

## Article

# Global-Local Feature-Based Rice Leaf Disease Classification Using Two-Stream Deep Neural Network Feature Fusion

Md Nahidur Rahaman <sup>1</sup>, Abdullah Al Mamun <sup>1</sup>, Md. Kamal Hossen <sup>1</sup>, Abdur Rouf <sup>1</sup>, Tumpa Rani Shaha <sup>2</sup>, Jungpil Shin <sup>3</sup>, Mohd Nizam Husen <sup>4,\*</sup> and Abu Saleh Musa Miah <sup>4,5,\*</sup>

<sup>1</sup> Department of Computer Science and Engineering, Dhaka University of Engineering & Technology (DUET), Gazipur 1707, Bangladesh; 204010@student.duet.ac.bd (M.N.R.); 194011@student.duet.ac.bd (A.A.M.); 204021@student.duet.ac.bd (M.K.H.); 204014@student.duet.ac.bd (A.R.)

<sup>2</sup> Department of Computer Science and Technology, Gopalganj Science and Technology University, Gopalganj 8105, Bangladesh; tumpa.cse49@gmail.com

<sup>3</sup> School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-0006, Fukushima, Japan; jpshin@u-aizu.ac.jp

<sup>4</sup> Malaysian Institute of Information Technology (MIIT), Universiti Kuala Lumpur (UniKL), 1016 Jalan Sultan Ismail, Kuala Lumpur 50250, Malaysia

<sup>5</sup> Department of Computer Science and Engineering, Rajshahi University, Rajshahi 6205, Bangladesh

\* Correspondence: mnizam@unikl.edu.my (M.N.H.); musa.cse@ru.ac.bd (A.S.M.M.)

## Abstract

Rice leaf diseases significantly affect crop health and yield potential, creating a need for accurate and timely disease diagnosis. Many existing approaches still rely on single-stream feature extraction architectures, which may limit the ability to simultaneously capture global contextual information and fine-grained disease characteristics. Moreover, limited interpretability and decision-support capability hinder their practical deployment in real-world rice farming. To address these limitations, we employed a framework consists of two parallel feature extraction streams designed to capture different characteristics of disease patterns. The first stream uses a Swin Transformer to learn global contextual information and long-range spatial relationships across the leaf image. The second stream employs ConvNeXt to extract local texture features, including lesion details, spots, and color variations. By combining these complementary representations, the proposed framework effectively integrates global semantic information with local disease-specific features for improved classification performance. The extracted features from the two streams are fused through concatenation followed by an attention-based feature refinement module, enabling adaptive weighting of discriminative features. The refined representation is then used by a fully connected classifier for disease prediction. To enhance model interpretability, Grad-CAM visualization is incorporated to highlight disease-relevant regions and provide visual explanations for the model decisions. Furthermore, an LLM-based advisory module is integrated as a post-diagnosis decision-support component to provide contextualized disease management information and suggestions based on the predicted disease category. The generated suggestions are intended to support, rather than replace, expert agronomic recommendations and should be validated by agricultural professionals before practical application. The proposed framework was evaluated on two rice leaf disease datasets, achieving accuracies of 99.55% and 97.06% on Dataset-1 and Dataset-2, respectively, which are higher than those reported in previous studies. Additionally, five-fold cross-validation on Dataset-1 achieved an average accuracy of  $98.91\% \pm 0.47$ , demonstrating the stability of the proposed approach. Cross-dataset evaluation using nine common disease classes across both datasets achieved 90.80% accuracy, indicating improved generalization across different data distributions. The proposed framework provides an accurate and explainable approach for rice leaf disease diagnosis in smart agriculture applications.



Academic Editor: Xueyi Li

Received: 28 July 2026

Revised: 10 August 2026

Accepted: 10 August 2026

Published: 14 August 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** rice disease; hybrid deep learning; attention fusion; Grad-CAM; LLM advisory

## 1. Introduction

Rice (*Oryza sativa*) is one of the most important staple crops, providing a primary food source for more than 50% of the global population. However, rice production is continuously affected by various foliar diseases, including Neck Blast, Leaf Scald, Sheath Blight, Narrow Brown Leaf Spot, Leaf Blast, Rice Hispa, Brown Spot, and Bacterial Leaf Blight. These diseases significantly reduce crop productivity and quality, creating serious challenges for global food security, particularly in developing countries where agriculture plays a major role in the economy. Furthermore, favorable environmental conditions can accelerate disease propagation, emphasizing the importance of accurate and timely disease diagnosis for effective crop management [1–3].

Traditionally, rice disease identification has relied on expert visual inspection and laboratory-based analysis [4]. Although these approaches can provide reliable diagnosis, they are often time-consuming, labor-intensive, and difficult to access for small-scale farmers in remote agricultural regions. Inaccurate or delayed disease identification may lead to inappropriate pesticide application, increasing production costs and causing potential environmental impacts [5]. Therefore, automated and accessible disease recognition systems have become increasingly important for supporting efficient agricultural practices.

In recent years, deep learning (DL)-based approaches, particularly Convolutional Neural Networks (CNNs), have demonstrated significant potential for plant disease classification using leaf images. CNN-based methods effectively extract local spatial patterns, including color variations, textures, and lesion characteristics, enabling high classification performance under controlled experimental conditions [6]. However, CNN-based models may face challenges in capturing long-range dependencies and global contextual relationships due to their inherently local receptive field operations [2]. These limitations can affect their ability to fully represent complex disease patterns that involve both localized symptoms and broader leaf-level characteristics.

To address the limitations associated with convolution-based feature extraction, Vision Transformers (ViTs) have recently gained attention due to their ability to model global relationships through self-attention mechanisms. Unlike CNNs, transformer-based architectures represent images as sequences of patches, enabling comprehensive analysis of spatial dependencies across the entire image [7]. Consequently, transformer-based methods have demonstrated improved robustness and competitive performance in various plant disease recognition tasks [8]. Furthermore, a deep learning-based ROA-DM framework has been proposed for automated rice leaf disease classification, integrating a Deep Autoencoder (DAE), Deep Max-out Network (DMN), and Remora Optimization Algorithm (ROA) [9]. However, despite their high predictive performance, many existing DL-based agricultural diagnosis systems still face challenges related to model interpretability. These models often operate as black-box systems, providing predictions without sufficient explanations regarding disease-related visual evidence. Such limitations reduce user confidence and hinder practical adoption in agricultural applications. Therefore, Explainable Artificial Intelligence (XAI) techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM) and attention visualization, have become increasingly important for identifying disease-relevant regions and improving model transparency [10].

Beyond accurate and interpretable disease recognition, practical agricultural applications require intelligent decision-support capabilities that can provide understandable and actionable recommendations to farmers. Recently, Large Language Models (LLMs) have

demonstrated remarkable capabilities in natural language understanding and knowledge-based reasoning. The integration of LLMs with disease recognition systems provides opportunities to generate disease-specific management recommendations, preventive strategies, and treatment guidance, thereby bridging the gap between automated diagnosis and agricultural decision-making [11,12]. Nevertheless, many existing studies remain primarily focused on offline model evaluation, while challenges related to real-time deployment, user interaction, interpretability, and practical usability remain insufficiently addressed. Therefore, an integrated framework combining accurate classification, explainable prediction, and intelligent advisory support is highly desirable for real-world smart farming applications [13].

To address these challenges, this study proposes a dual-backbone complementary representation learning framework for accurate and interpretable rice leaf disease diagnosis. The proposed framework integrates Swin Transformer and ConvNeXt [14] architectures to exploit complementary visual representations. Specifically, Swin Transformer is employed to capture global contextual dependencies and long-range spatial relationships, whereas ConvNeXt is utilized to extract fine-grained local texture features associated with disease symptoms. An attention-based feature fusion module is further applied to adaptively combine these heterogeneous feature representations and enhance discriminative disease feature learning. Additionally, Grad-CAM is incorporated to provide visual explanations by highlighting disease-relevant regions contributing to model predictions. Furthermore, a web-based LLM advisory module is introduced as a post-diagnosis decision-support component to generate disease-specific management recommendations based on the predicted disease category. The LLM module is used exclusively for advisory purposes and does not participate in feature extraction, classification, or performance evaluation.

The main contributions of this study are summarized as follows:

- A dual-backbone complementary representation learning framework is proposed by integrating Swin Transformer and ConvNeXt to capture global contextual information and local disease-specific texture characteristics.
- An attention-based adaptive feature fusion mechanism used to combine complementary feature representations and improve discriminative disease classification.
- A comprehensive evaluation framework is established using multiple state-of-the-art baseline models, cross-validation, and cross-dataset testing for robustness and generalization capability.
- An explainable disease diagnosis mechanism based on Grad-CAM is incorporated to improve model transparency and identify disease-relevant image regions.
- A practical web-based decision-support extension with an LLM-based advisory module is introduced to provide disease-specific agricultural recommendations after diagnosis.

The remainder of this paper is organized as follows. Section 2 presents the related work, including recent advances in rice leaf disease diagnosis, transformer-based architectures, explainable artificial intelligence, and intelligent agricultural decision-support systems. Section 3 describes the proposed methodology, including dataset preparation, preprocessing, dual-backbone Swin Transformer-ConvNeXt architecture, attention-based feature fusion, Grad-CAM-based explainability, and the post-diagnosis LLM advisory module. Section 4 presents experimental results and detailed analysis, including comparisons with baseline models, ablation studies, cross-validation results, cross-dataset evaluation, confusion matrix analysis, ROC-AUC analysis, and explainability assessment. Section 5 discusses the web-based deployment framework and its applicability for smart agricultural practices. Finally, Section 6 concludes the paper and highlights future research directions.

## 2. Related Work

The growing need for intelligent and scalable agricultural solutions has made automated rice leaf disease detection an active research topic. In recent years, deep learning-based solutions have greatly enhanced the precision and effectiveness of plant disease diagnosis. Nevertheless, the current approaches differ in their capacity to reflect complicated features, be interpretable, and be applicable in the real world. Convolutional Neural Networks (CNNs) have been widely applied to rice leaf disease classification because they are powerful at extracting local spatial features like texture, color, and lesion patterns. For example, Nguyen et al. [15] introduced an LCSNet–FedPerGC framework for rice leaf disease classification, which is a combination of a lightweight CNN and an optimized federated learning strategy. The framework was shown to be highly accurate for classification, efficient for inference and robust with respect to both IID and non-IID data distributions. A similar approach was used to introduce a Concurrency Attention-Based CNN for rice leaf disease recognition under different lighting conditions, which uses a multiple-parallel attention mechanism to capture discriminative features and ignore background interference, with a classification accuracy of 96.2% [16]. Moreover, Twin-CNN structures [17] have shown enhanced performance through the application of dual-stream feature extraction. Despite the promising results obtained using CNN-based models, they mainly use local receptive fields, which restricts their capacity to extract long-range dependencies and global contextual information. As a result, they can perform poorly in real-life situations with complex backgrounds, occlusions, and different illumination. Transformer-based models have been suggested to address the shortcomings of CNNs in detecting plant diseases. Vision Transformers (ViTs) are self-attention models that are used to capture global relations between image patches, enhancing contextual knowledge. As shown by Singha Roy and Kukreja [18], ViT-based models are more robust when it comes to rice leaf disease classification. Likewise, AISOA-SSformer [19] uses a transformer-based segmentation framework with sparse attention to boost disease localization. In addition, Swin Transformer-based approaches [20] have shown strong performance by incorporating hierarchical feature representation and shifted window attention mechanisms. Explainable transformer models [21] further improve transparency by incorporating attention visualization. Nevertheless, the benefits of transformer-based models are not without drawbacks as they typically need massive datasets and high-performance computing, and most current implementations cannot be explained or deployed. More recent studies have concentrated on hybrid architectures that combine CNN and transformer models to utilize both local and global feature representations. As an example, GRG-ViT [22] combines generative adversarial networks (GANs) with Vision Transformers to deal with the issue of class imbalance and enhance generalization. Likewise, hybrid CNN–transformer models [23] exhibit better performance due to the integration of convolutional feature extraction and global attention mechanisms. In addition, ConvNeXt-based architectures [24] are also suggested as CNN backbone modernizations that enable a transition between CNNs and transformers. Hybrid models based on multiple AI models that are green and emphasize computational efficiency and sustainability are also highlighted as green AI-based hybrid models [25]. Nevertheless, A considerable number of hybrid models do not include integrated explainability analysis and do not encompass the use of better classification performance, even though they are more effective than their counterparts. Explainable Artificial Intelligence (XAI) has become a key element of enhancing trust and transparency in AI-based agricultural systems. Grad-CAM and other techniques have been extensively applied to identify disease-related areas on leaf images [26]. Transformer-based explainability techniques [21] further increase interpretability through the use of attention maps. Despite the fact that these methods offer great insight into how models make decisions, they are not usually incorporated into a



holistic decision-support system but are used as modules on their own. The combination of Large Language Models (LLMs) with vision-based systems is a recent development that has introduced new opportunities in intelligent agriculture. Advisory frameworks assisted by an LLM [27] can be used to produce disease-specific recommendations, such as treatment plans and prevention. Multimodal vision-language models [28,29] also add more capabilities to the system by integrating image comprehension with natural language reasoning. Also, research emphasizes the need to go beyond dataset-based assessment to real-world usage [30]. Practical AI solutions for farmers are feasible as illustrated by web-based deployment frameworks and real-time mobile systems [31] and [32], respectively. Nevertheless, the current LLM implementations tend to be disconnected with the model of explainable visions, which prevents them from connecting predictions with practical insights. Although recent progress has been made, a number of problems still exist. First, the existing approaches focus mainly on classification performance and do not fully consider the robustness of the system in its real-world applications involving changes in lighting conditions, background complexity and other environmental factors. Second, many hybrid CNN–Transformer models have been proposed to improve feature representation, but most do not yet fully leverage the complementary characteristics of global information and fine-grained local information related to disease textures via adaptive feature learning. Third, there has been little time spent on incorporating explainable AI approaches, which could boost the openness and credibility of disease prognosis. Finally, although there are many research papers that look at automatic disease detection, there are limited studies have explored integrating diagnosis models with farmer-oriented decision-support systems. In view of this, a dual-backbone complementary representation learning network is proposed to tackle the above problems, which involves using both Swin Transformer and ConvNeXt with an attention-based feature fusion module to provide a robust and interpretable rice leaf disease diagnosis. A list of suggested management recommendations is also included using an LLM-based advisory system, which is further augmented with a web-enabled system, and Grad-CAM to offer visual explanations.

Table 1 shows the quantitative comparison of various methods used for detecting diseases on rice leaves. Some of the important characteristics include model structure, explainability, advisability, deployment, robustness, and scalability.

**Table 1.** Comparison of recent rice leaf disease detection methods.

| Study                         | Dataset                     | Method                     | Deployment | XAI | Accuracy (%) | Limitation                                      |
|-------------------------------|-----------------------------|----------------------------|------------|-----|--------------|-------------------------------------------------|
| Nguyen et al. (2026) [15]     | Multi-class Rice Dataset    | FedPerGC FL                | No         | No  | 98.25        | No real-time deployment                         |
| Premkumar et al. (2026) [16]  | Rice Leaf Dataset           | CNN + Attention            | Yes        | Yes | 96.2         | Lacks global feature learning                   |
| Singha Roy et al. (2025) [18] | Rice Disease Dataset        | Vision Transformer (ViT)   | No         | No  | 77.94        | Low accuracy                                    |
| Dai et al. (2024) [19]        | Rice Segmentation Dataset   | Transformer (SSFormer)     | No         | No  | N/R          | No deployment support                           |
| Felicitia et al. (2026) [22]  | Augmented Rice Dataset      | GAN + ViT Hybrid           | No         | Yes | 96.0         | No advisory system                              |
| Pai et al. (2025) [17]        | Rice Leaf Dataset           | Dual CNN Fusion            | No         | Yes | 96.8         | Limited interpretability                        |
| Al-Falluji et al. (2025) [25] | Rice Disease Dataset        | Hybrid ConvNeXt            | No         | Yes | 90.69        | No real-time deployment                         |
| Shuai et al. (2026) [29]      | Plant disease image dataset | AgriMAPO                   | No         | No  | 90.3         | Weak explainability                             |
| Singh et al. (2026) [32]      | Self-collected              | MobileViT                  | No         | Yes | 99.00        | No advisory integration                         |
| Srinivasan et al. (2026) [21] | D-I and D-II                | ResViT-152                 | No         | Yes | 99.23        | No advisory system                              |
| Wang et al. (2026) [27]       | Agricultural Text Dataset   | LLM Advisory System        | No         | No  | N/R          | No vision model integration                     |
| Proposed Work                 | Multi-source Rice Dataset   | SwinV2 + ConvNeXt (Hybrid) | Yes        | Yes | 99.55        | Computational cost (future optimization needed) |

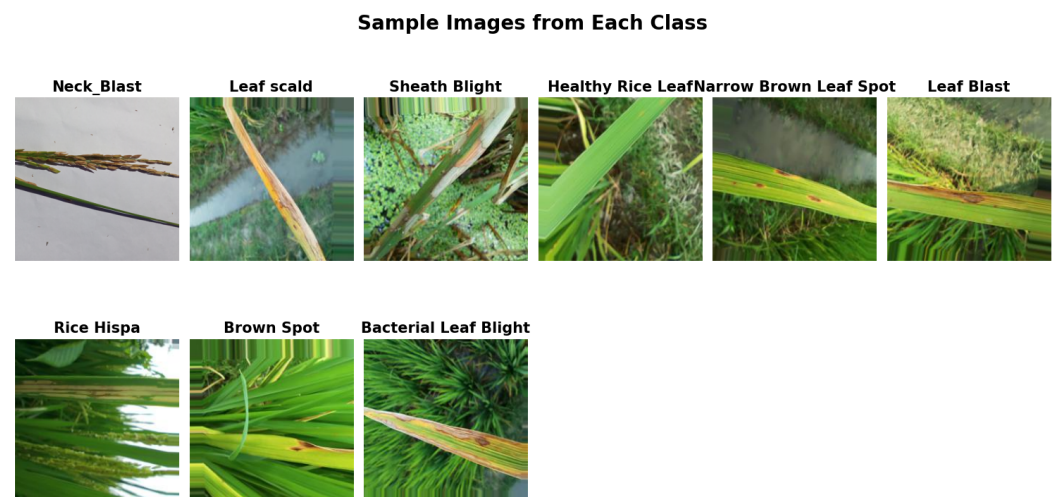
As summarized in Table 1, most of the current techniques are mainly employed for classification accuracy and/or feature extraction, lacking a holistic and verifiable diagnostic framework. CNN-based approaches [15,16] well model local visual patterns with limited capability of modeling long-range contextual dependencies. Transformer-based methods [19] have been employed to enhance the modeling of global context, but they generally have limited interpretability and applicability. Current hybrid methods [22,23,25] blend CNNs and Transformers; however, they most commonly integrate features using static or identity

mappings and fail to substantially leverage the complementary feature representations using adaptive attention mechanisms. The proposed model, on the contrary, adopts the dual-backbone complementary representation learning scheme and attention-based feature fusion for simultaneous learning of global contextual and local texture representation and generalization of disease characterization. Additionally, Grad-CAM makes the model more interpretable and the LLM advisory module, accessible via the web, is used only as a post-diagnosis decision-support component (presenting management suggestions to the farmer rather than involvement in disease classification). The entire system is implemented on the Hugging Face platform, facilitating fast and on-the-go inference and deployment in the field of smart agriculture.

### 3. Dataset Description

#### 3.1. Dataset-1

This study utilizes a publicly available rice leaf disease image dataset collected from real-world agricultural field environments. The dataset was obtained from the Kaggle platform and contains high-resolution images representing different healthy and diseased rice leaf conditions. Since the images were collected in practical field environments, the dataset includes considerable variations in illumination, background complexity, leaf orientation, and image appearance, providing a challenging benchmark for evaluating the robustness and generalization capability of deep learning-based rice disease classification models. To analyze the dataset characteristics, class distribution and visual examples are presented in Figures 1–3. Figure 1 illustrates representative images from different disease categories, demonstrating variations in lesion patterns, color changes, and leaf appearances. Figure 2 presents the number of samples available in each disease category, while Figure 3 shows the corresponding percentage distribution.



**Figure 1.** Representative samples from Dataset-1.

The dataset consists of 11,790 images categorized into nine classes, including both healthy and diseased rice leaves. The disease categories include Neck Blast, Leaf Scald, Sheath Blight, Healthy Rice Leaf, Narrow Brown Leaf Spot, Leaf Blast, Rice Hispa, Brown Spot, and Bacterial Leaf Blight. These categories represent common rice diseases that significantly affect crop productivity and quality.

The diversity of the dataset, including variations in environmental conditions and disease appearances, provides a suitable benchmark for developing robust deep learning models capable of generalizing beyond controlled laboratory conditions. Furthermore, careful dataset organization was performed to prevent data leakage by ensuring that

visually identical or highly correlated samples were not simultaneously included in training and testing subsets.

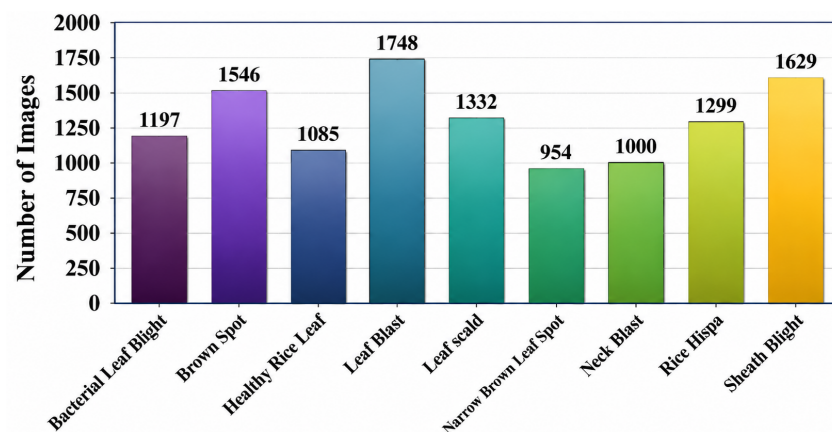


Figure 2. Class-wise distribution of Dataset-1.

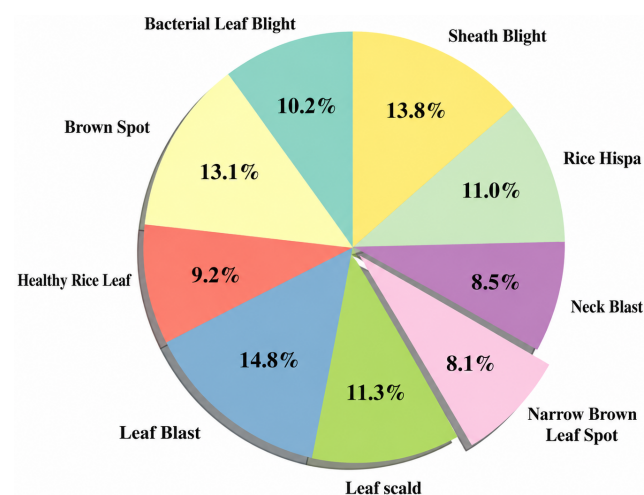


Figure 3. Percentage distribution of Dataset-1 classes.

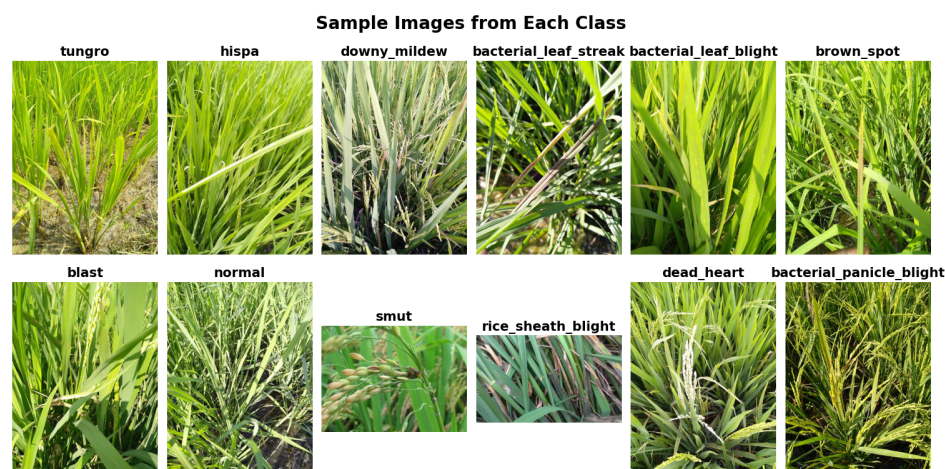
### 3.2. Dataset-2

A second rice disease image dataset shared by the public was used to further validate the proposed framework and its ability to generalize. The dataset was downloaded from the Kaggle platform and includes high-resolution image data captured in the field, where they represent images from multiple disease types in addition to the plant images that are healthy. The images were taken under actual cultivation conditions, resulting in significant differences in light conditions, camera angles, background noise, various vegetative growth stages, and disease symptoms, making it a difficult reference for deep learning rice disease classification.

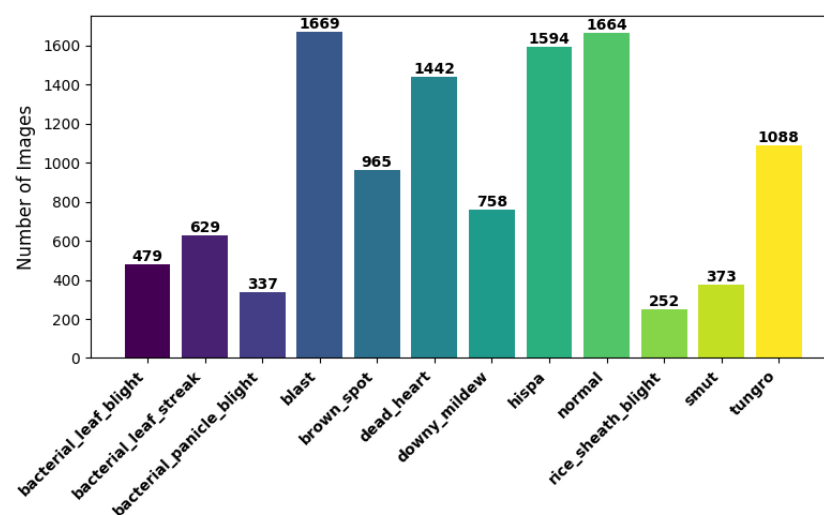
These images consist of 11,250 diseased and normal rice plants across 12 categories: Bacterial Leaf Blight, Bacterial Leaf Streak, Bacterial Panicle Blight, Blast, Brown Spot, Dead Heart, Downy Mildew, Rice Hispa, Healthy Rice Plant (Normal), Rice Sheath Blight, Smut, and Tungro. The following disease categories include major rice diseases occurring commonly under practical cultivation conditions which have a significant effect on yield and quality.

To investigate the features of Dataset-2, representative samples of images along with the distribution of images among the classes are shown in Figures 4–6. Eleven representative images were obtained from each disease category and are presented in Figure 4 showing differences in lesion morphology, the colors of lesions, lesion texture, and background in

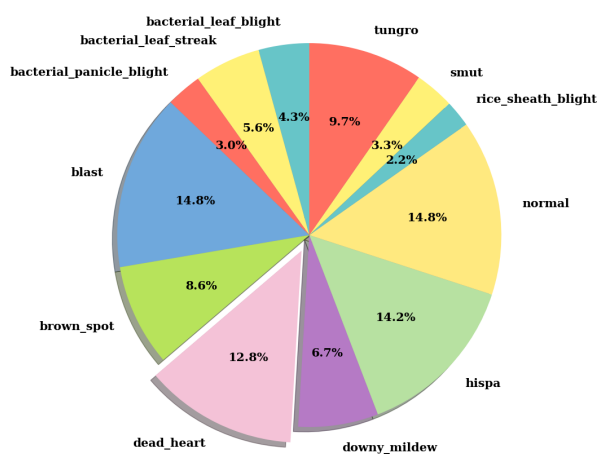
the fields. The number of images that are available per class is shown in Figure 5, while the percentage distribution shows the number of images in each of the categories in Figure 6.



**Figure 4.** Representative samples from Dataset-2.



**Figure 5.** Class-wise distribution of Dataset-2.



**Figure 6.** Percentage distribution of Dataset-2 classes.

The high diversity of Dataset-2 in disease manifestations, field variations, light conditions, and acquisition angles offers a suitable reference point for evaluating the usefulness and robustness of the deep learning models in the real farming environment. Furthermore,



the datasets were organized and split into training, validation, and testing sets to reduce potential data leakage by not having two or more samples similar to each other in both training and testing sets at the same time.

### 3.3. Dataset Partitioning

To evaluate the performance of the proposed framework, the dataset was divided into training, validation, and testing subsets using a stratified splitting strategy. The dataset-1 was partitioned with a ratio of 70% for training, 15% for validation, and 15% for testing, whereas Dataset-2 was divided using a 60%, 20%, 20%. Stratified sampling was applied to maintain the original class distribution across all subsets and ensure balanced representation of each rice disease category.

The dataset distribution after partitioning is summarized in Table 2.

**Table 2.** Dataset split distribution.

| Dataset        | Split Ratio | Dataset-1 | Dataset-2 |
|----------------|-------------|-----------|-----------|
| Training Set   | 70%/60%     | 8253      | 6750      |
| Validation Set | 15%/20%     | 1768      | 2250      |
| Testing Set    | 15%/20%     | 1769      | 2250      |
| Total          | 100%        | 11,790    | 11,250    |

For each disease category  $c$ , the stratified partitioning can be represented as

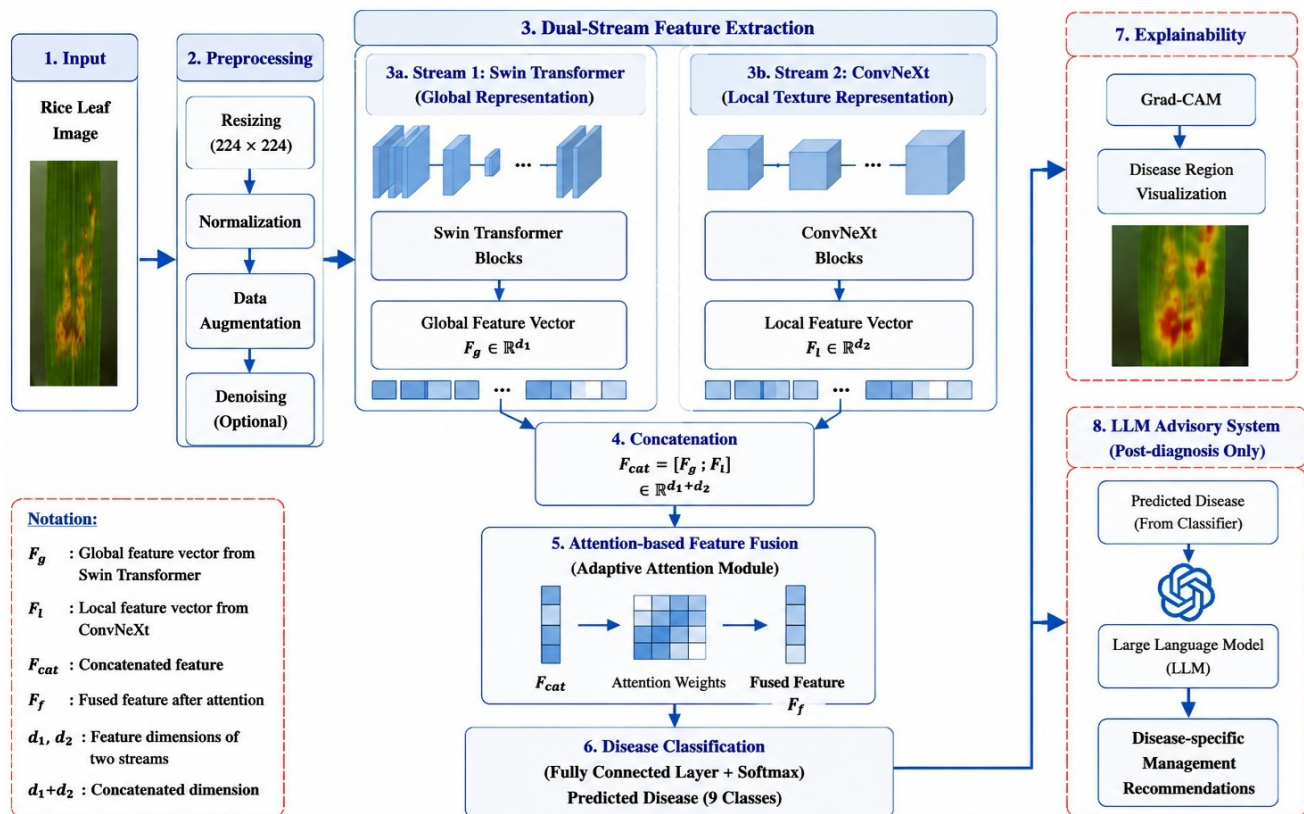
$$\begin{aligned}
 N_{\text{train}}^{(c)} &= \begin{cases} 0.70N^{(c)}, & \text{for Dataset-1,} \\ 0.60N^{(c)}, & \text{for Dataset-2,} \end{cases} \\
 N_{\text{val}}^{(c)} &= \begin{cases} 0.15N^{(c)}, & \text{for Dataset-1,} \\ 0.20N^{(c)}, & \text{for Dataset-2,} \end{cases} \\
 N_{\text{test}}^{(c)} &= \begin{cases} 0.15N^{(c)}, & \text{for Dataset-1,} \\ 0.20N^{(c)}, & \text{for Dataset-2.} \end{cases}
 \end{aligned} \tag{1}$$

where  $N^{(c)}$  represents the total number of images belonging to class  $c$ . This strategy ensures that each subset maintains similar class proportions, enabling reliable training and unbiased performance evaluation.

## 4. Methodology

The proposed framework consists of four major stages: (1) data preparation and dual-backbone complementary representation learning, (2) adaptive feature fusion and disease classification, (3) explainability and reliability assessment, and (4) post-diagnosis agricultural decision support through an LLM-based advisory module, as illustrated in Figure 7. In the first stage, the rice leaf disease dataset is prepared and divided into training, validation, and testing subsets. The proposed framework employs two parallel feature extraction streams to capture complementary disease characteristics [14]. The first stream utilizes a Swin Transformer-based architecture to learn global contextual dependencies and long-range spatial relationships across the entire leaf image. The second stream adopts ConvNeXt to extract fine-grained local representations, including lesion patterns, texture variations, and disease-specific visual characteristics. By combining these two complementary representations, the proposed framework effectively integrates global semantic information with local disease-related features.





**Figure 7.** Workflow of the proposed model in the process of rice leaf disease classification. The blue and light-blue elements represent the processing modules and feature representations, while the blue arrows show information flow. Red dashed borders distinguish the explainability and advisory modules. In Grad-CAM, red/yellow indicate higher activation, while green/blue indicate lower activation [14].

In the second stage, the extracted feature representations from the Swin Transformer and ConvNeXt streams are concatenated and processed through an attention-based feature fusion module. Unlike conventional feature concatenation approaches that assign equal importance to all extracted features, the proposed attention mechanism adaptively learns the contribution of different feature components and emphasizes the most discriminative information for disease recognition. The fused feature representation is subsequently passed through a fully connected classification layer to predict the corresponding rice leaf disease category. The performance of the proposed framework is evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC analysis. In the third stage, Grad-CAM is incorporated to improve the interpretability of the proposed model by visualizing disease-relevant regions contributing to the final prediction. This explainability analysis provides insights into the decision-making process of the model and enhances confidence in AI-assisted agricultural diagnosis. Furthermore, the robustness and generalization capability of the proposed framework are evaluated through comprehensive experiments, including five-fold cross-validation, independent cross-dataset evaluation, and comparison with state-of-the-art baseline models. In the final stage, the trained disease classification model is integrated into a web-based agricultural decision-support system. A large language model (LLM)-based advisory module is incorporated as a post-diagnosis component that generates disease-specific management recommendations based on the predicted disease category. The LLM does not participate in image feature extraction, model training, or disease classification; rather, it

functions as an application-level assistant to improve the practical usability and accessibility of the proposed framework for smart farming applications.

#### 4.1. Data Preprocessing [3]

Before being provided to the proposed dual-backbone framework, all input images undergo preprocessing operations to ensure compatibility with the backbone networks and improve model robustness against field-level variations. The preprocessing pipeline consists of image resizing, normalization, and data augmentation.

Initially, all images are resized to a fixed resolution of  $224 \times 224$  pixels to satisfy the input requirements of both the Swin Transformer and ConvNeXt architectures:

$$I' = \text{Resize}(I, H, W) \quad (2)$$

where, in Equation (2),  $I$  represents the original image and  $I'$  denotes the resized image.

The resized images are subsequently normalized using ImageNet mean and standard deviation values to improve training stability:

$$I''_c = \frac{I'_c - \mu_c}{\sigma_c}, \quad c \in \{R, G, B\} \quad (3)$$

where, in Equation (3),  $\mu_c = [0.485, 0.456, 0.406]$  and  $\sigma_c = [0.229, 0.224, 0.225]$  represent the RGB channel mean and standard deviation values.

To increase training diversity and reduce overfitting, data augmentation techniques are applied only to the training images. The augmentation process includes random horizontal flipping with a probability of 0.5, random rotation within  $[-30^\circ, 30^\circ]$ , and color jittering to simulate variations in brightness, contrast, saturation, and hue.

The augmented image can be represented as

$$\tilde{I} = T(I'') \quad (4)$$

where, in Equation (4),  $T(\cdot)$  represents the applied stochastic transformation.

The complete preprocessing pipeline is formulated as

$$\tilde{I} = \mathcal{F}_{aug}(\mathcal{F}_{norm}(\mathcal{F}_{resize}(I))) \quad (5)$$

where, in Equation (5),  $\mathcal{F}_{resize}$ ,  $\mathcal{F}_{norm}$ , and  $\mathcal{F}_{aug}$  denote resizing, normalization, and augmentation operations, respectively.

The final processed images are then forwarded into the proposed dual-stream architecture, where Swin Transformer and ConvNeXt independently extract complementary global and local disease representations.

#### 4.2. Dual-Backbone Feature Extraction

The proposed framework employs a dual-backbone complementary representation learning strategy to extract discriminative features from rice leaf images. Unlike conventional approaches that rely on a single feature extractor, the proposed method utilizes two parallel deep learning streams to capture different characteristics of disease patterns [14]. Specifically, the Swin Transformer stream is responsible for learning global contextual relationships and long-range spatial dependencies, whereas the ConvNeXt stream focuses on extracting local texture information and fine-grained disease-specific patterns.

Given an input rice leaf image  $I \in R^{H \times W \times 3}$ , the two backbone networks independently generate complementary feature representations:

$$F_{swin} = Swin(I) \quad (6)$$

$$F_{conv} = ConvNeXt(I) \quad (7)$$

where, in Equations (6) and (7),  $F_{swin}$  and  $F_{conv}$  represent the extracted feature maps from the Swin Transformer and ConvNeXt networks, respectively.

The overall motivation for using dual backbones is that disease symptoms in rice leaves contain both global and local visual characteristics. Global patterns, such as disease distribution across the entire leaf region, require contextual understanding, while local symptoms, including lesion boundaries, spots, discoloration, and texture changes, require detailed spatial feature extraction.

#### 4.2.1. Swin Transformer Stream [14,33]

The first feature extraction stream utilizes the Swin Transformer architecture to capture global contextual representations from rice leaf images. Unlike conventional convolutional networks that mainly operate on local receptive fields, the Swin Transformer employs a hierarchical transformer structure with shifted window-based self-attention, enabling effective modeling of long-range dependencies while maintaining computational efficiency.

Initially, the input image is divided into non-overlapping patches and projected into an embedding space. The patch embedding operation can be formulated as

$$Z_0 = PatchEmbed(I) \quad (8)$$

where, in Equation (8),  $Z_0$  represents the initial patch-level feature representation.

The Swin Transformer processes the embedded patches through multiple hierarchical stages. Each stage consists of shifted window multi-head self-attention (SW-MSA) and multilayer perceptron (MLP) blocks. The self-attention operation is defined as

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (9)$$

where, in Equation (9),  $Q$ ,  $K$ , and  $V$  denote query, key, and value matrices, respectively, and  $d$  represents the feature dimension.

The shifted window mechanism enables information exchange between neighboring windows, improving the capability of the model to understand spatial relationships among different disease regions.

Through hierarchical feature transformation, the Swin Transformer generates a final global feature representation:

$$F_{swin} \in R^{7 \times 7 \times C_s} \quad (10)$$

where, in Equation (10),  $C_s$  denotes the channel dimension of the final Swin Transformer feature map.

The extracted representation provides global semantic information related to disease distribution patterns and overall leaf structure.

#### 4.2.2. ConvNeXt Stream [14,34]

The second feature extraction stream employs ConvNeXt to capture local disease-specific characteristics from rice leaf images. ConvNeXt is a modern convolutional architecture that integrates several design principles inspired by transformer mod-

els, including large-kernel depth-wise convolution, layer normalization, and inverted bottleneck structures.

The ConvNeXt backbone processes the input image through multiple convolutional stages to extract hierarchical local representations. The convolution operation is expressed as

$$Y(i, j) = \sum_m \sum_n X(i + m, j + n)W(m, n) + b \quad (11)$$

where, in Equation (11),  $X$  represents the input feature map,  $W$  denotes convolution kernels, and  $b$  represents the bias term.

The depth-wise convolution operation enables efficient extraction of spatial texture patterns while reducing computational complexity. The ConvNeXt block can be represented as

$$F_{out} = LN(DWConv(F_{in}) + F_{in}) \quad (12)$$

where, in Equation (12),  $DWConv(\cdot)$  represents depth-wise convolution and  $LN(\cdot)$  denotes layer normalization.

After hierarchical feature extraction, the final ConvNeXt representation is obtained as

$$F_{conv} \in R^{7 \times 7 \times C_c} \quad (13)$$

where, in Equation (13),  $C_c$  represents the feature channel dimension.

The ConvNeXt stream provides detailed local representations, allowing the framework to effectively recognize fine-grained disease characteristics such as lesion shapes, texture variations, and color changes.

#### 4.3. Attention-Based Feature [14,35]

After extracting complementary representations from the Swin Transformer and ConvNeXt streams [14], an effective feature fusion [17] strategy is required to integrate global contextual information and local disease-specific characteristics. A simple feature concatenation approach treats all extracted features equally and may introduce redundant or less discriminative information. To overcome this limitation, this study proposes an attention [16]-based feature fusion module that adaptively learns the contribution of each backbone representation during the training process.

Initially, the extracted feature maps from both streams are converted into compact feature vectors using Global Average Pooling (GAP). The GAP operation for a feature map  $F \in R^{H \times W \times C}$  is defined as

$$f_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_{i,j,c} \quad (14)$$

where, in Equation (14),  $f_c$  represents the output feature value of channel  $c$ , and  $H$  and  $W$  denote the spatial dimensions of the feature map.

After applying GAP, the Swin Transformer and ConvNeXt features are represented as

$$f_{swin} = GAP(F_{swin}) \quad (15)$$

$$f_{conv} = GAP(F_{conv}) \quad (16)$$

where, in Equation (15),  $f_{swin}$  represents the global contextual feature vector obtained from the Swin Transformer, while, in Equation (16),  $f_{conv}$  represents the local texture feature vector extracted from ConvNeXt.

The Swin Transformer and ConvNeXt backbones extract complementary feature representations from the input rice leaf images. The Swin Transformer provides hierarchical self-attention to capture the global contextual information, while ConvNext focuses on fine-

grained local texture and spatial patterns. To combine these complementary representations, we use an attention-based feature fusion module.

The proposed module aims to undertake adaptable weighting at the Backbone-Branch level. Specifically, two learnable scalar parameters,  $\omega_1$  and  $\omega_2$ , are introduced to determine the relative contributions of the Swin Transformer and ConvNeXt feature representations, respectively. The following parameters are normalized with a softmax function to calculate branch-wise corresponding fusion parameters:

$$\alpha_i = \frac{e^{\omega_i}}{\sum_{j=1}^2 e^{\omega_j}}, \quad i \in 1, 2 \quad (17)$$

where, in Equation (17),  $\alpha_1$  and  $\alpha_2$  represent the normalized fusion weights assigned to the Swin Transformer and ConvNeXt feature representations, respectively. The softmax operation ensures that the coefficients are positive and satisfy the following:

$$\alpha_1 + \alpha_2 = 1 \quad (18)$$

Let  $F_{\text{Swin}}$  and  $F_{\text{ConvNeXt}}$  denote the feature representations extracted by the Swin Transformer and ConvNeXt backbones, respectively. The learned coefficients are applied to the corresponding feature representations prior to concatenation:

$$F_{\text{fused}} = [\alpha_1 F_{\text{Swin}}, \parallel, \alpha_2 F_{\text{ConvNeXt}}], \quad (19)$$

where, in Equation (19),  $\parallel$  denotes the feature concatenation operation. The weights of the backbone branches are then learned, so that the resulting  $F_{\text{fused}}$  representation is a combined representation of the two complementary backbone representations according to the learned branch-level weights.

The learnable branch-level weighting system enables the model to tune the contribution from the two feature representations during training. A larger value of  $\alpha_1$  increases the contribution of the global contextual representation extracted by the Swin Transformer, whereas a larger value of  $\alpha_2$  increases the contribution of the fine-grained local representation extracted by ConvNeXt. Hence, the proposed module offers a simple yet easy-to-learn mechanism for balancing the complementary representations of features before the last classification stage.

A crucial feature of the proposed fusion mechanism is that the weights are common to all input samples. The parameters  $\omega_1$  and  $\omega_2$  are global learnable parameters within the fusion module and are optimized during model training. This means that the resulting coefficients  $\alpha_1$  and  $\alpha_2$  are not computed individually for each input image and its feature representation. Instead, the learned coefficients are then applied to the corresponding Swin Transformer and ConvNeXt feature streams for all the input samples. Hence, the proposed mechanism is not an attention mechanism that is dependent on the input, but is a learnable branch-weighting strategy that can be applied globally.

It is important to clarify that the proposed attention-based feature fusion module is not a conventional spatial- or channel-wise attention mechanism, and attention-based feature fusion in this work refers to the learnable branch-level weighting of the two backbone feature representations. The two backbone branches get their own respective scalar attention weights,  $\alpha_1$  and  $\alpha_2$ . In this work, the phrase attention-based feature fusion is used specifically for the usage of learnable weighting of the two feature representations at the backbone level, instead of using the traditional spatial attention or channel attention.

The proposed fusion mechanism makes the network not only adaptively weight the global contextual representation learned by Swin Transformer and the local fine-grained representation from ConvNeXt, but also learn the relative importance of the two feature



streams. This complementary feature integration is designed to create a more discriminative, fused representation for rice leaf disease classification, especially for disease classes where the leaf disease symptoms are similar.

#### 4.4. Classification Module

The fused feature representation generated by the attention-based feature fusion module is forwarded to a fully connected classification network for final rice disease prediction. The classification module is designed as a multilayer perceptron (MLP) consisting of fully connected layers, batch normalization, ReLU activation, and dropout regularization.

#### 4.5. Grad-CAM Explainability [10,26]

To improve the interpretability of the proposed framework, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to visualize the image regions that contribute most significantly to the model prediction. Grad-CAM provides a heatmap representation by using the gradients of the predicted class with respect to the final convolutional feature maps.

The importance weight of each feature channel is calculated as

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (20)$$

where, in Equation (20),  $y^c$  represents the predicted score for class  $c$ ,  $A^k$  denotes the  $k$ -th feature map, and  $Z$  represents the number of spatial locations.

The Grad-CAM heatmap is generated by combining the weighted feature maps:

$$L_{Grad-CAM}^c = ReLU \left( \sum_k \alpha_k^c A^k \right) \quad (21)$$

where, in Equation (21),  $L_{Grad-CAM}^c$  represents the class-specific activation map. The generated heatmaps highlight disease-related regions of rice leaves, providing visual evidence for model decisions and improving trustworthiness in AI-assisted agricultural diagnosis.

#### 4.6. LLM-Based Advisory System [11,12,27]

To extend the practical usability of the proposed framework, a Large Language Model (LLM)-based advisory module is integrated as a post-diagnosis decision-support component. The LLM is used solely as an informational assistance tool and is not involved in the image feature extraction, disease classification, or final decision-making process. The LLM receives the predicted disease category from the classification module and generates disease-specific agricultural recommendations, including possible management strategies and preventive measures.

The advisory process can be represented as

$$R_{adv} = LLM(P, D_p) \quad (22)$$

where, in Equation (22),  $P$  represents the generated prompt,  $D_p$  denotes the predicted disease category, and  $R_{adv}$  represents the generated recommendation.

The LLM module operates only at the application level and does not participate in image feature extraction, model training, or disease classification. Its primary purpose is to transform AI-based diagnosis results into understandable recommendations, improving the accessibility and practical deployment of the proposed smart farming system.

## 5. Results and Analysis

The proposed dual-backbone complementary representation learning framework for rice leaf disease classification is evaluated extensively in this section. The proposed framework integrates the Swin Transformer and the fine-grained local texture extraction ability of ConvNeXt. Furthermore, an attention-based feature fusion module is employed to adaptively combine these complementary representations and enhance discriminative feature learning for accurate disease classification.

### 5.1. Experimental Setup and Evaluation Protocol

To ensure a comprehensive and unbiased evaluation, the proposed framework was assessed using multiple experimental protocols, including standard train-validation-test evaluation, ablation study, cross-validation, independent cross-dataset evaluation, confusion matrix analysis, ROC-AUC analysis, and explainability evaluation using Grad-CAM. Additionally, the trained classification framework was integrated into a real-time web-based agricultural decision-support system, where a Large Language Model (LLM) was incorporated as a post-diagnosis advisory module to provide disease-specific management recommendations. The LLM does not participate in feature extraction or disease classification; rather, it operates only at the application level to enhance the practical usability of the proposed framework. The overall evaluation workflow is illustrated in Figure 7.

The performance of the proposed dual-backbone complementary representation learning framework was compared with four representative baseline architectures: Swin Transformer, ConvNeXt, EfficientNetV2, and ResNet50. Two publicly available rice leaf disease datasets, referred to as Dataset-1 and Dataset-2, were used for evaluation. Dataset-1 contains 11,790 RGB images, while Dataset-2 consists of 11,250 RGB images collected from agricultural environments. To maintain consistency across all experiments, all input images were resized to a fixed spatial resolution of  $224 \times 224$  pixels before training and evaluation.

The classification performance of all models was evaluated using standard metrics, including accuracy, precision, recall, F1-score, confusion matrix, and area under the receiver operating characteristic curve (ROC-AUC). These evaluation criteria provide a comprehensive assessment of classification capability, robustness, and generalization performance of the proposed framework compared with existing approaches.

### 5.2. Performance Evaluation on Dataset-1

The performance of the proposed dual-backbone complementary representation learning framework was first evaluated on Dataset-1 to assess its effectiveness in rice leaf disease classification. The evaluation includes ablation study, class-wise performance investigation, learning curve analysis, and confusion matrix evaluation. These experiments aim to validate the contribution of the proposed dual-stream feature extraction and attention-based feature fusion strategy in improving classification accuracy and robustness.

#### 5.2.1. Baseline Comparison on Dataset-1

Table 3 presents the performance comparison between the proposed dual-backbone complementary representation learning framework and state-of-the-art baseline models, including EfficientNetV2, ConvNeXt-Tiny, ResNet50, and Swin Transformer, under identical experimental conditions. The evaluation was conducted using accuracy, precision, recall, F1-score, and validation loss. The proposed framework achieved the highest classification accuracy of 99.55%, then ConvNeXt-Tiny (98.69%), Swin Transformer (98.30%), EfficientNetV2 (98.24%), and ResNet50 (97.90%). Similarly, the proposed model obtained the highest F1-score of 99.56% with the lowest validation loss of 0.493, indicating improved classification capability and optimization stability.

**Table 3.** Performance comparison of the baseline models and the proposed dual-backbone framework on Dataset-1.

| Model            | Accuracy (%) | Precision | Recall | F1-Score (%) | Validation Loss |
|------------------|--------------|-----------|--------|--------------|-----------------|
| EfficientNet-VS2 | 98.24        | 98.22     | 98.44  | 98.33        | 0.528           |
| ConvNext-tny     | 98.69        | 98.77     | 98.77  | 98.77        | 0.519           |
| ResNet50         | 97.90        | 97.88     | 97.33  | 97.88        | 0.547           |
| Swin Tranformer  | 98.30        | 98.22     | 98.44  | 98.33        | 0.526           |
| Proposed Model   | 99.55        | 99.50     | 99.63  | 99.56        | 0.493           |

The improved performance can be attributed to the complementary representation learning strategy. Specifically, the Swin Transformer branch captures global contextual dependencies across the entire leaf image, while the ConvNeXt branch extracts fine-grained local disease patterns, including lesion textures and color variations. The attention-based feature fusion module adaptively integrates these complementary representations by assigning higher importance to more discriminative features, resulting in improved disease classification. Although the benchmark evaluation demonstrates the effectiveness of the proposed framework, performance on a single dataset may not fully represent real-world generalization capability. Therefore, additional experiments, including cross-validation and independent cross-dataset evaluation, are conducted in subsequent sections to verify robustness under unseen data distributions.

### 5.2.2. Ablation Study

To investigate the contribution of each component of the proposed framework, an ablation study was conducted by gradually increasing the model complexity. The evaluation began with individual backbone networks, followed by simple feature concatenation and finally the proposed attention-based feature fusion strategy.

Table 4 presents the performance of different model configurations. When ConvNeXt and Swin Transformer were used independently, they achieved accuracies of 98.69% and 98.30%, respectively. These results demonstrate that each backbone can effectively learn disease-related visual patterns; however, each architecture captures different aspects of image representation. Combining both backbones through simple feature concatenation improved the accuracy to 98.75%, confirming that global contextual information and local texture characteristics provide complementary information for disease recognition. However, traditional concatenation assumes all feature components are of equal importance and does not differentiate between informative and less discriminative representations. In order to further highlight the role of the proposed attention-based feature fusion module, an equal-weight fusion experiment was performed by setting the weight of both the feature streams of Swin Transformer and ConvNeXt to 0.5. As shown in Table 4, the equal-weight fusion achieved an accuracy of 99.20% and an F1-score of 99.21%, which is higher than simple feature concatenation. After introducing the proposed attention-based feature fusion module, the accuracy increased further to 99.55% with an F1-score of 99.56%. This improvement demonstrates that the attention mechanism effectively learns feature importance and selectively emphasizes disease-relevant representations. Therefore, the ablation results validate that both complementary dual-backbone learning and adaptive feature fusion contribute significantly to the final classification performance.

**Table 4.** Ablation study of the proposed framework.

| Model Variant                   | Swin | ConvNeXt | Fusion | Attention | Accuracy (%) | F1-Score (%) |
|---------------------------------|------|----------|--------|-----------|--------------|--------------|
| ConvNeXt (Only)                 | ×    | ✓        | ×      | ×         | 98.69        | 98.77        |
| Swin (Only)                     | ✓    | ×        | ×      | ×         | 98.30        | 98.33        |
| Swin + ConvNeXt (Concatenation) | ✓    | ✓        | ✓      | ×         | 98.75        | 98.79        |
| Swin + ConvNeXt (Equal-Weight)  | ✓    | ✓        | ✓      | ×         | 99.20        | 99.21        |
| Proposed                        | ✓    | ✓        | ✓      | ✓         | 99.55        | 99.56        |

### 5.2.3. Class-Wise Performance Analysis on Dataset-1

Beyond overall performance metrics, class-wise evaluation was conducted to investigate the classification capability of the proposed framework across individual rice disease categories. The precision, recall, and F1-score values for each class are presented in Table 5.

**Table 5.** Per-class performance of the proposed framework on the rice leaf disease dataset.

| Class | Disease Category       | Precision | Recall | F1-Score |
|-------|------------------------|-----------|--------|----------|
| c0    | Bacterial Leaf Blight  | 0.9894    | 1.0000 | 0.9947   |
| c1    | Brown Spot             | 1.0000    | 0.9954 | 0.9977   |
| c2    | Healthy Rice Leaf      | 1.0000    | 1.0000 | 1.0000   |
| c3    | Leaf Blast             | 0.9964    | 0.9892 | 0.9928   |
| c4    | Leaf Scald             | 0.9950    | 0.9900 | 0.9925   |
| c5    | Narrow Brown Leaf Spot | 0.9779    | 1.0000 | 0.9888   |
| c6    | Neck Blast             | 1.0000    | 1.0000 | 1.0000   |
| c7    | Rice Hispa             | 1.0000    | 1.0000 | 1.0000   |
| c8    | Sheath Blight          | 0.9960    | 0.9921 | 0.9941   |

The proposed framework achieved consistently high performance across all nine disease categories, with F1-scores ranging from 0.9800 to 1.0000. Particularly, Brown Spot(c1), Healthy Rice Leaf(c2), Neck Blast(c6) and, Rice Hispa(c7) achieved perfect classification performance with precision, recall, and F1-score values of 1.00.

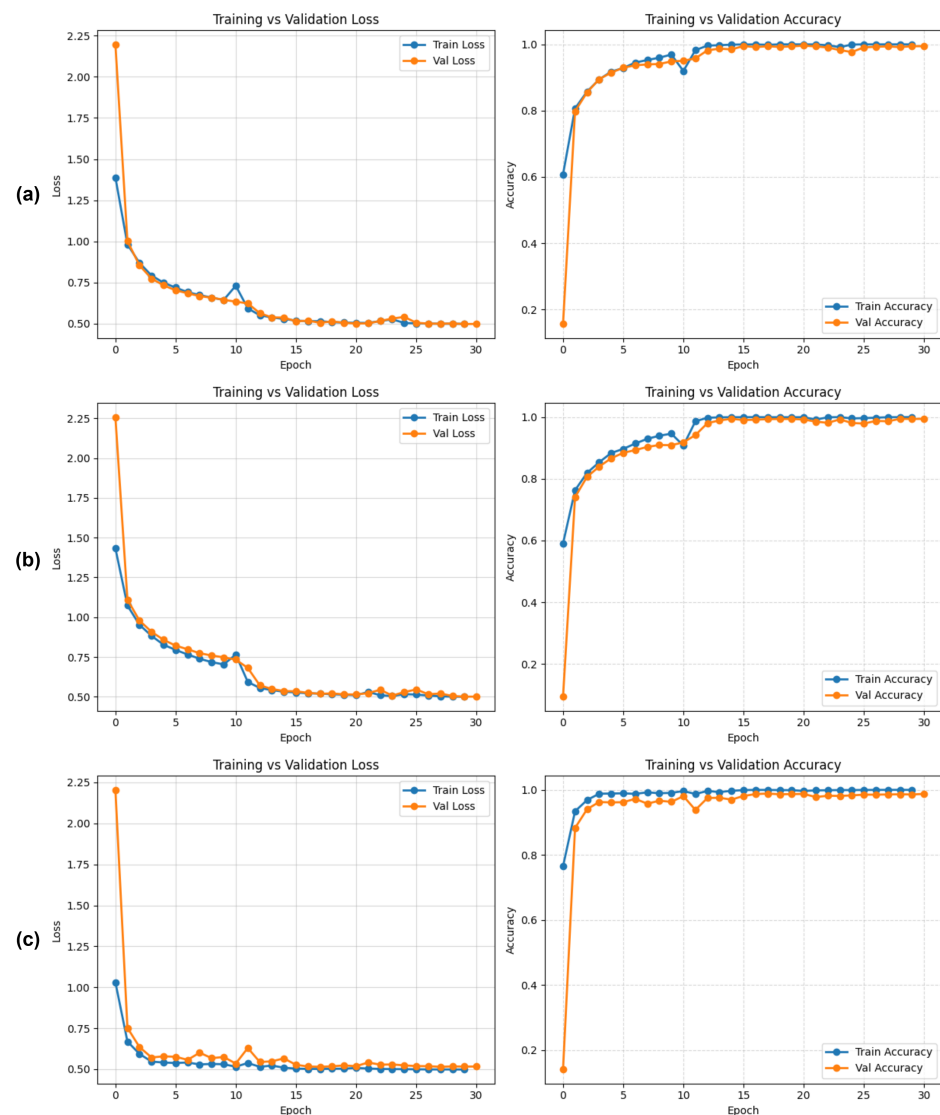
The remaining disease categories also achieved highly balanced results. Bacterial Leaf Blight(c0), Leaf Blast(c3), Leaf Scald(c4) and Sheath Blight(c8) obtained F1-scores of 0.99, demonstrating the ability of the proposed framework to distinguish visually similar disease patterns.

Minor performance variations were observed for Narrow Brown Leaf Spot(c5). These variations are mainly associated with similarities in visual appearance, including overlapping texture patterns, lesion structures, and color characteristics among different categories.

Overall, the class-wise results demonstrate that the proposed framework provides balanced classification performance without bias toward specific disease categories.

### 5.2.4. Learning Curve Analysis for Dataset-1

The training and validation loss and accuracy curves of the proposed framework and baseline models are illustrated in Figure 8. These curves are analyzed to evaluate optimization stability, convergence behavior, and generalization capability during training.



**Figure 8.** Training and validation loss and accuracy curves of (a) the proposed framework, (b) ConvNeXt, and (c) Swin Transformer for rice leaf disease classification.

The proposed framework shows faster convergence and smoother optimization behavior compared with individual Swin Transformer and ConvNeXt models. Both training and validation curves demonstrate consistent improvement with limited fluctuation, indicating stable learning and reduced optimization uncertainty.

The lower validation loss achieved by the proposed framework (0.493) compared with the baseline models further confirms improved feature learning capability. This stable convergence is mainly attributed to the attention-based fusion mechanism, which effectively combines global semantic information from Swin Transformer with local disease-specific representations from ConvNeXt.

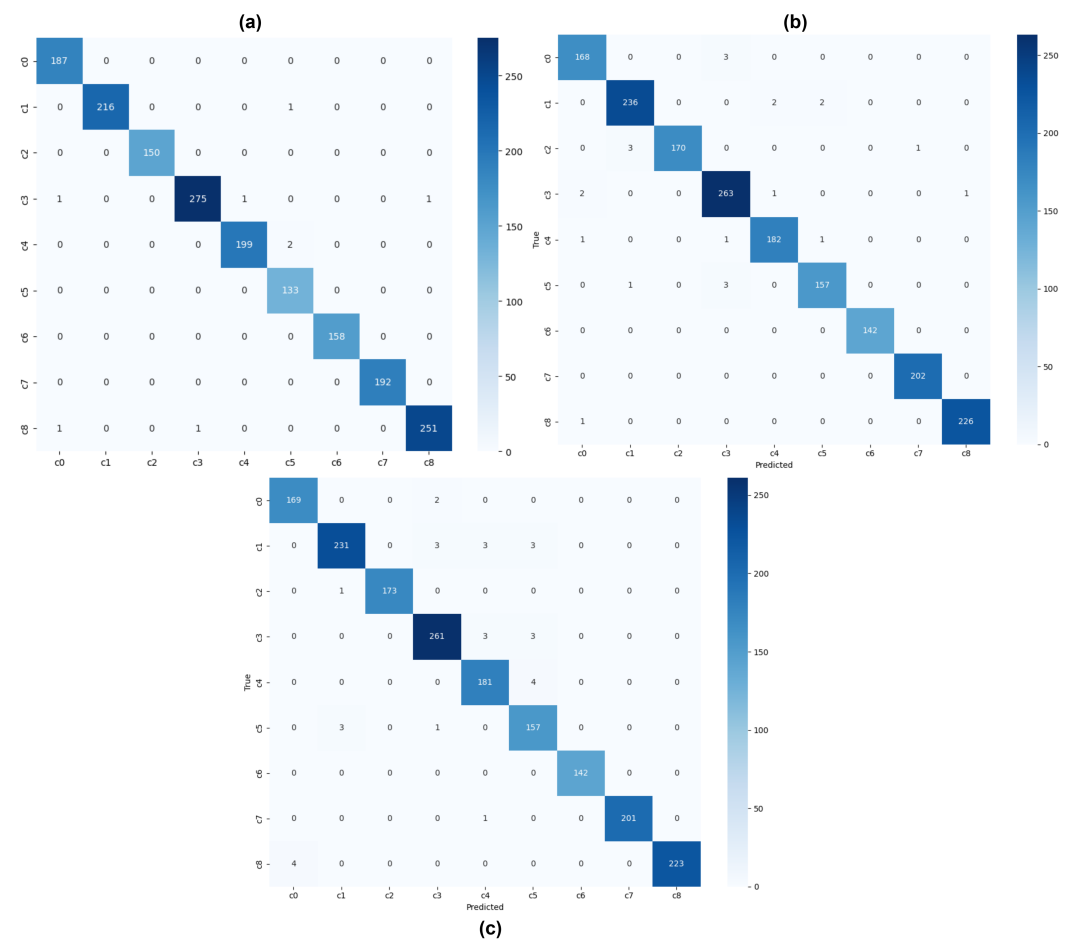
Therefore, the learning curves provide additional evidence that the proposed complementary representation learning strategy enables more robust and discriminative feature extraction.

#### 5.2.5. Confusion Matrix Analysis for Dataset-1

The confusion matrices of the proposed framework, ConvNeXt-Tiny, and Swin Transformer are presented in Figure 9. The diagonal elements represent correctly clas-



sified samples, whereas off-diagonal elements indicate classification errors between disease categories.



**Figure 9.** Confusion matrices of (a) the proposed framework, (b) ConvNeXt-Tiny, and (c) Swin Transformer for rice leaf disease classification.

As shown in Figure 9a, the proposed framework demonstrates strong diagonal dominance, indicating accurate classification across almost all disease categories. Only a small number of misclassifications were observed, mainly among categories with visually similar symptoms, such as Bacterial Leaf Blight, Brown Spot, Narrow Brown Leaf Spot, and Sheath Blight. Compared with the baseline models, ConvNeXt-Tiny and Swin Transformer show relatively higher off-diagonal responses, indicating increased inter-class confusion. Although both models successfully capture important disease extraction characteristics, their individual representations are limited either by local feature extraction or global contextual understanding. The proposed framework reduces such confusion by integrating complementary representations from both architectures. Swin Transformer provides global structural information, whereas ConvNeXt captures detailed local disease patterns. The attention-based fusion mechanism further enhances feature discrimination by emphasizing the most relevant disease characteristics. The confusion matrix analysis is consistent with the quantitative results reported in Tables 3 and 5, confirming the effectiveness of the proposed framework in achieving reliable multi-class rice disease classification.

### 5.3. Performance Evaluation on Dataset-2

To further validate the robustness and generalization capability of the proposed dual-backbone complementary representation learning framework, additional experiments were conducted on Dataset-2. Compared with Dataset-1, Dataset-2 contains a larger number of

disease categories with increased intra-class variation and inter-class similarity, making the classification task more challenging. The evaluation includes baseline comparison, class-wise performance analysis, learning curve investigation, and confusion matrix analysis to comprehensively assess the effectiveness of the proposed framework.

### 5.3.1. Baseline Comparison on Dataset-2

Table 6 presents the performance comparison between the proposed framework and representative state-of-the-art deep learning models, including EfficientNetV2, ConvNeXt-Tiny, ResNet50, and Swin Transformer, under identical experimental settings. The comparison was performed using accuracy, precision, recall, F1-score, and validation loss to provide a comprehensive evaluation of classification performance.

**Table 6.** Performance comparison of the baseline models and the proposed dual-backbone framework on Dataset-2.

| Model            | Accuracy (%) | Precision | Recall | F1-Score (%) | Validation Loss |
|------------------|--------------|-----------|--------|--------------|-----------------|
| EfficientNet-VS2 | 95.75        | 95.61     | 94.87  | 95.23        | 0.645           |
| ConvNext-tny     | 96.26        | 96.12     | 95.84  | 95.97        | 0.638           |
| ResNet50         | 95.36        | 95.21     | 94.78  | 94.99        | 0.651           |
| Swin Tranformer  | 96.31        | 95.46     | 96.14  | 95.73        | 0.635           |
| Proposed Model   | 97.06        | 96.70     | 96.82  | 96.75        | 0.598           |

The proposed framework achieved the highest accuracy of 97.06%, outperforming Swin Transformer (96.31%), ConvNeXt-Tiny (96.26%), EfficientNetV2 (95.75%), and ResNet50 (95.36%). Similarly, the proposed model achieved the highest F1-score of 96.75% and the lowest validation loss of 0.598, indicating improved classification reliability and optimization stability.

The performance improvement can be attributed to the complementary feature learning strategy adopted in the proposed framework. The Swin Transformer branch captures global contextual relationships and overall disease distribution patterns, while the ConvNeXt branch focuses on local disease-specific characteristics, including lesion textures, color variations, and fine-grained visual patterns. Furthermore, the attention-based feature fusion module adaptively emphasizes the most informative feature representations, allowing the model to generate more discriminative features for distinguishing visually similar rice disease categories.

### 5.3.2. Class-Wise Performance Analysis on Dataset-2

To analyze the classification capability of the proposed framework across individual disease categories, a class-wise performance evaluation was conducted using precision, recall, and F1-score. This analysis is particularly important for Dataset-2 because it contains twelve disease categories with complex visual variations and overlapping symptoms. The detailed class-wise performance results are presented in Table 7.

As shown in Table 7, the proposed framework achieved consistently high performance across all twelve disease categories, with F1-scores ranging from 0.9398 to 0.9983. The Dead Heart class achieved the highest performance with an F1-score of 0.9983, demonstrating the ability of the proposed framework to effectively capture distinctive disease characteristics.

**Table 7.** Per-class performance of the proposed framework on Dataset-2.

| Class | Disease Category         | Precision | Recall | F1-Score |
|-------|--------------------------|-----------|--------|----------|
| c0    | Bacterial Leaf Blight    | 0.9592    | 0.9895 | 0.9741   |
| c1    | Bacterial Leaf Streak    | 0.9758    | 0.9680 | 0.9719   |
| c2    | Bacterial Panicle Blight | 0.9565    | 0.9851 | 0.9706   |
| c3    | Blast                    | 0.9659    | 0.9798 | 0.9728   |
| c4    | Brown Spot               | 0.9793    | 0.9793 | 0.9793   |
| c5    | Dead Heart               | 1.0000    | 0.9965 | 0.9983   |
| c6    | Downy Mildew             | 0.9360    | 0.9435 | 0.9398   |
| c7    | Hispa                    | 0.9710    | 0.9465 | 0.9586   |
| c8    | Normal                   | 0.9529    | 0.9773 | 0.9649   |
| c9    | Rice Sheath Blight       | 0.9400    | 0.9400 | 0.9400   |
| c10   | Smut                     | 0.9726    | 0.9595 | 0.9660   |
| c11   | Tungro                   | 0.9952    | 0.9539 | 0.9741   |

Most disease categories achieved F1-scores above 0.96, including Bacterial Leaf Blight, Bacterial Leaf Streak, Bacterial Panicle Blight, Blast, Brown Spot, Smut, and Tungro. These results indicate that the proposed framework can effectively learn diverse disease patterns despite variations in appearance.

Relatively lower performance was observed for Downy Mildew and Rice Sheath Blight, with F1-scores of 0.9398 and 0.9400, respectively. These reductions are mainly associated with visual similarities in lesion appearance, texture patterns, and background variations within field-acquired images. However, the balanced precision and recall values demonstrate that the proposed framework maintains reliable classification capability even for challenging disease categories.

Overall, the class-wise analysis confirms that the proposed dual-backbone framework provides robust and balanced recognition performance across diverse rice disease categories.

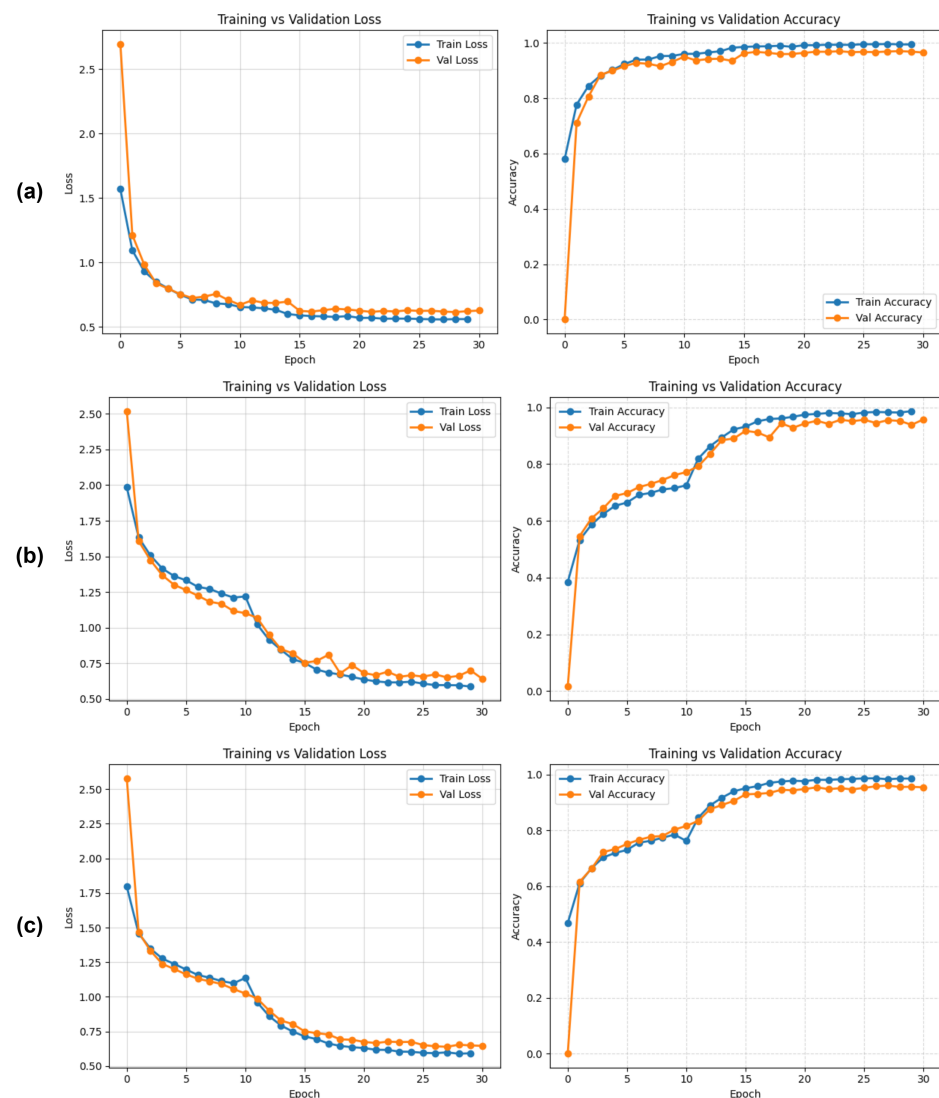
### 5.3.3. Learning Curve Analysis for Dataset-2

The training and validation accuracy and loss curves of the proposed framework and baseline models on Dataset-2 are presented in Figure 10. The learning curves are analyzed to investigate convergence behavior, optimization stability, and generalization capability during the training process.

As shown in Figure 10a, the proposed framework demonstrates smooth and stable convergence throughout the training process. Both training and validation losses gradually decrease and reach a stable state after approximately 15 epochs, while the accuracy curves continuously improve with limited fluctuation. The small gap between training and validation performance indicates effective learning without significant overfitting.

Compared with ConvNeXt-Tiny and Swin Transformer shown in Figure 10b,c, the proposed framework exhibits more stable optimization behavior and faster convergence. Although both baseline models achieve competitive performance, their validation loss curves show comparatively larger fluctuations, indicating less stable feature optimization.

The improved convergence behavior is mainly attributed to the attention-based feature fusion mechanism, which effectively integrates complementary global and local representations. By dynamically selecting informative features from Swin Transformer and ConvNeXt, the proposed framework achieves more discriminative feature learning and improved optimization stability.



**Figure 10.** Training and validation loss and accuracy curves of (a) the proposed framework, (b) ConvNeXt-Tiny, and (c) Swin Transformer on Dataset-2.

The quantitative results in Table 6 further support these observations, where the proposed framework achieves the highest accuracy, F1-score, and lowest validation loss among all evaluated models.

#### 5.3.4. Confusion Matrix Analysis for Dataset-2

The confusion matrices of the proposed framework, ConvNeXt-Tiny, and Swin Transformer on Dataset-2 are presented in Figure 11. The confusion matrix provides detailed insight into class-level prediction behavior, where diagonal values represent correctly classified samples and off-diagonal values indicate misclassification among disease categories.

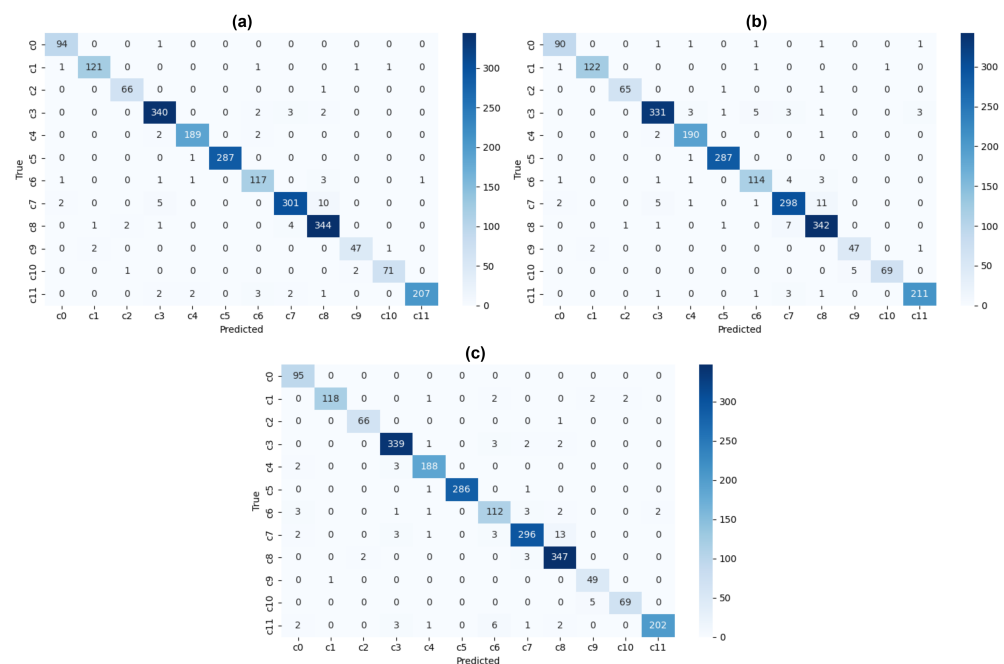
As shown in Figure 11a, the proposed framework demonstrates strong diagonal dominance across all twelve disease categories, indicating reliable classification performance. Only a limited number of samples are misclassified, mainly among visually similar disease categories such as Downy Mildew, Hispa, Normal, and Rice Sheath Blight. These errors are expected because several rice diseases share similar lesion structures, color patterns, and environmental effects in field images.

Compared with the proposed framework, the confusion matrices of ConvNeXt-Tiny and Swin Transformer contain relatively higher off-diagonal responses, indicating increased inter-class confusion. Although these baseline models successfully capture important dis-

ease features, their individual representation capabilities are limited because they primarily focus on either local texture information or global contextual relationships.

The lower confusion rate achieved by the proposed framework is mainly due to the complementary representation learning strategy. The Swin Transformer captures long-range contextual dependencies, whereas ConvNeXt extracts fine-grained local disease characteristics. The attention-based fusion module further enhances feature discrimination by adaptively emphasizing the most relevant disease representations, resulting in improved separation between visually similar disease classes.

The confusion matrix analysis is consistent with the quantitative results reported in Table 6, confirming the effectiveness of the proposed framework for robust rice leaf disease classification under complex image conditions.



**Figure 11.** Confusion matrices of (a) the proposed framework, (b) ConvNeXt-Tiny, and (c) Swin Transformer on Dataset-2.

#### 5.4. Robustness and Generalization Evaluation

To evaluate the robustness and generalization capability of the proposed framework beyond standard test-set performance, additional experiments were conducted using five-fold cross-validation and independent cross-dataset evaluation. These experiments aim to investigate the stability of the model under different data partitions and its ability to maintain performance when exposed to unseen data distributions.

##### 5.4.1. Five-Fold Cross-Validation

Dataset-1 contains 11,790 RGB images belonging to nine rice leaf disease categories. To evaluate the stability and reliability of the proposed framework, five-fold cross-validation was performed. In this experiment, the dataset was divided into five subsets, where each subset was used as validation data once while the remaining subsets were used for training. The final performance was obtained by averaging the results across all five folds. The five-fold cross-validation results presented in Table 8 demonstrate that the proposed framework achieves consistent performance across different data partitions. The model obtained an average accuracy of  $98.91\% \pm 0.47\%$  and an average F1-score of  $99.12\% \pm 0.42\%$ . The small variation among different folds indicates that the proposed framework is not dependent on a particular training-testing split and provides stable feature learning capability.



**Table 8.** Five-fold cross-validation results of the proposed framework on the Kaggle rice leaf disease dataset.

| Fold    | Accuracy (%) | F1-Score (%) |
|---------|--------------|--------------|
| 1       | 98.47        | 98.48        |
| 2       | 99.02        | 99.05        |
| 3       | 99.19        | 99.21        |
| 4       | 98.38        | 99.36        |
| 5       | 99.48        | 99.49        |
| Average | 98.91 ± 0.47 | 99.12 ± 0.42 |

#### 5.4.2. Independent Cross-Dataset Evaluation

To further investigate the practical generalization capability of the proposed framework, an independent cross-dataset evaluation was performed. For independent cross-dataset evaluation, a class label space was aligned with Dataset-1 (with nine classes representing nine disease categories). Dataset-2 contains 12 categories, but not all of them correspond to the output classes of the model trained on Dataset-1. As a result, only 9 classes were subsequently chosen for cross-dataset evaluation, that is, those classes related to the Dataset-1 label space. For a nine-class evaluation set, six of these classes were already in Dataset-2, but the other three were also collected from Kaggle. To ensure cross-dataset evaluation, categories such as Dead Heart, Downy Mildew, Smut, and Tungro that were not present in the Dataset-1 label space were excluded. Therefore, the model was tested with the same 9-class output space used in Dataset-1, and the performance measures were computed for the 9 classes on Dataset-1 instead of the 12 classes in Dataset-2. This evaluation scenario introduces domain variation in terms of illumination, background complexity, image acquisition conditions, and disease appearance, providing a more realistic assessment of model robustness.

#### 5.4.3. Baseline Comparison Under Cross-Dataset Evaluation

Table 9 presents the performance comparison of the proposed framework with baseline architectures under independent cross-dataset evaluation. The proposed framework achieved the highest accuracy of 90.80%, outperforming ConvNeXt-Tiny (88.81%), Swin Transformer (86.46%), ResNet50 (85.06%), and EfficientNetV2 (84.82%).

The proposed framework also achieved the highest precision (87.22%), recall (84.88%), and F1-score (86.03%), demonstrating improved discrimination capability under unseen data conditions. The performance improvement can be attributed to the complementary representation learning strategy, where Swin Transformer captures global disease structures while ConvNeXt extracts fine-grained local texture patterns. The attention-based feature fusion module adaptively integrates these complementary representations, enabling the model to maintain reliable classification performance despite domain variations.

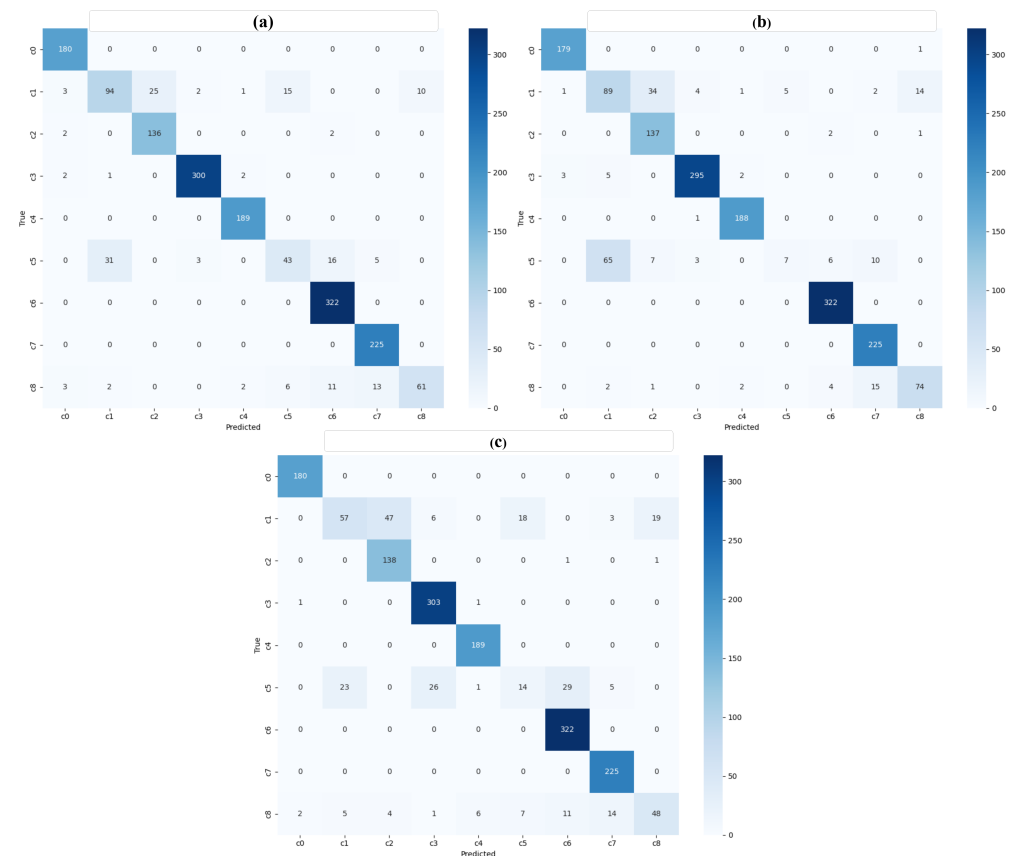
**Table 9.** Performance comparison of the baseline models and the proposed dual-backbone framework for rice leaf disease classification using standard evaluation metrics.

| Model            | Test Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|------------------|-------------------|---------------|------------|--------------|
| EfficientNet-VS2 | 84.82             | 78.88         | 75.66      | 77.24        |
| ConvNext-tny     | 88.81             | 83.22         | 81.66      | 82.43        |
| ResNet50         | 85.06             | 76.77         | 75.88      | 76.32        |
| Swin Tranformer  | 86.46             | 79.00         | 77.66      | 78.32        |
| Proposed Model   | 90.80             | 87.22         | 84.88      | 86.03        |

The obtained results indicate that the proposed framework maintains strong performance even when evaluated on an unseen dataset with different image characteristics. This demonstrates its capability to learn generalized disease representations rather than dataset-specific patterns, which is essential for real-world agricultural deployment.

#### 5.4.4. Cross-Dataset Confusion Matrix Analysis

To further analyze class-level generalization capability, confusion matrices obtained from the independent cross-dataset evaluation are presented in Figure 12. Unlike conventional test evaluation, this analysis reflects the model's ability to distinguish disease categories under domain shift without retraining or fine-tuning.



**Figure 12.** Confusion matrices obtained under independent cross-dataset evaluation: (a) Proposed framework, (b) ConvNeXt-Tiny, and (c) Swin Transformer.

As shown in Figure 12a, the proposed framework demonstrates stronger diagonal dominance compared with baseline models, indicating improved class-wise recognition capability on unseen samples. The limited off-diagonal responses suggest reduced inter-class confusion, although some misclassification remains among visually similar disease categories.

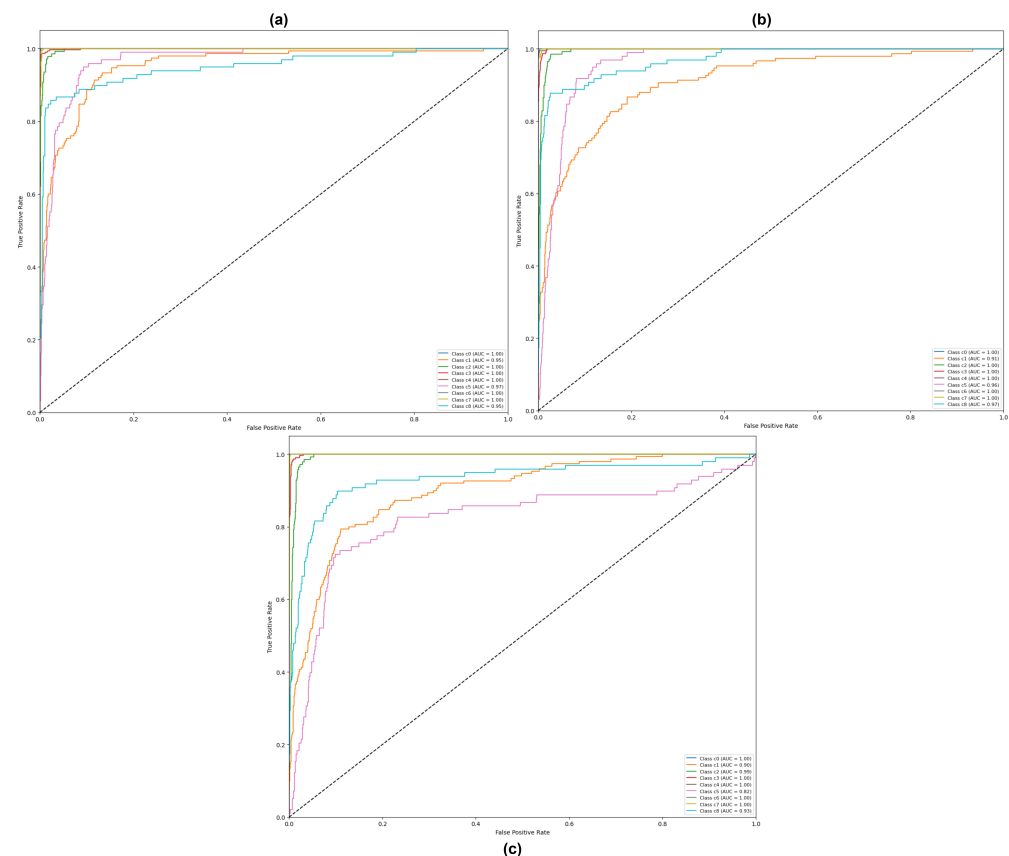
The baseline models exhibit relatively higher off-diagonal values, particularly for diseases with similar lesion patterns and visual characteristics. These errors are mainly associated with domain variations, including differences in illumination, background, camera viewpoint, and disease severity.

The improved confusion matrix performance of the proposed framework is attributed to the complementary feature learning mechanism. The combination of global contextual information from Swin Transformer and local disease texture features from ConvNeXt provides richer representations, while attention-based fusion enhances feature discrimination under complex field conditions.

#### 5.4.5. ROC-AUC Analysis

ROC-AUC analysis was additionally performed to evaluate the class discrimination capability of the proposed framework. Unlike accuracy-based evaluation, ROC-AUC provides a threshold-independent measurement of model separability between disease categories.

As illustrated in Figure 13, the proposed framework achieves higher AUC performance compared with the baseline models, indicating stronger discrimination capability across different rice disease categories. These findings further confirm that the proposed complementary representation learning strategy improves robustness and generalization under diverse imaging conditions.



**Figure 13.** ROC curves and AUC scores of (a) the proposed framework, (b) ConvNeXt-Tiny, and (c) Swin Transformer for rice leaf disease classification. The dotted diagonal line represents the random-classifier baseline.

#### 5.5. Comparison with State-of-the-Art Methods

Although the baseline comparison demonstrates the effectiveness of the proposed framework against commonly used deep learning architectures, comparison with recently published studies is necessary to position the proposed method within the current research landscape. Therefore, the proposed framework was further compared with representative state-of-the-art rice leaf disease classification approaches reported in the literature, as summarized in Table 10.

As presented in Table 10, the proposed framework achieves competitive performance compared with existing rice leaf disease classification methods, obtaining an accuracy of 99.55%. Most previous studies mainly rely on single CNN-based architectures, transformer-based models, or knowledge distillation strategies. Although these approaches have demonstrated promising classification performance, they generally extract either local texture features or global contextual representations independently.

In contrast, the proposed framework introduces a dual-backbone complementary representation learning strategy by integrating Swin Transformer and ConvNeXt. The Swin Transformer branch captures long-range contextual relationships and overall disease distribution patterns, while the ConvNeXt branch extracts fine-grained local characteristics, such as lesion boundaries, texture variations, and color changes. Furthermore, the attention-based feature fusion module adaptively combines these complementary representations instead of treating all features equally, resulting in more discriminative disease representations.

**Table 10.** Comparison of the proposed model with previously published rice leaf disease classification studies for dataset-1.

| Reference                   | Dataset                          | Method                                                                | Accuracy (%) |
|-----------------------------|----------------------------------|-----------------------------------------------------------------------|--------------|
| T. Gobinath et al. [6]      | Rice Leaf Dataset                | EfficientNet-B4                                                       | 96.25        |
| T. Gobinath et al. [6]      | Rice Leaf Dataset                | MobileNetV3                                                           | 96.25        |
| P. S. Mallipudi et al. [36] | Rice Leaf Dataset                | Custom CNN                                                            | 98.16        |
| Richhariya et al. [37]      | Rice Leaf Dataset                | CNN                                                                   | 95.00        |
| Richhariya et al. [37]      | Rice Leaf Dataset                | Hybrid Ensemble                                                       | 98.00        |
| Richhariya et al. [37]      | Rice Leaf Dataset                | K-Nearest                                                             | 97.00        |
| Richhariya et al. [37]      | Rice Leaf Dataset                | AlexNet                                                               | 99.00        |
| P.-H. Hoang et al. [38]     | Rice Leaf Dataset                | Hybrid Knowledge Distillation Framework                               | 99.07        |
| R. K. Naseer et al. [39]    | Rice Leaf Dataset                | Vision Transformer (ViT)                                              | 99.00        |
| Proposed Framework          | Kaggle Rice Leaf Disease Dataset | Dual-Backbone Complementary Representation Learning (Swin + ConvNeXt) | 99.55        |

Beyond classification accuracy, the proposed framework was further validated through cross-validation, independent cross-dataset evaluation, and Grad-CAM analysis to demonstrate robustness and interpretability. These additional evaluations highlight that the proposed framework provides not only high predictive performance but also improved reliability for practical agricultural applications.

#### 5.6. Computational Complexity Analysis

While classification accuracy is an important evaluation criterion, computational efficiency is also essential for practical agricultural deployment. Therefore, computational complexity analysis was conducted to evaluate the trade-off between predictive performance and computational requirements. The proposed framework was compared with baseline models in terms of the number of parameters, GFLOPs, inference latency, and frames per second (FPS), as presented in Table 11.

**Table 11.** Computational complexity comparison of the proposed framework and baseline models.

| Model              | Parameters (M) | GFLOPs | Latency (ms) | FPS    |
|--------------------|----------------|--------|--------------|--------|
| EfficientNetV2     | 22.16          | 2.94   | 16.14        | 61.96  |
| ResNet50           | 23.53          | 4.13   | 5.43         | 184.13 |
| ConvNeXt-Tiny      | 27.83          | 4.45   | 5.88         | 170.15 |
| Swin Transformer   | 27.53          | 4.37   | 11.05        | 90.48  |
| Proposed Framework | 57.45          | 8.83   | 17.65        | 56.66  |

Table 11 shows that the proposed framework requires more computational resources than individual backbone networks due to the integration of two complementary feature extraction branches. Specifically, the proposed framework contains 57.45 million parameters and requires 8.83 GFLOPs, resulting in an inference speed of 56.66 FPS. Although the computational complexity is higher than single-backbone models, this additional cost provides significant performance improvements, including 99.55% benchmark accuracy, 98.91% five-fold cross-validation accuracy, and 90.80% cross-dataset accuracy. Therefore, the increased computational requirement is justified by the improvement in accuracy, robustness, and generalization capability. Furthermore, the achieved inference speed of 56.66 FPS demonstrates that the proposed framework remains suitable for real-time web-based agricultural decision-support applications, where reliable diagnosis is more critical than minimal computational complexity. It is also worth analyzing the computational efficiency and applicability of the proposed architecture. The model has 57.45M trainable parameters, and the performance measured on the experimental setup with GPU hardware equals to 56.66 FPS. At the same time, such performance metrics do not reflect the speed of inference on mobile and embedded devices because they were measured on the experimental platform using a GPU. In the current research, the proposed architecture was applied within a web-based application when the server performs the task of inference and delivers prediction results via the web interface. Inference on mobile or embedded devices has not been experimentally proven yet.

#### 5.7. Explainability Analysis Using Grad-CAM

Although deep learning models achieve high classification accuracy, their black-box nature remains a major limitation for real-world agricultural applications. Farmers and agricultural experts require visual evidence to understand whether the model focuses on meaningful disease symptoms rather than irrelevant background information. Therefore, Grad-CAM analysis was employed to investigate the decision-making behavior of the proposed framework.

The Grad-CAM visualization results are presented in Figure 14. The highlighted regions indicate the image areas that contribute most significantly to the predicted disease category, where warmer colors represent higher importance and cooler regions indicate lower contribution.

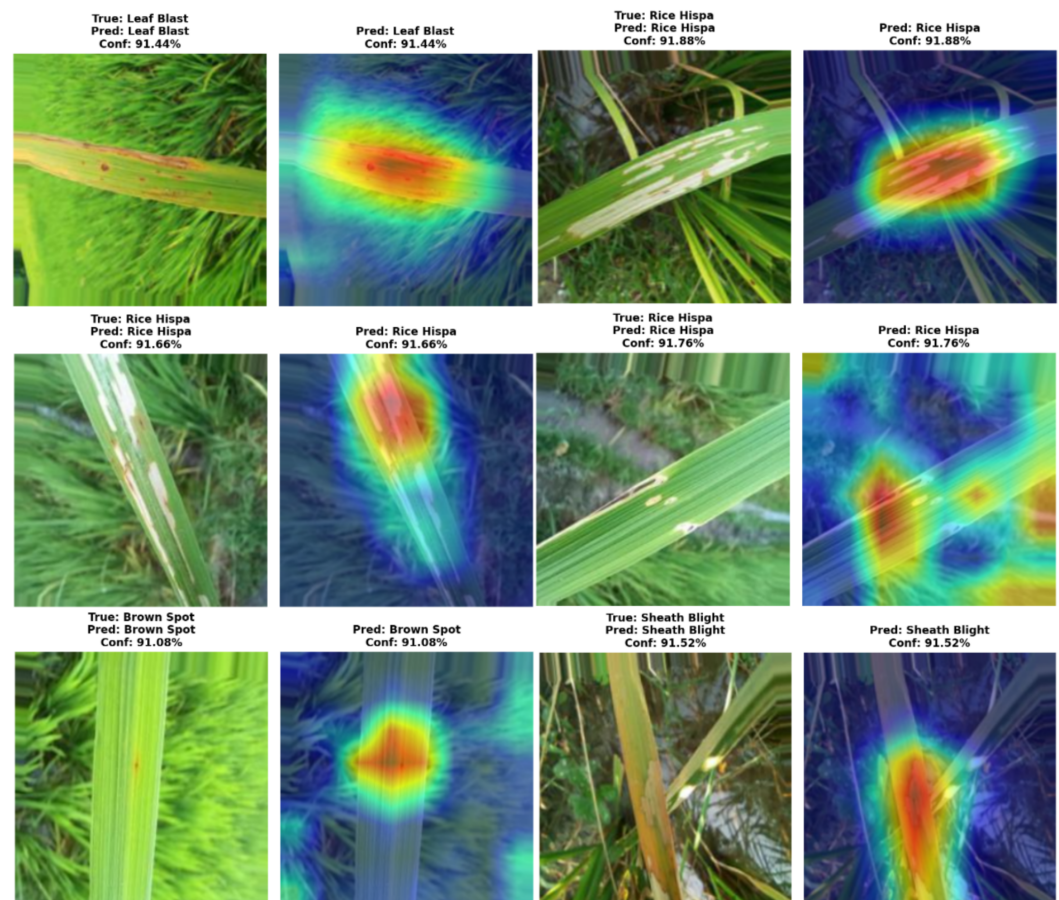
The visualization demonstrates that the proposed framework primarily focuses on disease-related regions, including lesions, damaged tissues, and abnormal color patterns, while assigning limited importance to background areas. This finding confirms that the learned representations are associated with meaningful biological characteristics rather than environmental artifacts.

The improved interpretability is attributed to the complementary feature learning strategy and attention-based fusion mechanism. By combining global contextual information from Swin Transformer and local disease characteristics from ConvNeXt, the model learns more informative and interpretable representations. Therefore, Grad-CAM analysis provides additional evidence supporting the reliability and practical usability of the proposed framework.

Additional Grad-CAM examples were examined to better understand which regions of the rice leaves influenced the model predictions. As shown in Figure 14, all the presented samples were correctly classified, with the predicted labels matching the ground-truth labels. The Grad-CAM maps show that the model mainly focuses on visibly affected regions of the leaves. For instance, the Leaf Blast and Rice Hispa samples show strong activation around the elongated and damaged portions of the leaves, while the Brown Spot and Sheath Blight samples show higher activation around localized symptomatic regions.



These highlighted areas are generally consistent with the visible disease symptoms in the corresponding images.



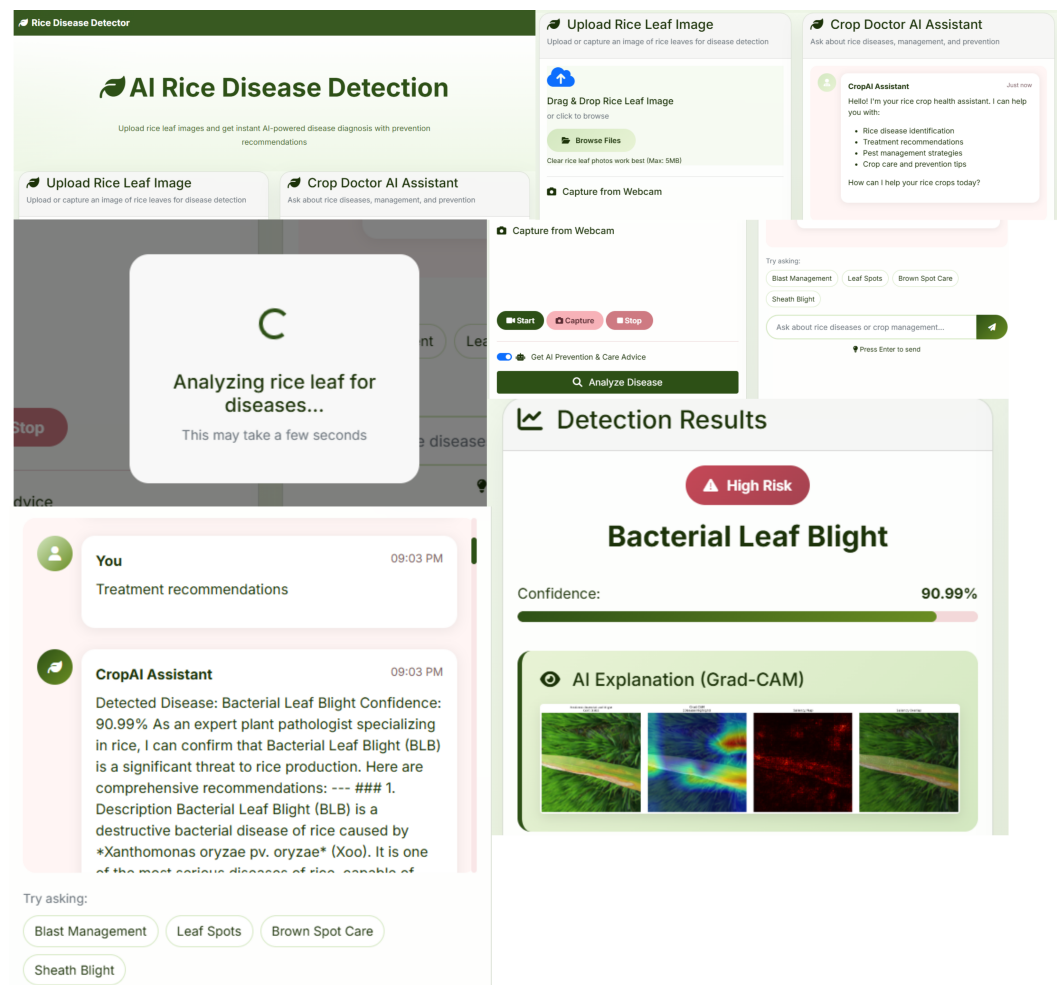
**Figure 14.** Grad-CAM visualization of disease-relevant areas detected by the proposed dual-backbone complementary representation learning framework for disease diagnosis in rice leaves.

Examples also indicate that symptoms of different diseases give rise to different reactions in the model, such as spots, elongated lesions, discolored lesions, and damaged sections of the leaves. Qualitative evidence that the model has learned the relevant visual features to classify a disease can be obtained from the agreement between the predicted and ground-truth labels, as well as the focused activation on the Grad-CAM maps produced by the model. However, the main drawback of Grad-CAM is that it is an interpretability tool and does not provide pixel-level disease annotations. Therefore, the highlighted regions should be considered an indication of the areas that influenced the prediction rather than exact boundaries of the actual disease symptoms. The number of representative cases in the present figure is confined to those that were correctly classified to demonstrate this trend.

### 5.8. Real-Time Deployment and LLM-Based Advisory System

High classification accuracy alone is insufficient for practical agricultural applications unless the model can support real-world decision-making. Therefore, the proposed framework was integrated into a web-based deployment system that combines automated disease diagnosis, visual explanation, and post-diagnosis agricultural advisory.

The workflow of the proposed system is illustrated in Figure 15. Users upload rice leaf images through the web interface, and the trained dual-backbone classification model performs disease prediction by extracting complementary global and local representations followed by attention-based feature fusion.



**Figure 15.** Web-based real-time deployment interface of the proposed rice leaf disease detection system showing image upload, prediction output with confidence score, Grad-CAM and the integrated LLM-based assistant for disease explanation and recommendations.

After disease identification, the predicted disease category is provided to the LLM-based advisory module. The LLM does not participate in image feature extraction, classification, or performance evaluation. Instead, it functions exclusively as a post-diagnosis decision-support component to generate farmer-oriented recommendations, including disease explanation, possible causes, management strategies, and preventive measures. This separation between the classification model and advisory module ensures that the quantitative performance of the proposed framework remains fully determined by the deep learning model, while the LLM improves practical usability by transforming prediction results into understandable agricultural guidance. Therefore, the proposed framework extends the proposed method from a classification model into an intelligent agricultural decision-support system, demonstrating its potential for real-world smart farming applications.

To clarify the implementation of the LLM-based advisory module, the system uses the Gemini-2.5-Flash language model with a structured system–human prompting strategy. The system prompt is used to identify the model as an agricultural plant-disease specialist, and it includes the predicted class of disease and the confidence score as input. The model is trained to produce information about what to do in seven distinct categories: description, symptoms, causes, treatment, prevention, organic solutions, and immediate actions. The human-message portion of the prompt is then used to provide the user’s input. This structured prompting strategy ensures a consistent organization of the generated recommendations for farmers. The LLM is used only after the disease classification stage to

generate advisory information and does not contribute to the image classification process. In the current study, the recommendations generated were not validated formally by agricultural experts, so the LLM module is outlined here as an advisory support system and does not constitute an agricultural recommendation system. Website Link: [https://huggingface.co/spaces/nahidur415/Plant\\_disease\\_detector](https://huggingface.co/spaces/nahidur415/Plant_disease_detector) (accessed on 5 July 2026).

## 6. Conclusions

This study presents a dual-backbone complementary representation learning framework for accurate, explainable, and practically deployable rice leaf disease diagnosis. Rice diseases significantly affect agricultural productivity and food security, highlighting the necessity of intelligent and reliable automated diagnosis systems. To address the limitations of existing single-backbone approaches, which often struggle to simultaneously capture global disease patterns and fine-grained local visual characteristics, the proposed framework integrates Swin Transformer and ConvNeXt in a complementary feature learning strategy. The proposed framework effectively combines global contextual representations from the Swin Transformer stream and local texture representations from the ConvNeXt stream. Furthermore, an attention-based feature fusion module is introduced to adaptively learn the importance of different feature components, enabling the model to generate more discriminative disease representations compared with conventional feature concatenation strategies. In addition, Grad-CAM visualization improves model interpretability by highlighting disease-relevant regions, while the LLM-based advisory module extends the framework toward practical agricultural decision support by generating disease-specific management recommendations after diagnosis. Extensive experimental evaluation on a rice leaf disease datasets demonstrates the effectiveness of the proposed approach. Despite the promising performance, the current framework is primarily designed for image-based rice leaf disease classification and relies on the availability of representative training samples. Future research will focus on expanding the framework toward larger and more diverse field datasets, developing lightweight versions for mobile and edge-based agricultural applications, and integrating additional environmental information, such as weather conditions and soil characteristics, for more comprehensive crop health monitoring. Furthermore, future studies will investigate advanced multimodal learning approaches that combine visual diagnosis with domain-specific agricultural knowledge to develop more intelligent and personalized decision-support systems for smart farming.

**Author Contributions:** Conceptualization, M.N.R., M.K.H. and A.A.M.; validation, M.N.R., A.A.M. and A.R.; formal analysis, M.N.R., A.A.M., A.S.M.M. and A.R.; investigation, M.N.R., A.A.M. and A.R.; resources, M.K.H., J.S., M.N.H. and A.S.M.M.; data curation, M.N.R., A.A.M. and A.R.; writing—original draft preparation, M.N.R., A.A.M., A.S.M.M., T.R.S. and A.R.; writing—review and editing, M.K.H., M.N.H., A.S.M.M., T.R.S. and A.A.M.; visualization, M.N.R., A.A.M. and M.K.H.; supervision, J.S., M.N.H. and A.S.M.M.; project administration, A.S.M.M., J.S. and M.N.R.; funding acquisition, J.S., A.S.M.M. and M.N.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by Universiti Kuala Lumpur, Malaysian Institute of Information Technology (UniKL MIIT) Research Funds.

**Data Availability Statement:** The data used in this study are publicly available at: Dataset-1: <https://www.kaggle.com/datasets/loki4514/rice-leaf-diseases-detection> (accessed on 18 May 2026), Dataset-2: <https://www.kaggle.com/datasets/trumanrase/rice-leaf-diseases> (accessed on 27 May 2026). The source code of the proposed framework is available at: [https://github.com/NahidurRahaman/Rice\\_Leaf.git](https://github.com/NahidurRahaman/Rice_Leaf.git) (accessed on 5 July 2026).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Yang, H.; Deng, X.; Shen, H.; Lei, Q.; Zhang, S.; Liu, N. Disease Detection and Identification of Rice Leaf Based on Improved Detection Transformer. *Agriculture* **2023**, *13*, 1361. [\[CrossRef\]](#)
2. Lee, S.H.; Jiang, Q.W.; Cheng, C.H.; Tsai, Y.S.; Huang, Y.F. Boosting Rice Disease Diagnosis: A Systematic Benchmark of Deep CNN Models in Precision Agriculture. *Agriculture* **2025**, *15*, 2494. [\[CrossRef\]](#)
3. Tasci, M. DBLA-MobileNetV2: Real-Time Rice Leaf Disease Detection on Edge Devices With Dual-Branch Lightweight Attention. *IEEE Access* **2026**, *14*, 22356–22371. [\[CrossRef\]](#)
4. Munny, I.J.; Majumder, A.; Piyas, B.R.C.; Farid, F.A.; Jani, M.R.; Tisha, F.J.; Jahan, I.; Miah, A.S.M.; Karim, H.A. An ROI-Guided Optimized Machine Learning Framework for Orange Disease Recognition with Feature Selection and Explainability. *Comput. Mater. Contin.* **2026**, *89*, 65. [\[CrossRef\]](#)
5. Li, R.; Chen, S.; Matsumoto, H.; Gouda, M.; Gafforov, Y.; Wang, M.; Liu, Y. Predicting rice diseases using advanced technologies at different scales: Present status and future perspectives. *ABIOTECH* **2023**, *4*, 359–371. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Ayyappan, A.B.; Gobinath, T.; Kumar, M.; Sivaramakrishnan, A. Rice Plant Disease Detection Using Convolutional Neural Networks. *Discov. Artif. Intell.* **2025**, *5*, 50. [\[CrossRef\]](#)
7. Murugavalli, S.; Gopi, R. Plant leaf disease detection using vision transformers for precision agriculture. *Sci. Rep.* **2025**, *15*, 22361. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Zhou, C.; Zhong, Y.; Zhou, S.; Song, J.; Xiang, W. Rice leaf disease identification by residual-distilled transformer. *Eng. Appl. Artif. Intell.* **2023**, *121*, 106020. [\[CrossRef\]](#)
9. Chaudhary, S.; Kumar, U.; Sahoo, B.M. A deep learning optimized model for classification and detection of rice leaf diseases. *Sci. Rep.* **2026**, *16*, 20539. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017.
11. Tzachor, A.; Devare, M.; Richards, C.; Pypers, P.; Ghosh, A.; Koo, J.; Johal, S.; King, B. Large Language Models and Agricultural Extension Services. *Nat. Food* **2023**, *4*, 941–948. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Zhang, Y.; Fan, Q.; Chen, X.; Li, M.; Zhao, Z.; Li, F.; Guo, L. IPM-AgriGPT: A Large Language Model for Pest and Disease Management. *Mathematics* **2025**, *13*, 566. [\[CrossRef\]](#)
13. Upadhyay, A.; Chandel, N.S.; Singh, K.P.; Chakraborty, S.K.; Nandede, B.M.; Kumar, M.; Subeesh, A.; Upendar, K.; Salem, A.; Elbeltagi, A. Deep learning and computer vision in plant disease detection: A comprehensive review of techniques, models, and trends in precision agriculture. *Artif. Intell. Rev.* **2025**, *58*, 92. [\[CrossRef\]](#)
14. Qezelbash-Chamak, J.; Hicklin, K. A Hybrid Learnable Fusion of ConvNeXt and Swin Transformer for Optimized Image Classification. *IoT* **2025**, *6*, 30. [\[CrossRef\]](#)
15. Nguyen, T.P.; Dang, T.P.; Tran-Thi, M.K. LCSNet–FedPerGC: An efficient federated learning framework for rice leaf disease classification in the Mekong Delta. *Comput. Electron. Agric.* **2026**, *254*, 112249. [\[CrossRef\]](#)
16. Premkumar, S.; Sharma, P. Development of a concurrent attention-based CNN model for enhanced crop disease analytics in rice. *AIP Conf. Proc.* **2026**, *3414*, 040008. [\[CrossRef\]](#)
17. Pai, P.; Amutha, S.; Basthikodi, M.; Ahamed Shafeeq, B.M.; Chaitra, K.M.; Gurpur, A.P. A twin CNN-based framework for optimized rice leaf disease classification with feature fusion. *J. Big Data* **2025**, *12*, 89. [\[CrossRef\]](#)
18. Roy, P.S.; Kukreja, V. Vision transformers for rice leaf disease detection and severity estimation: A precision agriculture approach. *J. Saudi Soc. Agric. Sci.* **2025**, *24*, 3. [\[CrossRef\]](#)
19. Dai, W.; Zhu, W.; Zhou, G.; Liu, G.; Xu, J.; Zhou, H.; Hu, Y.; Liu, Z.; Li, J.; Li, L. AISOA-SSformer: An effective image segmentation method for rice leaf disease based on the transformer architecture. *Plant Phenomics* **2024**, *6*, 0218. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Saood, S.; Khan, S.H.; Iqbal, R. A Customized Swin Transformer-Based Framework for Cassava Leaf Disease Classification. *Spectr. Eng. Sci.* **2026**, *4*, 37–48. [\[CrossRef\]](#)
21. Srinivasan, S.; Ranjith Kumar, A.; Nagashree, B.A.; Tanwar, J.; Singh, V.P.; Moorthy, U. Multi-class classification of plant leaf diseases using a hybrid deep neural transformer system and explainable AI techniques. *Sci. Rep.* **2026**, *16*, 18161. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Felicita, S.A.M.; Kavitha, B.R. Enhancing multiclass plant disease classification using GAN-boosted vision transformer with XAI insights. *Front. Plant Sci.* **2026**, *16*, 1649399. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Jawed, M.M.; Tufail, F.A.; Ahmed, M.Z.; R, A.S.; Nallusamy, P.; Raja, K.T. A hybrid deep learning framework using convolutional and transformer models for robust plant disease classification. *Sci. Rep.* **2026**, *16*, 9704. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Arshad, Z.; Javed, A.; Saudagar, A.K.J. ConvGeM-next: A deep learning framework for plant disease detection. *Front. Plant Sci.* **2026**, *17*, 1763739. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Al-Falluji, R.A.; Albahar, M.A. Enhancing green AI through explainable deep learning-based multi-model for automated rice leaf disease classification. *Discov. Comput.* **2025**, *28*, 269. [\[CrossRef\]](#)



26. Khiem, N.M.; The, H.T.; Quyen, P.N. REBANet: A Hybrid Deep Learning Model for Rice Leaf Disease Prediction with Grad-CAM Explainability. In *Computational Intelligence in Engineering Science*; Springer: Berlin/Heidelberg, Germany, 2026; pp. 535–549. [[CrossRef](#)]
27. Wang, Y.; Zhao, K.; Li, H.; Wang, R.; Chen, W. A condition-aware retrieval-augmented decision support framework for tomato cultivation management. *Front. Plant Sci.* **2026**, *17*, 1359–1374. [[CrossRef](#)] [[PubMed](#)]
28. Gautam, V.; Rani, J. Cross-Modal Apple Disease Recognition: Enhancing Apple Leaf Disease Detection Through Vision–Language Feature Alignment. *Appl. Fruit Sci.* **2026**, *68*, 245. [[CrossRef](#)]
29. Shuai, L.; Zhu, L.; Yang, Y.; Lin, T. AgriMAPO: A multimodal automatic prompt optimization approach for crop disease classification using large language models. *Artif. Intell. Agric.* **2026**, *16*, 911–925. [[CrossRef](#)]
30. Bozdağ, Ö.E.; Arabacı, H.E.; Keskin, M.B.; Kahraman, A. Publicly Available Datasets for Smart and Precision Agriculture: A Systematic Review. In *Proceedings of the 5th International Informatics and Software Engineering Conference (IISEC)*, Ankara, Turkey, 5–6 February 2026; pp. 694–699. [[CrossRef](#)]
31. Uyala, S.; Uyala, S.B.; James, B.M. Web-Machine Learning-API Integration for Agricultural Decision-Making and Optimization. In *Proceedings of the IEEE 9th International Conference on Information and Communication Technology (CICT)*, Chennai, India, 19–21 December 2025; pp. 1–6. [[CrossRef](#)]
32. Singh, N.; Malik, M.; Khokhar, M.K.; Kumar, A.; Choudhary, M.; Kumar, M.; Nayak, N.; Singh, A.K. Mobile-assisted deep learning framework for identification of insect pests and diseases of maize from field images. *Front. Plant Sci.* **2026**, *17*, 1803005. [[CrossRef](#)] [[PubMed](#)]
33. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022. [[CrossRef](#)]
34. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 11976–11986. [[CrossRef](#)]
35. Li, F.; Zou, C.; Yun, J.; Huang, L.; Liu, Y.; Tao, B.; Xie, Y. A Depth Awareness and Learnable Feature Fusion Network for Enhanced Geometric Perception in Semantic Correspondence. *Sensors* **2024**, *24*, 6680. [[CrossRef](#)] [[PubMed](#)]
36. Mallipudi, P.S.; Dara, S.; Banka, V.; Allam, M. Comparative Analysis of Deep Learning Architectures for Rice Leaf Disease Detection. In *Proceedings of the Second International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, Eride, India, 2–4 March 2026; pp. 1000–1006. [[CrossRef](#)]
37. Islam, M.N.; Richhariya, P. Deep Learning-Based Classification of Rice Leaf Diseases Using Hybrid Ensemble Models. *Int. J. Innov. Sci. Eng. Manag.* **2024**, *3*, 34–41. [[CrossRef](#)]
38. Hoang, P.H.; Trinh, N.T.; Tran, V.M.; Phan, T.T.H. Multi-objective Hybrid Knowledge Distillation for Efficient Deep Learning in Smart Agriculture. *arXiv* **2025**, arXiv:2512.22239. [[CrossRef](#)]
39. Naseer, R.K.; Afzal, A.; Arshaq, M.; M., R.K.M.; Abini, A.M. RiViT: Rice Leaf Disease Identification and Classification Using Vision Transformer (ViT). In *Proceedings of the 4th International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, Ernakulam, India, 11–13 June 2025; pp. 245–250. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.