Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3555~3566
e-ISSN: 2548-6861

Comparison of VGG16 and ResNet50 Performance in Rice Leaf Disease Classification

Valda Laura Uswary [1]*, Amriana Amriana [2]**
* Teknik Informatika, Universitas Tadulako
** Program Studi Teknologi Informasi, Fakultas Teknik, Universitas Tadulako, Palu
valdalaura30@gmail.com [1], amriana@untad.ac.id [2]

Article Info

Article history:

Received 2026-07-09
Revised 2026-07-28
Accepted 2026-08-10

Keyword:

Convolutional Neural Network,
Image Classification,
ResNet50,
Rice Leaf Disease,
Transfer Learning,
VGG16.

ABSTRACT

Paddy (Oryza sativa L.) is a strategic staple food commodity in Indonesia, yet its production is frequently disrupted by various plant diseases that cause significant yield losses each year. Conventional visual disease identification is inefficient and prone to error, necessitating the adoption of more reliable, automated diagnostic technology. This study compares the performance of two pre-trained convolutional neural network architectures, VGG16 and ResNet50, in classifying rice leaf images into four categories: Brown Spot, Leaf Blast, Healthy Rice Leaf, and Rice Hispa. Both models were fine-tuned using transfer learning under an identical experimental configuration, including the same data split, optimizer, learning-rate schedule, and augmentation pipeline, to ensure a controlled architectural comparison. Model evaluation was conducted using accuracy, precision, recall, F1-score, the Matthews Correlation Coefficient (MCC), confusion matrix analysis, training convergence behaviour, inference-time computational efficiency, and Grad-CAM interpretability visualization. Experimental results show that ResNet50 achieved a test accuracy of 99% and an MCC of 0.9874, outperforming VGG16, which achieved a test accuracy of 95% and an MCC of 0.9306. Confusion matrix analysis revealed that ResNet50's errors were concentrated almost exclusively within the visually similar brown spot–leaf blast class pair, whereas VGG16 exhibited additional confusion between the healthy rice leaf and rice hispa classes. ResNet50 also demonstrated faster and more stable training convergence, more spatially coherent Grad-CAM activation patterns, and substantially higher throughput under batched inference (356.27 vs. 207.63 images/second at batch size 32), while VGG16 retained a marginal latency advantage under single-image inference. These findings indicate that, under the configuration examined in this study, ResNet50 is the more suitable architecture for rice leaf disease classification, offering an advantageous combination of accuracy, interpretability, and computational efficiency for potential deployment in agricultural monitoring systems.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Paddy (Oryza sativa L.) is an agricultural commodity that forms as the backbone of food security in Indonesia. Given that rice is the principal staple for the majority of Indonesians, maintaining stable production in terms of both quantity and quality is vital for food security. The imbalance in rice production not only affects food availability but can also impact national economic stability more broadly [1].

One of the primary challenges in rice cultivation is the occurrence of plant diseases, which can hinder growth and substantially decrease production yields. Some common diseases that attack rice plants include blast, brown spot, tungro, and hispa. It is estimated that around 200,000 to 300,000 tons of rice are affected by pests and diseases every

year in Indonesia. [2]. This situation is exacerbated by the shortage of agricultural extension workers and the limited knowledge of farmers in identifying and managing plant diseases in a timely manner.

The conventional method of plant disease identification, which relies on visual observation by experts, has been proven to have significant limitations. This manual approach tends to be inefficient, labor-intensive, and susceptible to misinterpretation, particularly in the early stages of infection when the symptoms are still difficult to distinguish visually [3]. Manual visual observation methods not only require a long time and significant effort, but also heavily depend on the availability of limited expert personnel, making them difficult to implement widely, especially in large-scale agricultural lands [4]. If disease symptoms are detected late and not properly managed, it can lead to crop failure, which impacts rice production [1]. In addition, conventional methods are prone to subjective bias because they rely on the observer's experience and individual judgment, so the accuracy of disease diagnosis in the field is often inconsistent [5]. From an economic aspect, rice diseases have the potential to reduce harvest yields by up to 80%, which directly impacts farmers' income, food availability, and market price stability [6].

The recent growth of deep learning as a field of artificial intelligence has fostered new possibilities for developing plant disease detection using digital imagery. An increasingly popular practice is to reuse CNN architecture pre-trained on extensive datasets (for example, ImageNet) then fine-tune them for smaller, specific datasets through transfer learning. This approach yields strong accuracy results despite using a comparatively small training set. The success of CNNs in various cases, from mobile-based megalithic object detection [7], wood type classification [8], to skin cancer detection [9], confirms their flexibility and effectiveness in image classification.

This trend is corroborated by recent large-scale reviews of the field. A systematic review of 160 studies published between 2020 and 2024 found that transfer-learning-based CNN and Vision Transformer models have shown a consistent upward trend in classification accuracy over that period, with detection-oriented studies averaging above 92% accuracy and many recent works exceeding 94–95% [10]. Complementary reviews of deep learning and computer vision in plant disease detection further emphasize that pre-trained CNN architectures remain the dominant backbone choice in precision agriculture research, owing to their demonstrated ability to generalize from large-scale natural image datasets to smaller, domain-specific agricultural datasets through transfer learning, and to their scalability as more labeled data and computational resources become available [11]. These findings reinforce the rationale for adopting a transfer-learning-based approach in the present study, particularly given the practical constraints of limited labeled rice leaf disease imagery in the Indonesian agricultural context.

Among available pre-trained architectures, VGG16 and ResNet50 are two of the most frequently examined models when comparing performance across image classification tasks. Developed by Simonyan and Zisserman of the University of Oxford, VGG16 achieved strong performance in ILSVRC 2014 and has been shown to work well for multiple image classification problems. The VGG16 architecture applied to wood type classification by utilizing transfer learning demonstrates a high capability in extracting discriminative features from image textures [8]. Meanwhile, ResNet50 introduces the concept of residual learning through shortcut connections that are capable of reducing the impact of vanishing gradients in very deep network architectures. The implementation of ResNet50 for mobile-based skin cancer classification successfully achieved good performance, confirming the capability of this architecture in complex visual detection tasks [9]. A number of previous studies comparing these two architectures have yielded varied results. ResNet50 demonstrates superiority on certain datasets, while VGG16 is superior in other contexts. It implies that model superiority is not absolute but depends on the specific dataset characteristics and experimental context.

Although more recent architectures such as EfficientNet [12], ConvNeXt [13], and Vision Transformer (ViT) [14] have demonstrated state-of-the-art performance on large-scale benchmark datasets, this study deliberately limits its scope to VGG16 and ResNet50. This choice is motivated by three considerations. First, VGG16 and ResNet50 represent two fundamentally distinct feature-extraction paradigms—a uniform deep sequential convolutional design versus a residual learning design with shortcut connections—making their comparison more informative for isolating the effect of architectural strategy on classification performance than a comparison between architectures of similar design lineage. Second, both models have been extensively validated as transfer learning baselines on visually comparable domains, including wood texture classification [8] and skin lesion classification [9], allowing the present findings to be cross-referenced with prior studies. Third, architectures such as EfficientNet rely on a compound scaling strategy that requires careful tuning of depth, width, and resolution coefficients, while ConvNeXt and ViT typically demand substantially larger training datasets, extensive data augmentation, and specialized training recipes to reach their reported performance; ViT in particular exhibits weak image-specific inductive bias and tends to underperform when fine-tuned on small to medium-sized datasets unless pre-trained on very large-scale data [14]. Given the comparatively limited size of the rice leaf disease dataset used in this study, VGG16 and ResNet50 were considered more appropriate and reproducible choices, while the exploration of newer architectures is left as a direction for future work.

As outlined earlier, this study aims to measure and compare VGG16 and ResNet50 (pre-trained CNNs) in their ability to classifying rice leaf diseases. The research results are expected to provide recommendations for an optimal model and support the development of a deep learning-based plant disease detection system that is applicable to agriculture in Indonesia.

II. METHOD

This study uses a comparative experimental approach to examine and compare the effectiveness of two pre-trained CNN models (VGG16 and ResNet50) for rice leaf disease classification. Transfer learning was applied by starting the models with weights obtained from ImageNet prior to training on the rice leaf disease dataset. Figure 1 provides an overview of the study's workflow.

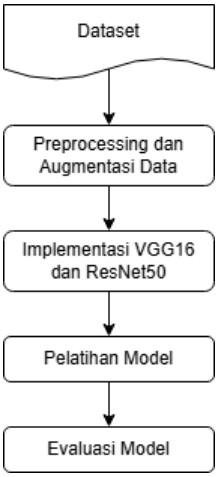

Figure 1. Research Flow

A. Dataset

The dataset used in this study is the same dataset employed in a previously published rice disease classification study that utilized a pre-trained DenseNet model, namely the "Rice Disease Leaves" dataset published on Kaggle [15]. It comprises approximately 5,677 labeled rice leaf images distributed across four classes: Brown Spot, Leaf Blast, Healthy Rice Leaf, and Rice Hispa. This classification reflects the variation of rice leaf diseases commonly encountered in the field, as well as healthy conditions as a control class. Using the same dataset as [15] allows the results of this study to be meaningfully cross-referenced against the DenseNet-based baseline reported therein, and ensures that any performance difference observed between VGG16 and ResNet50 in this study is not attributable to differences in data source or data quality.

The dataset folder structure is read and processed using a specialized function that traverses all image file locations, extracts label information from directory names, and compiles it into a structured data frame to facilitate further manipulation and processing. The dataset was partitioned using a stratified random split into training (70%), validation (15%), and testing (15%) subsets, using a fixed random seed (seed = 42) to guarantee that VGG16 and ResNet50 were trained and evaluated on exactly the same set of training, validation, and test images. This procedure resulted in 854 test images, distributed as 232 Brown Spot, 264 Leaf Blast, 163 Healthy Rice Leaf, and 195 Rice Hispa samples. Figure 2 shows example images from each dataset class.

Figure 2. Dataset

B. Data Preprocessing dan Augmentation

During preprocessing, all images were standardized identically for both networks by converting to RGB, resizing to 224×224 pixels, and normalizing with ImageNet's mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225] to ensure compatibility with the pretrained weights. The same normalization statistics and image size were applied to both VGG16 and ResNet50 without exception, eliminating preprocessing as a possible source of performance discrepancy between the two models.

Data augmentation is applied only to the training set, using an identical augmentation pipeline for both architectures, in order to artificially increase data variation and improve the models' ability to generalize. The augmentation techniques applied include random resized cropping to 224×224 pixels, random affine transformation (rotation up to 10°, translation up to 10% of image size, and scaling between 90% and 110%), random rotation up to 15°, random horizontal flipping, and color jitter (brightness, contrast, and saturation adjusted up to 20%, and hue up to 10%). This augmentation is performed on-the-fly so that each training batch contains different image variations, which is expected to reduce overfitting and improve robustness to real-world variation in leaf orientation, lighting, and camera angle. Because augmentation is applied stochastically and only during training, the reported training accuracy is consistently lower than the validation and test accuracy throughout the training process (see Section III),

since the model is evaluated on a comparatively harder, artificially perturbed version of the data during training, while validation and test images are evaluated in their original, unperturbed form. Meanwhile, the validation and test data are subjected only to basic transformations—resizing and normalization—in order to represent the original image conditions more objectively and to provide an unbiased estimate of generalization performance. All data is then loaded using a DataLoader with an identical batch size of 32 for both models, to enable efficient batch-wise data access during training.

C. Model Architecture and Configuration

ImageNet pre-trained weights are used to initialize both architectures, enabling them to extract fundamental features such as edges, textures, and patterns from the start of training. Both models undergo full fine-tuning, meaning that all layers—including the convolutional backbone and the classifier head—remain trainable throughout training, rather than freezing the backbone and training only the classifier head. This strategy allows both networks to adapt their low- and mid-level feature representations to the specific visual characteristics of rice leaf lesions, at the cost of higher computational demand and a greater risk of overwriting useful pretrained features if the learning rate is not properly controlled.

For each model, the original classifier head was replaced to match the required number of output classes (four), and a Dropout layer (p = 0.5) was inserted immediately before the final linear layer to reduce overfitting. For ResNet50, this dropout is applied within the modified fully connected (fc) layer; for VGG16, it is applied within the final block of the classifier sequential module, replacing the original last linear layer. Placing dropout before—rather than after—the final classification layer ensures that it regularizes the learned feature representation rather than being applied to the already-computed class logits, which would otherwise have no meaningful regularizing effect on the decision boundary.

Both architectures utilize Cross Entropy Loss, an appropriate criterion for multi-class classification. Training used momentum-based stochastic gradient descent (SGD, momentum = 0.9), beginning with a learning rate of 0.01. To maintain stable learning dynamics as training progresses, the learning rate is reduced by a factor of 10 ($\gamma = 0.1$) via a step-based scheduler every five epochs (StepLR). Over 20 training epochs, the checkpoint achieving the maximum validation accuracy was selected and saved as the representative model for that architecture; in the event of a tie in validation accuracy across epochs, the checkpoint with the lower validation loss was retained. This selection served as the test-time model because it generalized best on unseen data.

D. Ensuring a Fair Comparison

Because the objective of this study is to isolate the effect of model architecture on classification performance, every hyperparameter and procedural choice outside of the architecture itself was held constant between VGG16 and ResNet50. Table I summarizes the shared experimental configuration used to train both models.

TABLE I
TRAINING CONFIGURATION FOR VGG16 AND RESNET50

Hyperparameter	**Value (identical for VGG16 & ResNet50)**
Input image size	224 x 224 pixels
Batch size	32
Optimizer	SGD, momentum = 0.9
Initial learning rate	0.01
LR Scheduler	StepLR (step size = 5 epochs, $\gamma = 0.1$)
Loss function	Cross Entropy Loss
Training epochs	20
Dropout (before final linear layer)	p = 0.5
Pretrained weights	ImageNet-1k
Fine-tuning strategy	Full fine-tuning (all layer trainable)
Random seed	42
Train / Val / Test split	70% / 15% / 15%

E. Model Evaluation

Models were evaluated on the test set, using accuracy, precision, recall, F1-score, and the Matthews Correlation Coefficient (MCC) in the classification report, along with confusion matrix analysis. Classification report values are calculated from true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

1) Accuracy

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

2) Precision

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

3) Recall

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

4) F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5) Matthews Correlation Coefficient (MCC)

$$MCC = 2 \times \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC was included in addition to the standard classification-report metrics because, unlike accuracy or F1-score, it produces a single balanced score that accounts for all four confusion-matrix quadrants simultaneously and

remains informative even under class imbalance, making it a more conservative indicator of overall classification quality (MCC values range from −1 to 1, where 1 indicates perfect prediction).

In addition to numerical evaluation, confusion matrix analysis was also conducted and visualized in the form of a heatmap to identify classification error patterns in more detail. The performance comparison between VGG16 and ResNet50 was conducted based on all of these evaluation metrics in order to produce a comprehensive and objective conclusion.

F. Computational Efficiency Evaluation

In addition to predictive performance, this study evaluates the computational efficiency of both models at inference time, which is a practically relevant consideration for real-time or large-scale field deployment. Inference time was measured on the entire test set (854 images) using a CUDA-enabled GPU, under two scenarios: (1) single-image inference, where images are processed one at a time (batch size = 1) to simulate a real-time, on-demand prediction scenario, and (2) batch inference, where images are processed in batches of 32 using the same DataLoader configuration used during training, to evaluate throughput under a more computationally favorable, parallelized setting. Because CUDA operations execute asynchronously, torch.cuda.synchronize() was called immediately before and after each timed operation to ensure that reported timings reflect actual computation time rather than the time required merely to queue GPU instructions.

G. Model Interpretability with Grad-CAM

To assess whether the classification decisions of both models are grounded in biologically meaningful visual evidence rather than spurious background correlations, Gradient-weighted Class Activation Mapping (Grad-CAM) [16] was applied to both trained models. Grad-CAM remains one of the most widely adopted post-hoc visual explanation techniques for convolutional neural networks in recent biomedical and biological image classification studies, owing to its architecture-agnostic nature and its ability to localize class-discriminative regions without requiring model retraining [17]–[19]. Grad-CAM computes the gradient of the target class score with respect to the activation maps of the last convolutional layer, performs global average pooling on these gradients to obtain per-channel importance weights, and uses these weights to produce a weighted combination of the activation maps followed by a ReLU operation. The result is a coarse heatmap that highlights the image regions most influential to the model's prediction for a given class. For ResNet50, the target layer used is the last convolutional layer of the final residual block (layer4[2].conv3); for VGG16, the target layer is the last convolutional layer of the feature-extraction block, immediately preceding the final max-pooling operation (features[28], corresponding to the conv5_3 layer). Both Grad-CAM models were instantiated directly from the fine-tuned weights of the corresponding best checkpoint of each architecture, ensuring that the visualized heatmaps faithfully reflect the decision behavior of the evaluated models.

III. RESULT AND DISCUSSION

The following section evaluates how VGG16 and ResNet50 perform at rice leaf disease image classification under the identical experimental configuration described in Section II. To quantitatively measure the performance of both models, a classification report is used, presenting accuracy, precision, recall, F1-score, and MCC values for each predicted class. In addition, the analysis utilizes confusion matrix visualization, computational efficiency measurements, and Grad-CAM interpretability analysis, allowing the strengths and weaknesses of each model to be understood from complementary perspectives.

A. Training Convergence Analysis

As shown in Figure 3, VGG16's training loss decreased steadily from 1.3786 at epoch 1 to 0.2738 at epoch 20, while validation loss declined more sharply over the same interval, from 1.1245 to 0.0975. A pronounced drop in both curves occurs between epochs 4 and 6, coinciding with the point at which the learning rate scheduler reduces the step size, after which both curves flatten and stabilize. Correspondingly, training accuracy rose from 31.54% to 89.73%, while validation accuracy increased from 53.93% to a peak of 96.60% at epoch 11. Beyond epoch 11, validation accuracy plateaued within a narrow band of approximately 96.3–96.5% for the remaining epochs, and validation loss remained consistently lower than training loss throughout this period. This persistent gap, with validation performance exceeding training performance, is consistent with the effect of dropout and data augmentation being applied only during training, which introduces additional optimization difficulty that is absent when the model is evaluated on the unaugmented validation set.

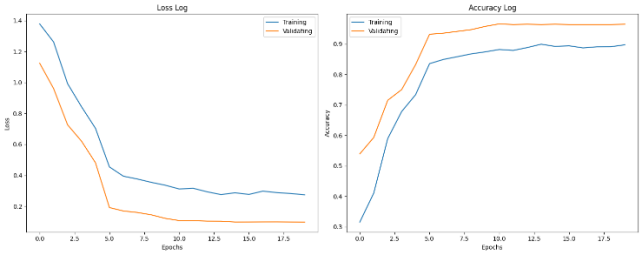

Figure 3. VGG16 Training and Validation Loss and Accuracy Curves

As shown in Figure 4, ResNet50 convergence occurred more rapidly and reached a substantially higher performance ceiling. Training loss fell from 0.7567 to 0.1817, while validation loss dropped from 0.2914 to 0.0213, with the steepest improvement occurring within the first six epochs. Training accuracy increased from 68.48% to a range

fluctuating around 93–95% in later epochs, while validation accuracy rose from 88.39% at epoch 1 to above 98% by epoch 6 and above 99% from epoch 10 onward, reaching a maximum of 99.30% at epoch 13. From epoch 10 through epoch 20, validation loss remained below 0.025 and validation accuracy stayed at or above 98.8%, indicating that the model reached a stable, high-performing state early in training and sustained this performance with only minor fluctuation through the remaining epochs.

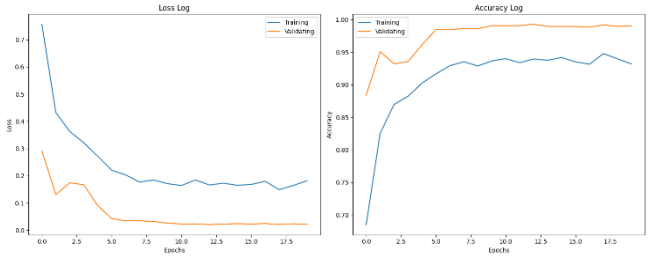

Figure 4. ResNet50 Training and Validation Loss and Accuracy Curves

Comparing the two architectures, ResNet50 exhibits both faster convergence and a substantially lower final validation loss than VGG16 (0.0213 vs. 0.0975), alongside a higher plateau in validation accuracy (approximately 99% vs. approximately 96.5%). This difference in convergence behaviour is consistent with the deeper representational capacity afforded by ResNet50's residual connections, which mitigate vanishing gradients and enable more effective optimization even as network depth increases. In both models, the absence of divergence between training and validation curves in later epochs, together with consistently low validation loss, indicates that neither model exhibits pronounced overfitting under the training configuration used in this study.

B. Evaluation Metrics of VGG16 and ResNet50 Models

Table II and Table III present the classification report for VGG16 and ResNet50, respectively, evaluated on the test set comprising 854 samples distributed across the four rice leaf conditions: Brown Spot (232 samples), Leaf Blast (264 samples), Healthy Rice Leaf (163 samples), and Rice Hispa (195 samples).

VGG16 achieved an overall test accuracy of 95%, with macro-averaged and weighted-averaged precision, recall, and F1-score all equal to 0.95, indicating balanced performance across classes without pronounced bias toward any particular category. For the brown spot class, the model attained a precision of 0.95 and a recall of 0.93, yielding an F1-score of 0.94, indicating that while most predicted brown spot samples were correct, a somewhat larger number of true brown spot samples were missed relative to the other classes. The leaf blast class achieved a precision of 0.95 and a recall of 0.96, resulting in an F1-score of 0.96, reflecting strong and well-balanced classification performance. The healthy rice leaf class recorded the lowest precision among all classes at 0.93, alongside a recall of 0.95, producing an F1-score of 0.94, suggesting a slightly elevated rate of false positives for this class despite good sensitivity. The rice hispa class achieved the highest F1-score among VGG16's four classes at 0.96, with a precision of 0.96 and a recall of 0.95, consistent with the generally balanced performance observed across the model.

TABLE II
VGG16 CLASSIFICATION REPORT

	Precision	**Recall**	**F1-score**	**Support**
Brown Spot	0.95	0.93	0.94	232
Blast	0.95	0.96	0.96	264
Healthy	0.93	0.95	0.94	163
Hispa	0.96	0.95	0.96	195
Accuracy			0.95	854
Macro avg	0.95	0.95	0.95	854
Weighted avg	0.95	0.95	0.95	854

TABLE III
RESNET50 CLASSIFICATION REPORT

	Precision	**Recall**	**F1-score**	**Support**
Brown Spot	0.9747	0.9957	0.9851	232
Blast	0.9962	0.9811	0.9885	264
Healthy	0.9939	1.0000	0.9969	163
Hispa	1.0000	0.9899	0.9948	195
Accuracy			0.9906	854
Macro avg	0.9912	0.9916	0.9914	854
Weighted avg	0.9908	0.9906	0.9906	854

ResNet50 demonstrated markedly stronger performance across all evaluation metrics, achieving an overall test accuracy of 99.06%, with macro-averaged precision, recall, and F1-score of 0.9912, 0.9916, and 0.9914, respectively, and weighted-averaged values of 0.9908, 0.9906, and 0.9906. The brown spot class achieved a precision of 0.9747 and a recall of 0.9957, yielding an F1-score of 0.9851; this indicates near-complete detection of true brown spot samples, although a comparatively larger number of other-class samples were misclassified as brown spot. The leaf blast class attained a near-perfect precision of 0.9962 paired with a recall of 0.9811, resulting in an F1-score of 0.9885, reflecting that predicted leaf blast samples were almost always correct, with only a small number of true leaf blast samples assigned to other categories. The healthy rice leaf class achieved a precision of 0.9939 and a perfect recall of 1.0000, producing the highest F1-score among ResNet50's four classes at 0.9969, indicating that the model correctly identified every true healthy sample in the test set. The rice hispa class recorded a perfect precision of 1.0000 alongside

a recall of 0.9899, resulting in an F1-score of 0.9948, meaning that every rice hispa prediction made by the model was correct, although a small number of true hispa samples were classified into a different category.

To complement the class-wise classification report, the Matthews Correlation Coefficient (MCC) was computed for both models, as this metric incorporates all four quadrants of the confusion matrix (true positives, true negatives, false positives, and false negatives) and is widely regarded as a more robust indicator of classification quality under potential class imbalance than accuracy alone. VGG16 obtained an MCC of 0.9306, while ResNet50 achieved a substantially higher MCC of 0.9874. The magnitude of this difference corroborates the classification report findings and indicates that ResNet50's superior performance is consistent and well-distributed across all four disease categories, rather than being driven disproportionately by strong performance on any single class.

Taken together, these results indicate that although both architectures achieved strong and practically usable classification performance, ResNet50 outperformed VGG16 across every evaluation metric examined, with particularly notable improvements in the healthy rice leaf class, where a perfect recall of 1.0000 was attained, and the rice hispa class, where a perfect precision of 1.0000 was achieved. This performance advantage is further examined in relation to the confusion matrix analysis presented in the following subsection.

C. Confusion Matrix Analysis

Confusion matrices provide a more detailed view of the misclassification patterns for each model. Based on Figure 5, out of 232 actual Brown Spot samples, 215 were correctly predicted by VGG16, while 12 samples were misclassified as leaf blast, 3 as healthy rice leaf, and 2 as rice hispa. This indicates that brown spot was most frequently confused with leaf blast, likely reflecting the visual similarity in lesion colouration and texture between these two diseases. In the leaf blast class, 254 of 264 samples were correctly identified, with 9 samples mistakenly predicted as brown spot and 1 as rice hispa, again reflecting the reciprocal confusion observed between the brown spot and leaf blast categories. For the healthy rice leaf class, 155 of 163 samples were correctly classified, while 2 samples were predicted as brown spot, 1 as leaf blast, and 5 as rice hispa; the latter misclassification suggests that certain healthy leaf textures were occasionally interpreted as bearing hispa-like surface irregularities. The rice hispa class showed the most notable confusion with the healthy rice leaf category, where 9 of 195 true hispa samples were misclassified as healthy, while no hispa samples were confused with either brown spot or leaf blast, indicating that VGG16's errors for this class were concentrated exclusively along the hispa–healthy boundary.

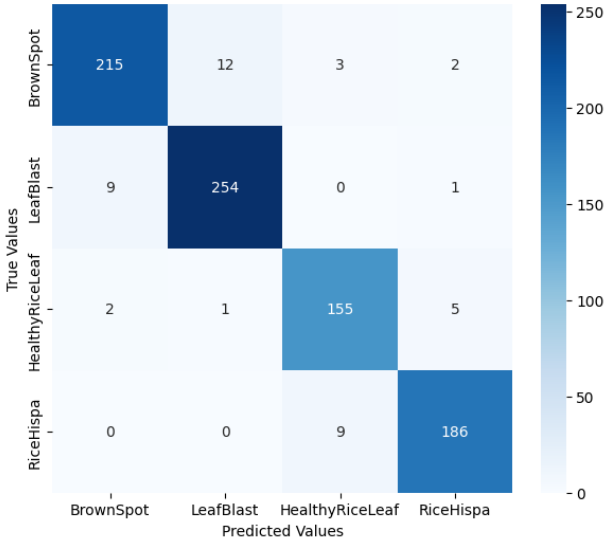

Figure 5. VGG16 Confusion Matrix

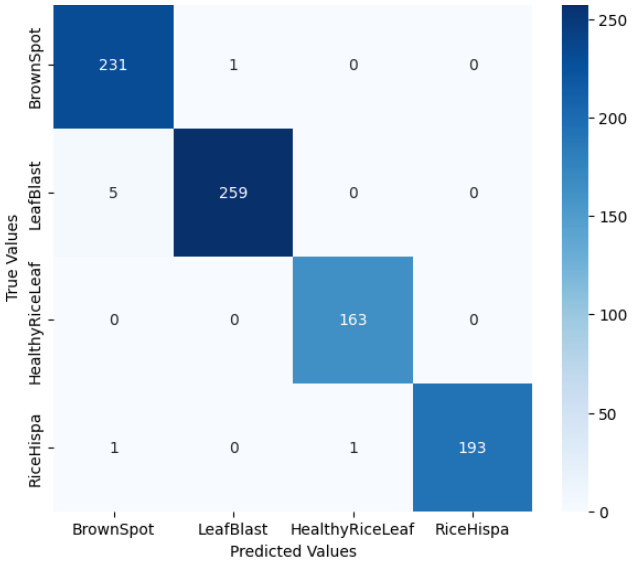

Figure 6. ResNet50 Confusion Matrix

ResNet50 exhibited a substantially cleaner confusion matrix, consistent with its stronger classification report and higher MCC. Of 232 actual brown spot samples, 231 were correctly classified, with only a single sample misclassified as leaf blast. In the leaf blast class, 259 of 264 samples were correctly predicted, with 5 samples misclassified as brown spot and no confusion observed with either the healthy or hispa categories, mirroring the same brown spot–leaf blast confusion pair identified for VGG16, albeit at a considerably reduced magnitude. The healthy rice leaf class achieved perfect classification, with all 163 samples correctly identified and no off-diagonal entries recorded. The rice hispa class showed a minimal degree of confusion, with 193 of 195 samples correctly classified, 1 sample misclassified

as brown spot, and 1 sample misclassified as healthy rice leaf.

Comparing the two matrices, both architectures share a consistent tendency to confuse brown spot and leaf blast, which can be attributed to overlapping visual characteristics such as elongated lesion shape and comparable colouration under certain lighting and background conditions; notably, this class pair accounts for the majority of ResNet50's errors and remains the single most prominent source of misclassification for VGG16 as well. However, VGG16 additionally displayed measurable confusion between the healthy rice leaf and rice hispa classes—a pattern almost entirely absent in ResNet50's predictions, where only one hispa sample was misclassified as healthy and the healthy class itself achieved perfect recall. This distinction suggests that ResNet50's residual learning mechanism enabled the extraction of more discriminative features for distinguishing subtle surface-level abnormalities associated with hispa feeding damage from otherwise healthy leaf tissue, an advantage not equally captured by VGG16's sequential convolutional architecture. Overall, the substantially reduced off-diagonal mass in the ResNet50 confusion matrix reinforces the higher accuracy and MCC values reported in Section III.B, indicating that ResNet50 not only classifies more samples correctly overall but also distributes its errors more narrowly across a single, visually justifiable class boundary rather than multiple sources of confusion.

D. Computational Efficiency Comparison

Table IV summarizes the inference-time efficiency of VGG16 and ResNet50, measured on the full test set of 854 images under two distinct evaluation scenarios: single-image inference (batch size = 1), which simulates a sequential, real-time prediction workflow, and batched inference (batch size = 32), which exploits the parallel processing capability of the GPU. All measurements were performed on a CUDA-enabled device, with torch.cuda.synchronize() invoked before and after each timed operation to ensure accurate wall-clock measurement given the asynchronous nature of CUDA execution.

TABLE IV
INFERENCE EFFICIENCY COMPARISON BETWEEN RESNET50 AND VGG16

Metric	ResNet50	VGG16
Single-image latency (ms/image)	8.55	8.16
Single-image throughput (image/s)	116.96	122.5
Batch latency, size = 32 (ms/image)	2.81	4.82
Batch throughput, size = 32 (image/s)	356.27	207.63

Under the single-image inference scenario, VGG16 and ResNet50 exhibited broadly comparable latency, with VGG16 processing each image in an average of 8.16 ms (122.50 images/second) and ResNet50 requiring a slightly higher average of 8.55 ms per image (116.96 images/second). This marginal difference of approximately 0.39 ms per image indicates that, when images are processed strictly one at a time, the additional depth and residual connections of ResNet50 do not translate into a meaningfully higher per-image computational cost relative to VGG16's comparatively shallower but parameter-heavy architecture. Notably, ResNet50 displayed a wider spread in per-image timing (standard deviation of 1.98 ms, with a maximum of 20.35 ms) compared to VGG16 (standard deviation of 0.80 ms, with a maximum of 13.65 ms), suggesting somewhat greater variability in single-image inference time for ResNet50, potentially attributable to its more complex layer structure and associated kernel-launch overhead under non-batched execution.

A markedly different pattern emerged under batched inference. With a batch size of 32, ResNet50 achieved an average latency of 2.81 ms per image, corresponding to a throughput of 356.27 images/second, whereas VGG16 achieved an average latency of 4.82 ms per image, corresponding to a throughput of 207.63 images/second. This represents an improvement of approximately 71.6% in throughput for ResNet50 relative to VGG16 under batched conditions. Both architectures benefited substantially from batching relative to single-image inference—VGG16's throughput increased by a factor of approximately 1.7×, while ResNet50's throughput increased by a factor of approximately 3.0×—but ResNet50 exhibited a considerably larger relative gain, indicating that its architecture is better optimized to exploit GPU-level parallelism when processing multiple images concurrently.

This divergence between single-image and batched performance can be attributed to architectural differences between the two networks. VGG16 relies on a large, densely connected fully connected classifier head following its convolutional layers, which contributes a substantial number of parameters and computation that does not parallelize as efficiently across a batch dimension. ResNet50, by contrast, employs a global average pooling layer prior to a comparatively lightweight fully connected output layer, together with residual blocks composed primarily of smaller convolutional kernels; this design is inherently more amenable to the parallelized matrix operations exploited during batched GPU inference, resulting in disproportionately higher throughput gains as batch size increases.

From a practical deployment perspective, these findings suggest that the two architectures present different trade-offs depending on the intended use case. For applications requiring immediate, single-image predictions—such as an interactive mobile or field-based diagnostic tool where images are captured and classified one at a time—VGG16 and ResNet50 offer comparable latency, with VGG16 holding a marginal advantage. However, for applications involving batch processing of multiple images, such as large-scale automated screening of stored agricultural imagery or high-throughput monitoring pipelines, ResNet50 offers a substantially more efficient solution, delivering higher

classification accuracy (as established in Section III.B) alongside superior computational throughput.

E. Grad-CAM Interpretability Analysis

To assess the extent to which each model's classification decisions are grounded in disease-relevant visual evidence rather than spurious background cues, Grad-CAM visualization was applied to one representative test sample from each of the four rice leaf classes, using VGG16 (Figure 7) and ResNet50 (Figure 8). The resulting heatmaps highlight the spatial regions that contributed most strongly to each model's prediction, with blue regions denoting the areas of highest gradient-weighted activation.

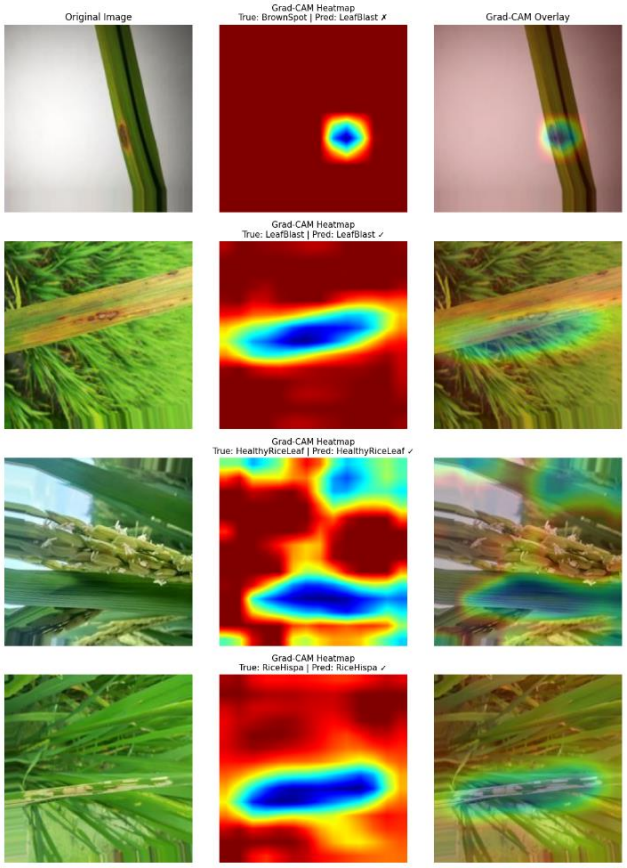

Figure 7. VGG16 Grad-CAM

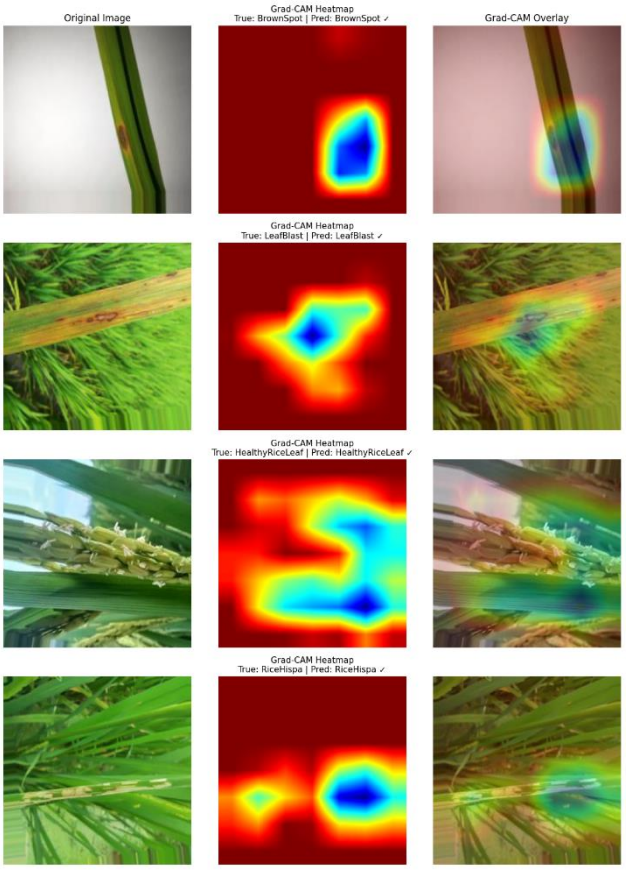

Figure 8. ResNet50 Grad-CAM

For the brown spot sample, VGG16 misclassified the leaf as leaf blast. Its Grad-CAM heatmap shows a small, tightly compact activation positioned below and to the side of the actual necrotic lesion visible on the stem, indicating that the model's attention was not well aligned with the true diagnostic region for this sample. ResNet50, by contrast, correctly classified the same sample, with its heatmap forming a more extensive, vertically elongated activation that directly overlaps the lesion itself. This discrepancy suggests that VGG16's misclassification in this instance stemmed at least partly from attending to a visual region only loosely associated with the actual disease symptom, whereas ResNet50 localized its attention more precisely on the causative feature.

In the leaf blast class, both models produced correct predictions. VGG16's heatmap forms a broad, diagonally elongated activation that follows the streak-like lesion characteristic of leaf blast, while ResNet50's heatmap is comparatively more compact and concentrated near the central portion of the lesion. Although both architectures correctly attend to the diagnostically relevant streak region, the tighter, more concentrated activation produced by ResNet50 suggests a somewhat sharper spatial discrimination of the lesion boundary relative to VGG16's broader activation spread.

For the healthy rice leaf sample, both models again produced correct predictions, but their attention patterns differed noticeably in structure. VGG16's heatmap is fragmented into two visually disconnected activation clusters—one near the upper portion of the panicle and another near its center—reflecting a less coherent attention pattern in the absence of a localized lesion. ResNet50's heatmap, in comparison, forms a continuous, S-shaped region of activation spanning from the lower-left to the mid-right of the image, suggesting that the model integrates evidence from a broader but more spatially connected area of the leaf and panicle surface when no specific disease

feature is present. This more coherent activation pattern in the healthy class may reflect ResNet50's greater capacity to aggregate distributed textural cues through its residual connections.

In the rice hispa class, both models correctly identified the sample, with heatmaps concentrated along the characteristic linear scraping damage produced by larval feeding. VGG16 produced a single, elongated activation region closely following the damaged leaf strip, while ResNet50 produced two distinct activation clusters—a smaller one to the left and a larger, more intense one directly over the central portion of the feeding damage. Despite this structural difference, both models' activations remain concentrated on the correct diagnostic feature, indicating that the hispa feeding pattern constitutes a visually distinctive cue that both architectures reliably capture regardless of underlying network design.

Overall, the Grad-CAM analysis indicates that both VGG16 and ResNet50 generally base their predictions on visually plausible, disease-relevant regions of the leaf rather than irrelevant background areas, supporting the interpretability of both models as practical diagnostic tools. However, the qualitative differences observed—most notably VGG16's misaligned attention on the misclassified brown spot sample and its fragmented activation pattern on the healthy sample, contrasted with ResNet50's more precise and spatially coherent localization across all four classes—offer a visual explanation consistent with ResNet50's superior quantitative performance.

F. Performance Comparison of VGG16 and ResNet50

Synthesizing the results presented in the preceding subsections provides a comprehensive basis for comparing VGG16 and ResNet50 under the identical experimental configuration adopted in this study. In terms of overall predictive performance, ResNet50 outperformed VGG16 across every measured criterion, achieving a higher test accuracy (99% vs. 95%) and a substantially higher Matthews Correlation Coefficient (0.9874 vs. 0.9306), indicating that ResNet50's advantage is consistent across all four disease categories rather than concentrated in a single class. This is further corroborated by the confusion matrix analysis (Section III.C), in which ResNet50's errors were confined almost entirely to the brown spot–leaf blast pair, whereas VGG16 additionally exhibited measurable confusion between the healthy rice leaf and rice hispa classes, indicating a comparatively narrower and more visually interpretable error profile for ResNet50.

This performance advantage is consistent with the training convergence behaviour of both models (Section III.A). ResNet50 reached a substantially lower final validation loss (0.0213 vs. 0.0975) and a higher, more stable validation accuracy plateau (approximately 99% vs. approximately 96.5%) earlier in training than VGG16, suggesting a more favourable optimization landscape for this classification task. The Grad-CAM interpretability analysis (Section III.E) offers a qualitative explanation for this gap: while both architectures generally based their predictions on visually plausible, disease-relevant leaf regions, ResNet50 produced more spatially coherent and precisely localized activations across all four classes, whereas VGG16 exhibited a misaligned activation on its misclassified brown spot sample and a fragmented, disconnected attention pattern on the healthy leaf sample.

The comparison is more nuanced with respect to computational efficiency (Section III.D). Under single-image inference, VGG16 held a marginal latency advantage over ResNet50 (8.16 ms vs. 8.55 ms per image, corresponding to 122.50 vs. 116.96 images/second), indicating that the two architectures impose a broadly comparable computational cost when images are processed strictly one at a time. Under batched inference with a batch size of 32, however, this relationship reversed decisively: ResNet50 achieved a substantially lower per-image latency (2.81 ms vs. 4.82 ms) and a markedly higher throughput (356.27 vs. 207.63 images/second), representing an improvement of approximately 71.6% over VGG16. This divergence reflects ResNet50's global-average-pooling classifier head and residual block structure, which are inherently more amenable to GPU-level parallelization than VGG16's parameter-heavy fully connected layers.

Taken together, these complementary lines of evidence—classification accuracy, MCC, error distribution, training convergence, interpretability, and batched inference throughput—consistently favour ResNet50 as the more suitable architecture for rice leaf disease classification under the configuration examined in this study, with VGG16 retaining a narrow advantage only in strictly single-image, non-batched inference scenarios. For deployment contexts involving large-scale or automated batch screening of rice leaf imagery, ResNet50 offers the more advantageous combination of accuracy and computational efficiency; for applications requiring immediate, one-at-a-time predictions—such as an interactive field diagnostic tool—the latency difference between the two architectures is small enough that other factors, such as model size or memory footprint, may reasonably inform the choice of architecture. Overall, given its consistently higher classification performance, more coherent Grad-CAM attention patterns, and superior batched-inference throughput, ResNet50 is recommended as the primary architecture for rice leaf disease classification in this study, while VGG16 remains a viable alternative in resource-constrained or strictly sequential inference settings.

IV. CONCLUSION

This study conducted a controlled comparison between VGG16 and ResNet50 for the classification of rice leaf diseases into four categories—Brown Spot, Leaf Blast,

Healthy Rice Leaf, and Rice Hispa—under a strictly identical experimental configuration, including the same data split, hyperparameters, optimizer, learning-rate schedule, and augmentation pipeline. Based on the classification report, confusion matrix analysis, Matthews Correlation Coefficient, training convergence behaviour, inference-time efficiency, and Grad-CAM interpretability results, ResNet50 consistently outperformed VGG16 across nearly every evaluated dimension. ResNet50 achieved a higher overall test accuracy (99% vs. 95%) and a substantially higher MCC (0.9874 vs. 0.9306), converged faster and more stably during training with a markedly lower final validation loss, and produced a confusion matrix in which errors were concentrated almost entirely within a single, visually justifiable class pair (brown spot and leaf blast), in contrast to VGG16, which additionally exhibited confusion between the healthy rice leaf and rice hispa classes. Grad-CAM visualization further indicated that both architectures generally based their predictions on visually plausible, disease-relevant leaf regions, supporting the interpretability and practical reliability of both models; however, ResNet50 exhibited more precise and spatially coherent activation patterns, offering a qualitative explanation for its superior quantitative performance.

With respect to computational efficiency, the two architectures presented complementary trade-offs. VGG16 held a narrow latency advantage under single-image inference (8.16 ms vs. 8.55 ms per image), making the two models broadly comparable in strictly sequential prediction scenarios. Under batched inference, however, ResNet50 achieved substantially higher throughput than VGG16 (356.27 vs. 207.63 images/second at batch size 32), an improvement of approximately 71.6%, owing to its residual block structure and lightweight classifier head, which are better suited to GPU-level parallelization than VGG16's parameter-heavy fully connected layers.

Taken together, these findings indicate that, under the configuration examined in this study, ResNet50 is the more suitable architecture for rice leaf disease classification, offering an advantageous combination of predictive accuracy, interpretability, and computational efficiency for potential deployment in real-time or large-scale agricultural monitoring systems, particularly in batch-processing contexts. VGG16 nonetheless remains a viable alternative in resource-constrained or strictly single-image inference settings, where its marginal latency advantage may be relevant. Given the near-perfect accuracy achieved by ResNet50, these results should be interpreted with appropriate caution regarding the relatively controlled conditions under which the underlying dataset was collected; future research should evaluate both architectures on independent, field-collected rice leaf images encompassing broader variation in lighting conditions, disease severity, and camera equipment to confirm the generalizability of these findings. Future work may also extend this comparison to more recent architectures such as EfficientNet, ConvNeXt, or Vision Transformer models, as well as explore model compression and edge-deployment strategies to further support the practical adoption of deep learning-based rice disease diagnosis tools in Indonesian agriculture.

REFERENCES

[1] P. Sitompul, H. Okprana, A. Prasetio, and G. Artikel, “Identifikasi Penyakit Tanaman Padi Melalui Citra Daun Menggunakan DenseNet 201,” JOMLAI: Journal of Machine Learning and Artificial Intelligence, vol. 1, no. 2, pp. 2828–9099, Sept. 2022.

[2] K. H. F. Ningrum. (2025) Mengenal Penyakit Penting pada Tanaman Padi on Mtanigroup. [Online]. Tersedia: https://mtanigroup.com/2025/03/11/mengenal-penyakit-penting-pada-tanaman-padi/.

[3] R. Suciani, D. A. Anugra, E. Faisal, and D. Nuswantoro, “Deteksi Penyakit Daun Padi Menggunakan Deep Learning Dengan Arsitektur CNN,” Journal of Information Technology and Computer Science (INTECOMS), vol. 8, no. 5, Sept. 2025.

[4] K. Hu, X. Zheng, X. Su, L. Wu, Y. Liu, and Z. Deng, “Identification of rice leaf disease based on DepMulti-Net,” Front. Plant Sci., vol. 16, no. 1522487, Mar. 2025

[5] M. Nawaz, I. Ahmed, S. Ali, and W. Khan, “Multi-model machine learning for automated identification of rice diseases using leaf image data,” PLOS ONE, vol. 19, no. 9, p. e0307461, Sept. 2025.

[6] S. N. Yousafzai, F. N. Al-Wesabi, H. Alsolai, S. A. Ebad, I. M. Nasir, E. Fadhal, and A. Thaljaoui, “Advanced clustering and transfer learning based approach for rice leaf disease segmentation and classification,” PeerJ. Computer science, vol 11, e3018, Jul. 2025.

[7] M. Fahmi, R. Rahim, A. Basri, and S. Samsul, “Penerapan Convolution Neural Network (CNN) Untuk Megalitikum Di Sulawesi Tengah Berbasis Mobile”, Jurnal Ilmiah Penelitian dan Pembelajaran Informatika (JIPI), vol. 10, no 3, pp. 2525-2536, Aug. 2025.

[8] N. A. Afiah, S. Syahrullah, R. Ardiansyah, R. Laila, and R. Pohontu, “Application Of Vgg16 Architecture In Wood Type Classification Using Convolutional Neural Network”, J. Tek. Inform. (JUTIF), vol. 6, no. 1, pp. 335–344, Feb. 2025.

[9] A. Asriani, N. T. Lapatta, D. W. Nugraha, A. Amriana, and W. Wirdayanti, “Implementation of ResNet-50-Based Convolutional Neural Network For Mobile Skin Cancer Classification”, JAIC, vol. 9, no. 4, pp. 1969–1577, Aug. 2025.

[10] I. Pacal, I. Kunduracioglu, M. H. Alma, M. Deveci, S. Kadry, J. Nedoma, V. Slany, and R. Martinek, "A systematic review of deep learning techniques for plant diseases," Artif. Intell. Rev., vol. 57, no. 11, art. 304, Sept. 2024.

[11] A. Upadhyay, N. S. Chandel, K. P. Singh, S. K. Chakraborty, B. M. Nandede, M. Kumar, A. Subeesh, K. Upendar, A. Salem, and A. Elbeltagi, "Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture," Artif. Intell. Rev., vol. 58, no. 3, art. 92, Jan. 2025.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. 36th Int. Conf. Machine Learning (ICML), Long Beach, CA, USA, 2019, pp. 6105–6114.

[13] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 11976–11986.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. 9th Int. Conf. Learning Representations (ICLR), Virtual Event, Austria, 2021.

[15] H. M. Yusuf, A. F. D. Kana, and S. A. Yusuf, "Enhanced rice disease classification through pre-trained densenet model using transfer learning," Franklin Open, vol. 13, Okt. 2025.
[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Int. J. Comput. Vis., vol. 128, no. 2, pp. 336–359, 2020.
[17] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, "Is Grad-CAM Explainable in Medical Images?," in Computer Vision and Image Processing (CVIP 2023), Communications in Computer and Information Science, vol. 2009, Springer, Cham, 2024, pp. 124–135.
[18] S. Tehsin, I. M. Nasir, and R. Damaševičius, "Explainability of Disease Classification from Medical Images Using XGrad-CAM," Communications in Computer and Information Science, vol. 2348, pp. 221–236, 2025.
[19] A. M. Fayyaz, S. J. Abdulkadir, N. Talpur, and S. M. Al-Selwi, "Grad-CAM (Gradient-weighted Class Activation Mapping): A Systematic Literature Review," Comput. Biol. Med., vol. 198, part B, art. no. 111200, 2025.