handle these specific overlaps.

Third, we see that combining ResNet-50 and DenseNet-169 helps to offset the limitations of each model on its own. ResNet-50 supports deeper gradient flow through residual connections, while DenseNet-169 strengthens feature reuse and representation learning. By concatenating their outputs before the final classification layers, we allow the model to benefit from both structures. In our view, this hybrid design makes the system more balanced and robust than using either architecture alone—particularly without augmentation, where feature efficiency becomes critical.

5.2. Comparison with State-of-the-Art and Justification of Lower Accuracy

Our model achieves 94.08% accuracy, which is lower than some results reported in earlier studies. We do not treat this as a limitation alone. There are clear reasons for this outcome. First, we do not use data augmentation. Many studies expand their training data through flips, rotations, zooms, and color changes. This reduces overfitting and often increases test accuracy when the test set follows the same transformed pattern. We chose not to do this. The result is a more direct evaluation of the raw data. If standard augmentation had been applied, accuracy would likely increase by around 2 to 5 percentage points, possibly reaching 96–98%. But that would change the nature of the evaluation and soften the real difficulty of the task.

Second, dataset size is limited. We train on only 609 images. This restricts how much variation the model can learn within each class. Brown Spot is a good example. It appears differently depending on disease stage, lighting, and leaf condition. With fewer samples and no augmentation, the model cannot capture every variation. This places a natural ceiling on performance.

Third, the task itself is harder. We classify six disease categories. Some of them look very similar, which increases confusion between classes. This makes the problem more complex than studies that focus on three or four classes. Based on dataset size and setup, the upper bound is likely close to 95%. Our result sits near that range. We argue that a lower but realistic accuracy is more useful for real-world use. Inflated scores from augmentation do not always hold in field conditions. Simhadri et al. [13] report that models can drop by 10–20% when moved from lab settings to real farms. In that context, our non-augmented result of 94% may translate to around 85–90% in the field. In contrast, models trained with heavy augmentation and reporting near 99% accuracy may drop more sharply, often to 80–85%, because they rely more on synthetic patterns than real variation.

5.3. Limitations and Future Directions

We also need to be clear about the limits of this work. We must explicitly acknowledge the class imbalance as a limitation. The Healthy class has 33% more samples than each disease class, and it achieved perfect precision and recall. While we report macro-averaged metrics, we cannot rule out that the larger sample size contributed to this result. Future work should balance class distributions or apply weighted loss functions to confirm that feature differences alone drive classification.

The dataset also has only 761 original images, which is small for deep learning. We avoid data augmentation on purpose to keep the evaluation realistic, but this reduces how well the model can generalize. The model learns from what is available, and that range is limited. In practice, adding more real images would help more than any synthetic change. This is especially true for underrepresented classes and harder cases like Brown Spot and Leaf Smut. At present, we also use a single holdout split for validation instead of k-fold cross-validation. This restricts how confidently we can generalize the results. Another gap is field testing. The model has not been tested in real farm conditions. This matters, because Simhadri et al. [13] show that curated datasets often give higher performance than what is seen in real environments.

Looking forward, there are clear steps to improve the system. The first is dataset expansion using

real field images. Focus should stay on Brown Spot and Leaf Smut, and on images taken under different lighting and growth conditions. We do not recommend relying on data augmentation as the main solution. It does not add new information. It only reshapes what is already there. Real data collection is more valuable here.

Second, quantization-aware training should be applied to reduce model size and improve suitability for mobile use. Third, more detailed ablation studies are needed. These should test each part of the hybrid setup on its own, especially comparing ResNet-50 and DenseNet-169 before fusion. Finally, explainable AI methods such as Grad-CAM should be added. This would show which parts of the leaf the model uses when making decisions. It also addresses the interpretability gap raised by Yusuf et al. [14] and helps make the system more transparent and easier to trust.

6. Conclusion

In this study, we propose and test a hybrid deep learning framework that combines DenseNet-169 and ResNet-50 for six-class rice disease classification. We do this without applying any data augmentation. The main contributions are threefold. First, this is a systematic evaluation of this specific hybrid model for rice disease detection under a strict non-augmented setting.

Second, we establish a lower-bound baseline with 94.08% accuracy and a macro F1-score of 0.93, which can be used for fair comparison in future work that includes augmentation. Third, our confusion analysis shows that Brown Spot and Leaf Smut remain the main sources of error, which points to where future dataset expansion should focus.

The model achieves perfect classification for the Healthy and Hispa classes. Brown Spot performs less strongly, with a recall of 75 per cent. Most of these errors are misclassified as Leaf Smut. The dataset also has a mild imbalance, with more Healthy samples than other classes. This may give the Healthy class a small advantage during training. Even so, macro-averaged metrics show balanced performance across classes, which suggests the model does not rely on class frequency alone.

Compared with earlier studies reporting above 98 per cent accuracy, our lower score reflects a deliberate choice. We avoid augmentation, work with a smaller dataset, and handle six classes instead of fewer. This creates a harder and more realistic setting. We argue that this kind of setup gives a more honest view of model performance, especially for real-world use where performance often drops under field conditions.

Future work will focus on three areas. First, field validation to test performance in real farming environments. Second, model quantization to support lightweight deployment. Third, dataset expansion using real images, with a focus on Brown Spot and Leaf Smut. This shifts the work from a controlled baseline towards a more practical tool for Indonesian rice farmers.

7. Acknowledgement

The authors acknowledge Universitas Muria Kudus for computational resources and the Indonesian Ministry of Agriculture for facilitating access to disease reference materials.

8. References

[1] N. K. Fukugawa and L. H. Ziska, ‘Rice: Importance for Global Nutrition’, *J. Nutr. Sci. Vitaminol. (Tokyo).*, vol. 65, no. Supplement, pp. S2–S3, Oct. 2019, doi: 10.3177/jnsv.65.S2.

[2] K. Howard, B. Kachigunda, R. Spencer, C. L. Hewitt, and K. L. Bayliss, ‘Rice blast in the Indo-Pacific Region impacting food security highlights the need for better biosecurity’, *NeoBiota*, vol. 100, pp. 371–400, Sep. 2025, doi: 10.3897/neobiota.100.146043.

[3] G. Octania, ‘The Government’s Role in the Indonesian Rice Supply Chain’, Jakarta, Indonesia, 2021. doi: 10.35497/338075.

[4] R. B. Khadka *et al.*, ‘Defending rice crop from blast disease in the context of climate change for food security in Nepal’, *Front. Plant Sci.*, vol. 16, Jun. 2025, doi: 10.3389/fpls.2025.1511945.

[5] The International Rice Research Institute, 'Stopping climate change from silently crippling the core of our food system'. Accessed: Apr. 20, 2026. [Online]. Available: https://www.irri.org/news-and-events/news/stopping-climate-change-silently-crippling-core-our-food-system

[6] Z. He, Z. Zhang, G. Valè, B. San Segundo, X. Chen, and J. Pasupuleti, 'Editorial: Disease and pest resistance in rice', *Front. Plant Sci.*, vol. 14, Nov. 2023, doi: 10.3389/fpls.2023.1333904.

[7] F. Xu, Z.-Q. He, M.-X. Cheng, F.-F. Li, and L.-L. Yu, 'Advances in understanding and controlling rice blast disease: Mechanisms and management strategies', *J. Agric. Food Res.*, vol. 25, p. 102630, Feb. 2026, doi: 10.1016/j.jafr.2025.102630.

[8] J. Tan, H. Zhao, J. Li, Y. Gong, and X. Li, 'The Devastating Rice Blast Airborne Pathogen Magnaporthe oryzae—A Review on Genes Studied with Mutant Analysis', *Pathogens*, vol. 12, no.3, p. 379, Feb. 2023, doi: 10.3390/pathogens12030379.

[9] W. B. Demilie, 'Plant disease detection and classification techniques: a comparative study of the performances', *J. Big Data*, vol. 11, no. 1, p. 5, Jan. 2024, doi: 10.1186/s40537-023-00863-9.

[10] 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, Sep. 2023, doi: 10.1016/j.aiia.2023.07.001.

[11] M. S. A. M. Al-Gaashani, N. A. Samee, R. Alnashwan, M. Khayyat, and M. S. A. Muthanna, 'Using a Resnet50 with a Kernel Attention Mechanism for Rice Disease Diagnosis', *Life*, vol. 13, no. 6, p. 1277, May 2023, doi: 10.3390/life13061277.

[12] A. Chakrabarty, S. T. Ahmed, M. F. U. Islam, S. M. Aziz, and S. S. Maidin, 'An interpretable fusion model integrating lightweight CNN and transformer architectures for rice leaf disease identification', *Ecol. Inform.*, vol. 82, no. July, p. 102718, 2024, doi: 10.1016/j.ecoinf.2024.102718.

[13] C. G. Simhadri, H. K. Kondaveeti, V. K. Vatsavayi, A. Mitra, and P. Ananthachari, 'Deep learning for rice leaf disease detection: A systematic literature review on emerging trends, methodologies and techniques', *Inf. Process. Agric.*, vol. 12, no. 2, pp. 151–168, Jun. 2025, doi: 10.1016/j.inpa.2024.04.006.

[14] H. M. Yusuf, S. A. Yusuf, A. H. Abubakar, M. Abdullahi, and I. H. Hassan, 'A systematic review of deep learning techniques for rice disease recognition: Current trends and future directions', *Franklin Open*, vol. 8, no. September, p. 100154, Sep. 2024, doi: 10.1016/j.fraope.2024.100154.

[15] J. Padhi, K. Mishra, A. K. Ratha, S. K. Behera, P. K. Sethy, and A. Nanthaamornphong, 'Enhancing paddy leaf disease diagnosis -a hybrid CNN model using simulated thermal imaging', *Smart Agric. Technol.*, vol. 10, no. November 2024, p. 100814, Mar. 2025, doi: 10.1016/j.atech.2025.100814.

[16] S. Senthil Pandi, A. Senthilselvi, J. Gitanjali, K. ArivuSelvan, J. Gopal, and J. Vellingiri, 'Rice plant disease classification using dilated convolutional neural network with global average pooling', *Ecol. Modell.*, vol. 474, no. October, p. 110166, Dec. 2022, doi: 10.1016/j.ecolmodel.2022.110166.

[17] 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, no. 2022, pp. 2026–2040, 2022, doi: 10.1016/j.procs.2023.01.179.

[18] J. Pan, T. Wang, and Q. Wu, 'RiceNet: A two stage machine learning method for rice disease identification', *Biosyst. Eng.*, vol. 225, pp. 25–40, 2023, doi: 10.1016/j.biosystemseng.2022.11.007.

[19] M. A. R. Khan, U. S. Raisa, M. A. Luthfa, I. Ahammad, and J. R. Das, 'Transfer learning-based comparative study of cnn architectures for automated rice disease detection in Bangladesh', *Franklin Open*, vol. 15, no. February, p. 100513, Jun. 2026, doi: 10.1016/j.fraope.2026.100513.

[20] D. C. Trinh, A. T. Mac, K. G. Dang, H. T. Nguyen, H. T. Nguyen, and T. D. Bui, 'Alpha-EIOU-YOLOv8: An Improved Algorithm for Rice Leaf Disease Detection', *AgriEngineering*, vol. 6, no. 1, pp. 302–317, Feb. 2024, doi: 10.3390/agriengineering6010018.

[21] F. S. Prity, M. Raquib, A. Al Shiam, M. M. Hossain, and K. M. A. Uddin, 'Rice disease classification using feed-forward neural network: Comparative analysis between hybrid and single feature extraction algorithms', *Franklin Open*, vol. 14, p. 100501, Mar. 2026, doi: 10.1016/j.fraope.2026.100501.

[22] M. E. Rana, V. A. Hameed, I. K. Y. Eng, H. K. Tripathy, and S. Mallik, 'Harnessing artificial intelligence for sustainable rice leaf disease classification', *Front. Plant Sci.*, vol. 16, no. September, pp. 1–27, Sep. 2025, doi: 10.3389/fpls.2025.1594329.

[23] M. K. Ambar, H. Öztoprak, and K. Yurtkan, 'Optimizing ResNet-50 for Multiclass Classification: A Multi-Stage Learning Approach', *IEEE Access*, vol. 13, pp. 142517–142534, 2025, doi: 10.1109/ACCESS.2025.3597227.

[24] A. B., M. Kaur, D. Singh, S. Roy, and M. Amoon, 'Efficient Skip Connections-Based Residual Network (ESRNet) for Brain Tumor Classification', *Diagnostics*, vol. 13, no. 20, p. 3234, Oct. 2023, doi: 10.3390/diagnostics13203234.

[25] R. Kaur and S. Singh, 'A comprehensive review of object detection with deep learning', *Digit. Signal Process.*, vol. 132, p. 103812, Jan. 2023, doi: 10.1016/j.dsp.2022.103812.

[26] H. Zeng *et al.*, 'DCAE: A dual conditional autoencoder framework for the reconstruction from EEG into image', *Biomed. Signal Process. Control*, vol. 81, p. 104440, Mar. 2023, doi: 10.1016/j.bspc.2022.104440.

[27] X. Liu and K. M. Goh, 'ResNet: Enabling Deep Convolutional Neural Networks through Residual Learning', Oct. 2025, [Online]. Available: http://arxiv.org/abs/2510.24036

[28] M. S. Alim *et al.*, 'Integrating convolutional neural networks for microscopic image analysis in acute lymphoblastic leukemia classification: A deep learning approach for enhanced diagnostic precision', *Syst. Soft Comput.*, vol. 6, p. 200121, Dec. 2024, doi: 10.1016/j.sasc.2024.200121.

[29] R. Zhang, B. Zhou, C. Lu, and M. Ma, 'The Performance Research of the Data Augmentation Method for Image Classification', *Math. Probl. Eng.*, vol. 2022, pp. 1–10, May 2022, doi: 10.1155/2022/2964829.

[30] V. Švábenský, C. Borchers, E. B. Cloude, and A. Shimada, 'Evaluating the Impact of Data Augmentation on Predictive Model Performance', in *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, New York, NY, USA: ACM, Mar. 2025, pp. 126–136. doi: 10.1145/3706468.3706485.

[31] R. More and J. S. Bradbury, 'Assessing Data Augmentation-Induced Bias in Training and Testing of Machine Learning Models', Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.01825

[32] B. Hu, C. Lei, D. Wang, S. Zhang, and Z. Chen, 'A Preliminary Study on Data Augmentation of Deep Learning for Image Classification', Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.11887

[33] S. Amosedinakaran, P. Anitha, R. Kannan, K. Karthikeyan, S. Suresh, and A. Bhuvanesh, 'Paddy leaf disease identification and K-means cluster segmentation using multi-class SVM techniques', *Meas. Sensors*, vol. 42, no. February 2024, p. 101979, Dec. 2025, doi: 10.1016/j.measen.2025.101979.

[34] R. Dogra, S. Rani, A. Singh, M. A. Albahar, A. E. Barrera, and A. Alkhayyat, 'Deep learning model for detection of brown spot rice leaf disease with smart agriculture', *Comput. Electr. Eng.*, vol. 109, no. October 2022, p. 108659, Jul. 2023, doi: 10.1016/j.compeleceng.2023.108659.