

Fine classification of rice diseases under field conditions based on improved ConvNeXt network

Shuaipeng Fei

China Agricultural University

Shiteng Han

China Agricultural University

Chunli Lv

China Agricultural University

Yan Guo

China Agricultural University

Yuntao Ma

`Yuntao.ma@cau.edu.cn`

China Agricultural University

Research Article

Keywords: Rice disease, ConvNeXt, HAT, Deep learning

Posted Date: May 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9636383/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Fine classification of rice diseases under field conditions based on improved ConvNeXt network

Shuaipeng Fei⁺, Shiteng Han⁺, Chunli Lv, Yan Guo, Yuntao Ma*

College of Land Science and Technology, China Agricultural University, Beijing 100193, China

***Correspondence: Yuntao Ma, yuntao.ma@cau.edu.cn**

⁺Shuaipeng Fei and Shiteng Han contributed equally to this work

Abstract

Rice disease identification is a critical technique for ensuring yield and quality in precision agriculture. However, complex field backgrounds, subtle lesion features, and similar symptomatic manifestations have led to low accuracy and poor robustness in traditional classification methods. To address these issues, this study proposes an improved ConvNeXt network model (Improve-ConvNeXt) for the fine classification of rice diseases under field conditions. A high-quality dataset containing six categories (healthy rice, rice blast, brown spot, bacterial leaf blight, bacterial leaf streak, and bacterial grain rot) was constructed from field images and public datasets, with a total of 5,663 samples. Using ConvNeXt-Tiny as the backbone, the model integrates a Hybrid Attention Transformer (HAT) to enhance the perception of lesion regions and key channels, and introduces Spatial and Channel Reconstruction Convolution (SCConv) to reduce feature redundancy and strengthen effective information expression. Experiments show that the Improve-ConvNeXt model achieves 96.27% accuracy on the test set, which is 4.85% higher than that of the original ConvNeXt and significantly outperforms ResNet and DenseNet. The precision, recall, and F1-score reach 95.84%, 96.11%, and 95.95%, respectively. Confusion matrix and Grad-CAM visualization prove that the model can accurately focus on lesion areas and effectively distinguish similar diseases. This method provides high precision and strong generalization for rice disease identification in complex field environments, and offers a reliable technical reference for intelligent monitoring and precise management of rice fields.

Keywords: Rice disease; ConvNeXt; HAT; Deep learning

1 Introduction

Rice (*Oryza sativa* L.) serves as a cornerstone of global food security by providing the primary caloric intake for over half of the world's population. As a staple crop widely cultivated across Asia, Africa, and Latin America, its stable production directly affects social stability, economic development, and the livelihoods of hundreds of millions of farming households (Muthayya et al., 2014). In the face of contemporary challenges such as rapid population growth, accelerating urbanization, diminishing arable land, and intensifying climate-induced stressors

including extreme temperatures, droughts, and floods, enhancing rice yield stability, stress resilience, and quality has become a strategic necessity rather than merely an agricultural objective (Mohidem et al., 2022). The continuous evolution of smart agriculture and the global transition toward Agriculture 4.0 further necessitate the integration of high-throughput phenotyping, intelligent monitoring, and automated diagnostic systems to accelerate breeding efficiency, optimize field management, and support sustainable crop protection (Huang et al., 2022; Zhai et al., 2020). Accurate and timely identification of rice health status, especially early and precise diagnosis of diseases, has become an indispensable link in the entire industrial chain of rice production.

However, the frequent occurrence and rapid spread of phytopathogens remain a primary obstacle restricting rice yield and quality, often leading to substantial yield losses, increased pesticide usage, and severe degradation of grain quality and market value. Rice blast, bacterial leaf blight, brown spot, and bacterial grain rot are among the most destructive diseases that frequently break out in field conditions, causing severe damage during critical growth stages. Traditional disease identification relies heavily on manual field scouting, experiential judgment, and expert visual assessment, which is not only time-consuming and labor-intensive but also highly subjective and inconsistent. When applied to large-scale, intensive, and standardized farming operations, such methods are unable to meet the demands for real-time, high-throughput, and accurate diagnosis (Bock et al., 2010). The inefficiency and lag of manual intervention often result in the delayed application of chemical treatments, which not only fails to suppress disease outbreaks effectively but also causes excessive pesticide residues, environmental contamination, and ecological damage. Early attempts to automate this process utilized conventional computer vision algorithms relying on handcrafted features such as color histograms, texture descriptors, shape parameters, and edge features (GUAN et al., 2010; MA et al., 2018). Although these techniques marked an initial step toward automatic disease recognition, their performance was frequently compromised by the high morphological diversity of lesions, uneven illumination, leaf occlusion, complex soil backgrounds, and other interference factors in dynamic field environments, resulting in poor robustness and limited generalization ability.

The paradigm shift in agricultural intelligence has been driven by Deep Learning (DL), especially Convolutional Neural Networks (CNNs), which automatically learn hierarchical visual features and effectively reduce the dependence on manual feature extraction and professional experience (Gu et al., 2018; Li et al., 2021). Compared with traditional machine learning methods, CNN-based models can capture low-level edge and texture features as well as high-level semantic information from disease images, enabling more robust and adaptive representation learning.

Representative architectures such as ResNet, DenseNet, and Inception have been widely applied in plant disease diagnosis, achieving promising classification accuracy under controlled laboratory environments with uniform illumination and simple backgrounds (Hu et al., 2020; Ji et al., 2020; Jung et al., 2023; Li et al., 2024; Pang et al., 2023; Upadhyay and Gupta, 2024; Xu and Li, 2024). These studies have confirmed that deep learning models possess strong feature expression capabilities and lay a solid foundation for intelligent disease identification.

Despite remarkable progress in laboratory environments, the practical deployment of CNN models in real field scenarios still faces enormous challenges. Under open-field conditions, continuous changes in natural light, complex canopy structures, overlapping leaves, soil clutter, weed interference, and variable shooting distances often lead to low signal-to-noise ratios and blurred lesion features. Standard CNN models are easily disturbed by these complex environmental factors, making it difficult to accurately distinguish real disease lesions from abiotic stresses such as nutrient deficiency, mechanical damage, water loss, or sunburn, resulting in high false detection rates and poor generalization (Yuan et al., 2026). Furthermore, numerous rice diseases exhibit highly similar visual characteristics during the early onset stage, such as tiny color changes, faint spots, or similar lesion shapes. This high inter-class similarity creates a significant bottleneck for fine-grained classification, which conventional CNNs can hardly overcome due to insufficient attention to local discriminative details.

To break through these limitations, researchers have gradually shifted their focus to fine-grained visual recognition and attention-augmented network structures. The emergence of ConvNeXt has renewed the potential of pure convolutional networks by introducing the design philosophy of Transformers while retaining the efficiency and stability of CNNs. With larger receptive fields, enhanced global information modeling, and adaptive weight adjustment mechanisms, ConvNeXt achieves stronger feature extraction ability than traditional convolutional models (Liu et al., 2022). In recent years, the integration of advanced attention modules has further improved the precision of disease recognition. Among them, Hybrid Attention Transformer (HAT) enhances the perception of critical lesion regions through spatial and channel collaborative attention, and Spatial-Channel Reconstruction Convolution (SCConv) reduces feature redundancy and strengthens effective information transmission. The combination of these modules enables the model to automatically focus on discriminative disease areas, suppress background noise, and improve classification accuracy and robustness in complex field environments (Cai et al., 2023).

In summary, although deep learning has laid the groundwork for rice disease classification, a robust model capable of balancing high precision with environmental resilience is still lacking. This research addresses this gap by proposing an enhanced Improve_ConvNeXt model

specifically engineered for fine-grained classification under complex field conditions. By introducing the HAT_SCConv module, we synergize the long-range modeling capabilities of Transformers with the efficiency of spatial-channel reconstruction. This study not only improves diagnostic precision but also investigates the model decision-making process through visualization techniques like Grad-CAM. Ultimately, this work provides a comprehensive technological framework for intelligent phenotypic monitoring and precision breeding, contributing to the broader goal of sustainable and resilient rice production systems.

2 Materials and Methods

2.1 Data Collection

The study was conducted at the experimental site of the Liaoning Saline-Alkali Land Utilization Research Institute, China (41°04' N, 122°18' E). The research area covers approximately 5,880 square meters, with dimensions of 150 m in length and 39.2 m in width. The field was organized into 287 plots; each plot occupied about 10.7 square meters (1.65 m x 6.5 m) and contained six rows of rice. A planting density of 0.3 m x 0.132 m was maintained. Using a randomized block design, the experiment aimed to identify elite varieties suitable for the region's saline-alkali soil. All varieties were transplanted around May 20, 2024, and harvested by October 10, 2024. Standardized field management was strictly applied throughout the growth cycle, with systematic documentation of growth data and disease patterns.

Image acquisition took place during two specific growth stages: the jointing stage and the early heading stage. To ensure image quality, photography was performed between 9:00 AM and 11:00 AM on clear days to minimize overexposure or shadows caused by direct sunlight. Images were captured using an iPhone 13 Pro Max smartphone, with the shooting height maintained between 30 and 50 centimeters from the leaves. The primary shooting angles were vertical (top-down) and 45-degree tilts, with a single-image resolution of 4032×3024 pixels. During the process, the focus was locked on the lesion areas to ensure they occupied the main portion of the frame, enabling the model to better capture disease features. To further enhance the diversity and representativeness of the dataset, this study integrated several public rice disease image datasets. Ultimately, a comprehensive dataset was constructed covering six categories: Healthy Rice, Rice Blast, Brown Spot, Bacterial Leaf Blight, Bacterial Leaf Streak, and Bacterial Grain Rot. All images have a resolution of at least 1024×768 pixels, and the shooting environments encompass various weather conditions and angles. This high degree of data diversity and representativeness effectively improves the generalization ability of the model. The rice disease dataset used in this study contains a total of 5,663 images. To ensure fairness in model evaluation and experimental reproducibility, the original dataset was partitioned into training, validation, and testing sets using a random sampling method at a ratio of 8:1:1. Specifically, all images were first shuffled and then

allocated to the three subsets according to the aforementioned ratio to ensure a balanced and representative distribution of disease categories in each subset. The number of samples for each category in the divided subsets is shown in Table 1.

Table 1 Distribution of the Rice Disease Classification Dataset

Category	Training / Validation / Test	Total Images
Healthy Rice	1412/176/176	1764
Rice Blast	1391/174/173	1738
Brown Spot	772/97/96	965
Bacterial Leaf Blight	384/48/47	479
Bacterial Leaf Streak	304/38/38	380
Bacterial Grain Rot	270/34/33	337
Total	4533/567/563	5663

2.2 Construction of the Improved Rice Disease Classification Model

To achieve high-precision classification of rice diseases, this section optimizes the network structure based on existing advanced deep learning architectures and the feature representation of disease images in complex field environments, constructing an improved rice disease classification model, Improve_ConvNeXt. This section will elaborate on the structural composition of the model, focusing on the backbone network ConvNeXt and the key improvement modules integrated into it.

2.2.1 ConvNeXt Backbone Network

ConvNeXt, as a convolutional neural network with outstanding performance in the field of computer vision in recent years, significantly raises the performance ceiling of traditional convolutional neural networks by drawing on the design concepts and successful experiences of Vision Transformers while retaining the efficiency of convolutional operations (Liu et al., 2022). Based on its powerful feature extraction capabilities, ConvNeXt was selected as the backbone network for the disease classification model, and its network structure is shown in Fig. 1. First, ConvNeXt adopts a hierarchical design similar to the Swin Transformer, dividing the network into four stages and configuring the number of ConvNeXt Blocks in each stage according to a ratio of 1:1:3:1. Second, it implements a Patchify structure, using a convolutional layer with a stride of 4 and a kernel size of 4×4 to directly process the input image, replacing the initial downsampling module of large convolutions and max pooling common in classic networks such as ResNet, capturing local information while controlling the amount of computation. Finally, the downsampling operations between stages are completed by a convolutional layer with a stride of 2 and a kernel size of 2×2 , which can better preserve spatial information compared to traditional pooling layers.

The core computational unit of ConvNeXt is the ConvNeXt Block. This module introduces an inverted bottleneck design: the input features first pass through a 1×1 convolutional layer to expand the channel dimension by 4 times, then a 7×7 depthwise convolution is used instead of the traditional 3×3 standard convolution for spatial feature extraction, and finally, another 1×1 convolutional layer restores the number of channels to the original dimension. This design provides a larger receptive field while controlling computational costs, helping the model capture broader contextual information. Inside the ConvNeXt Block, Layer Normalization is used instead of Batch Normalization, reducing the sensitivity of model performance to batch size. Meanwhile, the GELU activation function is used to replace the commonly used ReLU activation function to provide a smoother nonlinear mapping for gradient transmission and promote model convergence. To improve training stability and model generalization ability, the ConvNeXt Block further

integrates Layer Scale and Drop Path mechanisms. Among them, the Layer Scale mechanism introduces learnable scaling factors for the residual connection branches to adjust their contribution to the main path information, which helps stabilize the training process; the Drop Path mechanism randomly sets the output of certain residual connection paths to zero during training, enhancing the generalization ability of the model and alleviating overfitting.

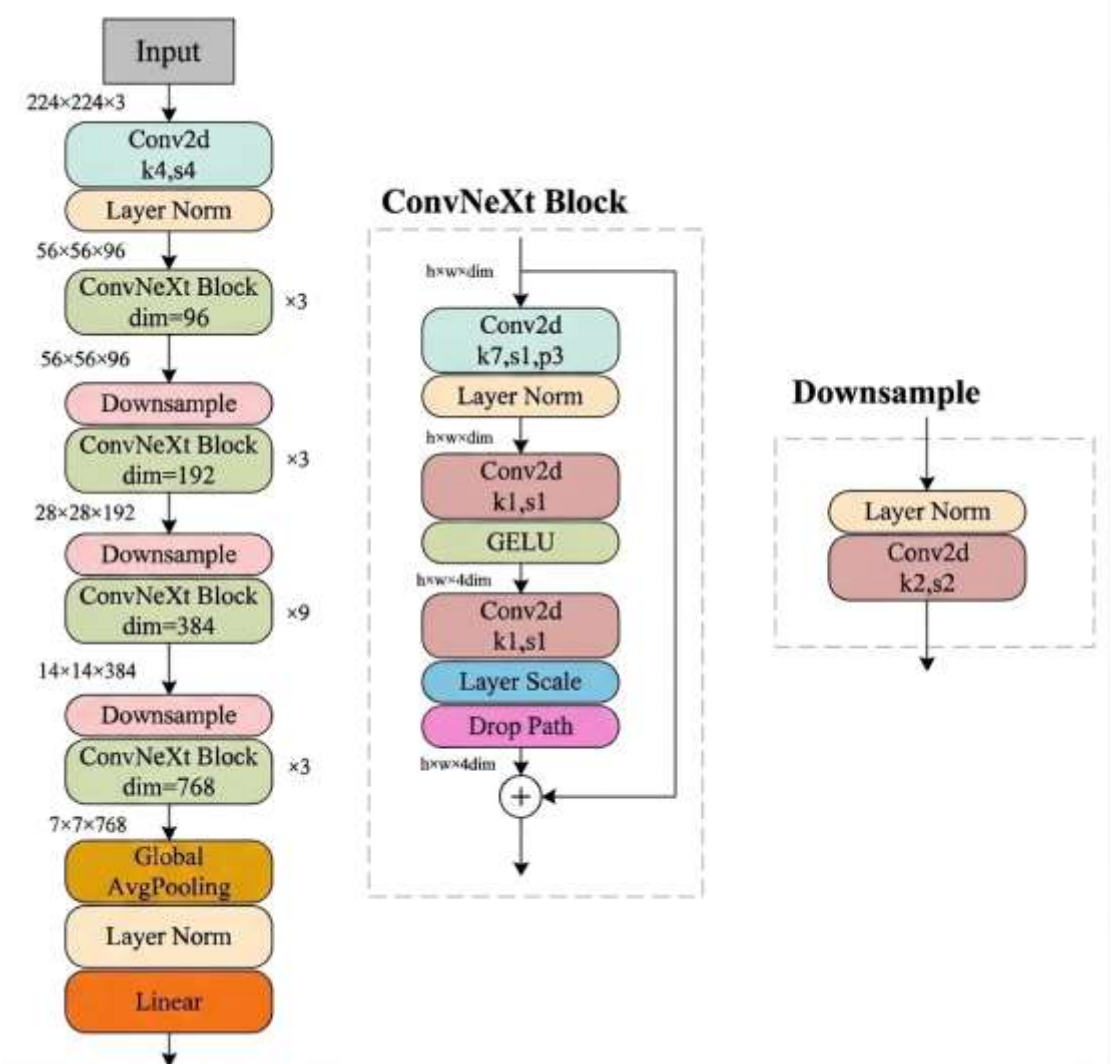


Fig.1 ConvNeXt network architecture

2.2.2 HAT Attention Mechanism

Although the ConvNeXt backbone network already possesses powerful global feature representation capabilities, there is still room for improvement when processing rice disease images with subtle lesions, high similarity, and complex backgrounds. Therefore, drawing on the design philosophy of HAT (Hybrid Attention Transformer), this model combines Window-based Multi-Head Self-Attention (W-MSA) with Channel Attention (CA) to further enhance the model's ability to focus on key discriminative regions and feature channels (Chen et al., 2023).

The overall framework of HAT consists of three stages: shallow feature extraction, deep

feature extraction, and image reconstruction, as shown in Fig. 2. First, the low-quality input image is processed through convolution to generate the initial feature map F_0 , which is then sent to a deep feature extraction module composed of multiple Residual Hybrid Attention Groups (RHAG). Through global residual connections, it is fused with the initial features to obtain the deep feature map F_{DF} . Finally, in the image reconstruction module, high-quality images are output through upsampling and convolution operations.

The core computational unit of HAT is the Residual Hybrid Attention Group, which typically consists of several Hybrid Attention Blocks (HAB) and an Overlapping Cross-Attention Block (OCAB) connected in series, with a 3×3 convolutional layer at the end for local feature fusion. The design objective of the Hybrid Attention Block is to refine information from both spatial and channel dimensions simultaneously. After the input features pass through Layer Normalization, two parallel paths are processed: one path employs Window-based Multi-head Self Attention (W-MSA) to calculate the spatial correlation between pixels within partitioned local windows, helping the model capture the local morphological characteristics of lesions; the other path utilizes a Channel Attention Block (CAB) to learn the importance of different feature channels and uses a Sigmoid activation function for adaptive weighting, enabling the model to prioritize feature channels that contribute more to distinguishing specific disease categories. The outputs of these two paths are fused through residual connections, achieving a synergistic effect between spatial and channel attention.

Standard window self-attention mechanisms may lead to information fragmentation and overlap because the windows are independent of each other. The Overlapping Cross-Attention Block introduces a cross-window information interaction mechanism-for example, by using query windows and key/value windows of different sizes-effectively promoting feature exchange between adjacent or overlapping regions. This allows the model to capture a wider range of contextual dependencies, which is potentially beneficial for identifying large-scale and irregularly shaped lesion features. The computational process is as follows: after the input features pass through Layer Normalization, the feature representation capability is enhanced through residual connections with the help of Overlapping Cross-Attention (OCA) and a Multilayer Perceptron (MLP) in sequence.

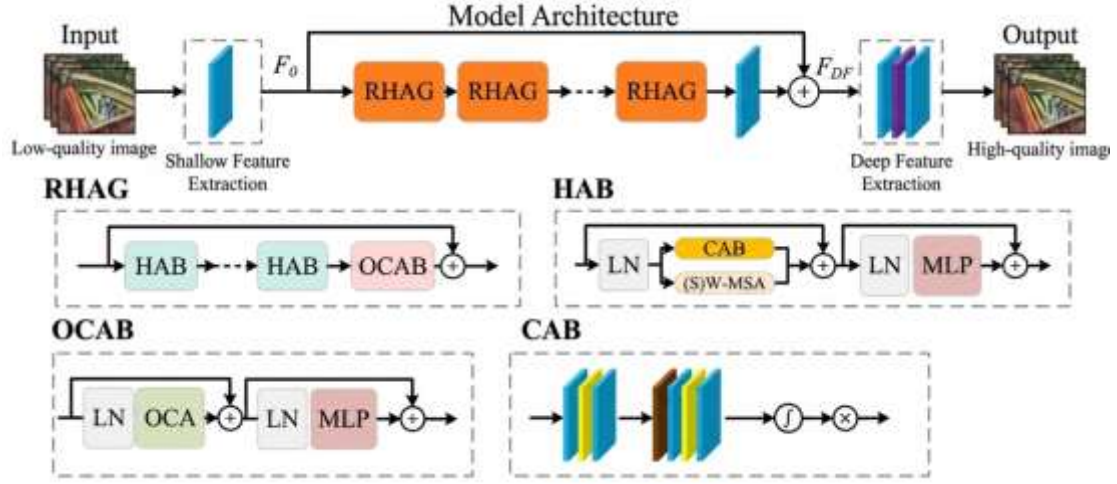


Fig. 2 HAT network architecture

2.2.3 Spatial and Channel Reconstruction Convolution

In the process of feature extraction in deep convolutional networks, a significant amount of redundant information often exists in both spatial and channel dimensions. To improve the computational efficiency and feature expression capability of the model, Spatial and Channel Reconstruction Convolution (SCConv) is further introduced.

From the perspective of the overall structure of SCConv, this module is typically embedded within the convolutional blocks of the network to replace standard convolutional layers. Its core consists of a Spatial Reconstruction Unit (SRU) and a Channel Reconstruction Unit (CRU) (Li et al., 2023), as shown in Fig. 3. The input feature X first undergoes a 1×1 convolution to complete channel compression and then enters the Spatial Reconstruction Unit. The SRU adopts a Separate-and-Reconstruct strategy: it first calculates the scaling factor for each channel through Group Normalization, then uses a Sigmoid activation and sets a threshold to divide the feature channels into information-rich channels W_1 and redundant channels W_2 . Finally, a Cross-Reconstruction is performed on both, reinforcing the information-rich channels while suppressing the redundant ones to output the spatially optimized feature X_W .

The spatially optimized feature X_W then enters the Channel Reconstruction Unit. The CRU adopts a Split-Transform-Fuse strategy, utilizing Grouped Convolution (GWC) and Pointwise Convolution (PWC) to reconstruct information in the channel dimension. It also introduces adaptive channel attention to weight different channels based on their importance, outputting the channel-optimized feature Y . Finally, a 1×1 convolution is used to restore the channel-optimized feature Y to the original number of channels, which is then passed to the next convolutional block after residual fusion with the main branch.

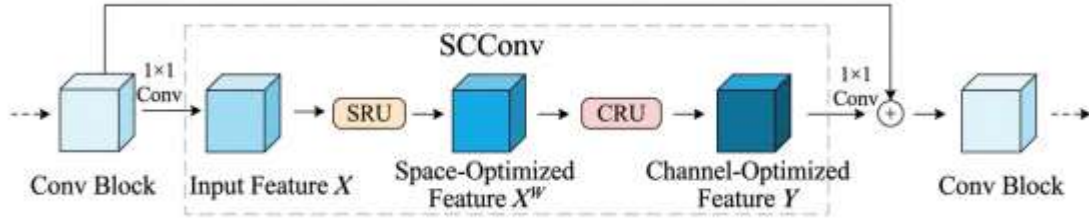


Fig.3 SCConvnetwork architecture

2.2.4 Improved Rice Disease Classification Model

The improved rice disease classification model, Improve_ConvNeXt, constructed in this section, uses ConvNeXt-Tiny as the backbone network for feature extraction and introduces the HAT_SCConv module in its deep stages to optimize classification performance. The specific structure is shown in Fig. 4. An input image with a size of $224 \times 224 \times 3$ first passes through the initial convolutional layer, Conv2dNormActivation, for downsampling and Layer Normalization. Subsequently, it sequentially enters the four core stages of the ConvNeXt backbone, which contain 3, 3, 9, and 3 ConvNeXt Blocks, respectively, with a Downsample module performing downsampling operations between every two stages.

The key improvement design is located after the output of the fourth stage, CNBlock4. At this point, the extracted high-level semantic feature map with a size of $7 \times 7 \times 768$ is fed into the HAT_SCConv module for deep feature refinement. This module first extracts global channel contextual information through a top-level AdaptiveAvgPool layer to provide a global perspective for subsequent feature reconstruction and attention calculation. Then, the features enter the core processing unit of the module. Unlike the traditional HAT structure, the SCConv module is used to replace the original hybrid attention module: specifically, the SRU replaces window-based self-attention to optimize local spatial features, and the CRU replaces the channel attention module to efficiently reconstruct channel information. This structure still retains the Overlapping Cross-Attention (OCA) module to strengthen cross-window information interaction. Additionally, a ReLU activation layer is introduced between the two residual-like hybrid attention group structures to enhance non-linear expression capabilities. The final output of the entire HAT_SCConv module is processed through a Sigmoid activation function.

After refinement by the HAT_SCConv module, the feature dimensions remain $7 \times 7 \times 768$. At this stage, the features enter a global average pooling layer, AdaptiveAvgPool2d, which compresses the spatial feature map of each channel into a single value, resulting in a $1 \times 1 \times 768$ feature vector. Finally, this vector passes through the Classifier Block—consisting of LayerNorm, Flatten, and Linear layers—to output a final vector of size $1 \times 1 \times 6$, completing the multi-classification task for rice diseases.

While ConvNeXt-Tiny provides powerful hierarchical feature extraction capabilities, the

HAT_SCConv module refines these features by integrating spatial-channel reconstruction with multi-level attention mechanisms. The improved model, Improve_ConvNeXt, which combines the strengths of both, aims to perform rice disease classification tasks more accurately, demonstrating significant advantages particularly in handling lesion details, distinguishing similar symptoms, and managing complex background interference.

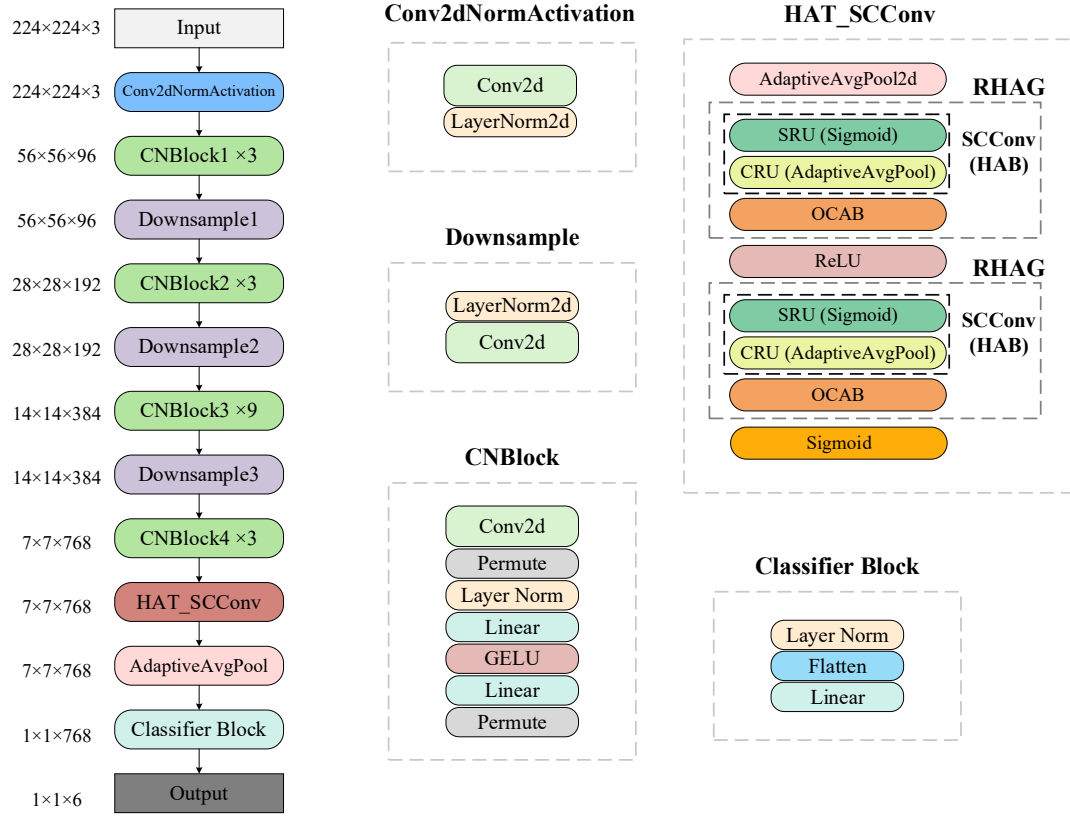


Fig. 3 Structural diagram of the improved rice disease classification model

2.3 Execution Environment

Table 2 details the primary hardware and software configuration parameters used for the rice disease classification research. The computing platform is based on a high-performance personal computer, utilizing the CUDA parallel computing architecture to fully leverage the acceleration capabilities of the GPU, ensuring efficient processing of large-scale image data and deep learning model operations. In terms of software, Python was used as the primary programming language, with code writing, debugging, and management conducted within the PyCharm integrated development environment (IDE). The construction of the deep learning models relies on the PyTorch framework, integrated with mainstream Python libraries such as Torchvision, OpenCV, and Matplotlib, providing a solid technical foundation for the successful implementation of the rice disease classification study.

Table 2 Execution Environment Configurations

Configuration Item	Configuration Parameter
Operating System	Microsoft Windows 11(64 Unit)
CPU	13th Gen Intel(R) Core(TM) i9-13900HX 2.20 GHz
GPU	NVIDIA GeForce RTX 4060 Laptop GPU
Available VRAM	8188MiB
Acceleration Environment	CUDA 12.6
Programming Language	Python 3.9
IDE	PyCharm Community Edition 2020.1.3
Deep Learning Framework	Pytorch 1.13.1

298

299 **2.4 Model Evaluation Metrics**

300 In the rice disease classification task, to accurately evaluate and compare the classification
301 performance of the models, the following four evaluation metrics were selected: Accuracy,
302 Precision, Recall, and the F1-score. Accuracy represents the proportion of samples correctly
303 classified by the model out of the total number of samples, directly reflecting the overall
304 classification performance of the model; the specific calculation steps are referred to in formula
305 (1). However, on datasets with imbalanced class distributions, it is necessary to perform a
306 comprehensive evaluation in conjunction with other metrics. Precision primarily focuses on the
307 proportion of samples truly belonging to a specific category among all samples predicted by the
308 model as that category; the specific calculation steps are referred to in formula (2). In disease
309 classification tasks, high precision implies higher credibility in the category prediction results,
310 which helps reduce misjudgments. Recall represents the proportion of samples successfully
311 classified by the model out of all samples truly belonging to that category; the specific calculation
312 steps are referred to in formula (3). In disease classification tasks, a high recall rate can reduce
313 missed detections. Nevertheless, precision and recall are often mutually restrictive. The F1-score
314 is the weighted harmonic mean of precision and recall, with values closer to 1 indicating better
315 model performance; the specific calculation steps are referred to in formula (4). The F1-score is
316 particularly effective for evaluating model performance on datasets with class imbalances.

317
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

318
$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

319
$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$F1\text{-score}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

Where TP (True Positive) represents the number of samples correctly predicted as the positive class, TN (True Negative) represents the number of samples correctly predicted as the negative class, FP (False Positive) represents the number of samples incorrectly predicted as the positive class, and FN (False Negative) represents the number of samples incorrectly predicted as the negative class.

3.1 Results

3.1.1 Training Parameters and Optimization Strategy

Rational hyperparameter configuration is critical in the training process of deep learning models. It not only helps accelerate the model convergence process but also prevents overfitting, ultimately leading to good generalization performance. During the training of the rice disease classification model, a set of hyperparameter combinations and optimization strategies were determined through multiple rounds of grid search; the specific configurations are detailed in Table 3. With an input size of 224×224 pixels, the batch size was set to 16 to maintain sufficient stability in gradient estimation, and 80 training epochs allowed the model to achieve fine-grained convergence at the end of the cosine annealing learning rate curve. Compared to the traditional Adam optimizer, the AdamW optimizer employs decoupled weight decay, resulting in faster convergence and superior generalization in convolutional networks. Furthermore, a weight decay setting of 0.01 was chosen as the option yielding the minimum validation error after a grid search within the 0.001 to 0.05 range.

Table 3 Hyperparameter Configuration

Hyperparameter	Value
Batch Size	16
Epoch	80
Input Size	224×224 Pixels
Optimizer	AdamW
Initial Learning Rate	0.001
Learning Rate Adjustment Strategy	Cosine Annealing
Weight Decay	0.01

3.1.2 Comprehensive Evaluation of Model Classification Performance

To comprehensively evaluate the effectiveness of the improved model, Improve_ConvNeXt, in the rice disease classification task, this section selects three classic convolutional neural networks—ResNet, DenseNet, and ConvNeXt—as control models. All models were trained using

the hyperparameter configurations described in Section 2.3 and evaluated on the training set, validation set, and an independent test set, respectively.

Fig. 5 illustrates the change in accuracy of the four models on the training and validation sets as training epochs progress, where subplot A represents the training set and subplot B represents the validation set. The improved model, Improve_ConvNeXt, demonstrated faster convergence in the early stages of training, with accuracy rising rapidly within the first 10 epochs and stabilizing after approximately 30 epochs. Furthermore, Improve_ConvNeXt exhibited particularly stable performance on the validation set, indicating strong generalization capabilities. ConvNeXt ranked second; although its curve was relatively smooth, its accuracy convergence speed was slightly slower, requiring nearly 40 epochs to stabilize, and its final accuracy was lower than that of Improve_ConvNeXt. In contrast, DenseNet and ResNet showed slower initial learning speeds and ultimately reached lower levels of convergence.

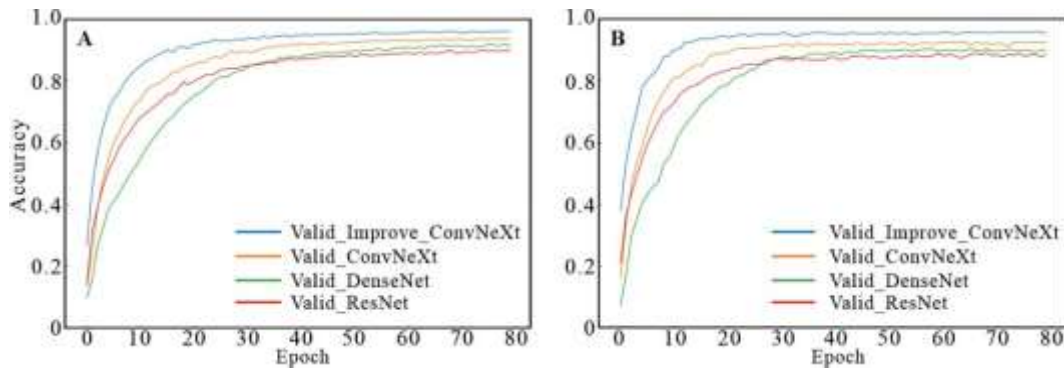


Fig.5 Accuracy curves of different models on training and validation sets

Fig. 6 illustrates the changes in the loss function for the four models on the training and validation sets over the training epochs, where subplot (a) represents the training set and subplot (b) represents the validation set. The loss function of the improved model, Improve_ConvNeXt, decreased the most rapidly in the early stages of training, reached its lowest point after approximately 30 epochs, and subsequently stabilized, ultimately converging to the lowest loss value on both the training and validation sets. ConvNeXt ranked second; its converged loss value was slightly higher than that of the improved model Improve_ConvNeXt, but it was able to stabilize after reaching its minimum. In contrast, DenseNet and ResNet were deficient in both loss reduction efficiency and the final levels achieved..

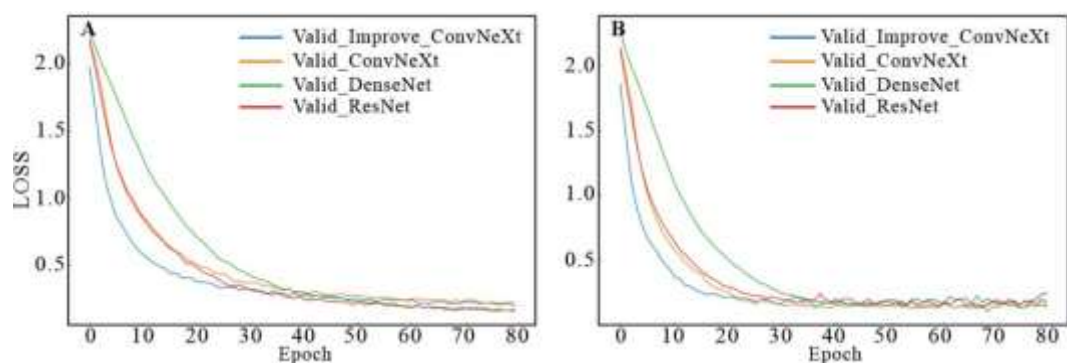


Fig.6 Loss curves of different models on training and validation sets

Figs. 7, 8, and 9 respectively illustrate the changes in precision, recall, and F1-score for the four models on the training and validation sets over the training epochs, where subplot (a) represents the training set and subplot (b) represents the validation set. Based on the comprehensive analysis of the aforementioned metrics, the improved model Improve_ConvNeXt demonstrated the most outstanding performance in the rice disease classification task. Whether on the training set or the validation set, its precision, recall, and F1-score reached the highest levels. Furthermore, its curves showed the fastest convergence speed, rising rapidly in the initial stages and then remaining stable at a high level. Although ConvNeXt performed well and remained stable during the training and validation processes, the performance levels it ultimately achieved were lower than those of the improved model Improve_ConvNeXt. The overall performance of DenseNet and ResNet was relatively weak; specifically, the precision curve of DenseNet exhibited significant fluctuations, while ResNet showed the slowest growth across all metrics and reached the lowest final performance levels, particularly performing poorly in the improvement of recall.

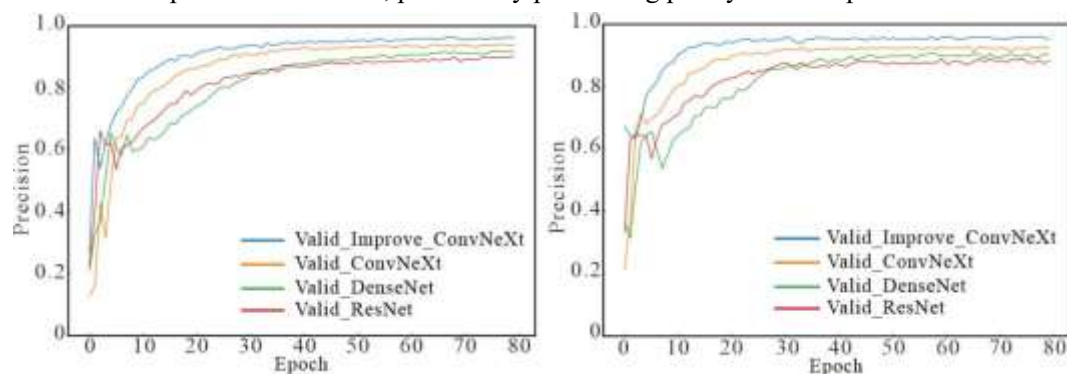


Fig.7 Precision curves of different models on training and validation sets

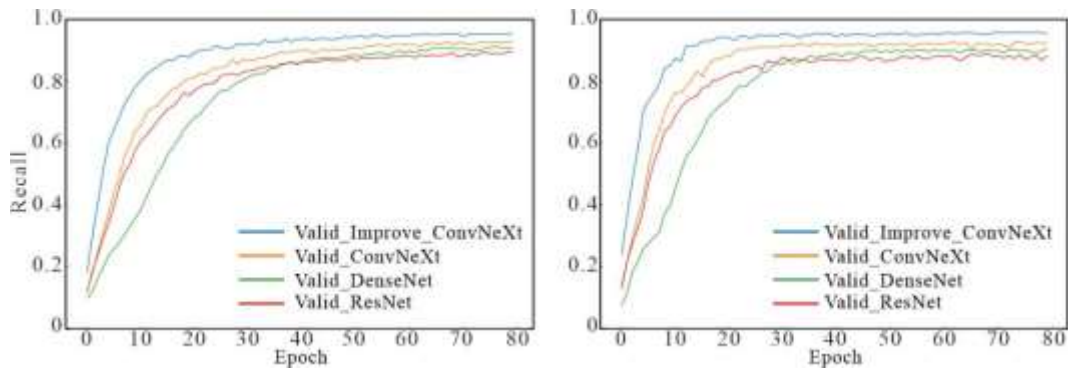


Fig.8 Recall curves of different models on training and validation sets

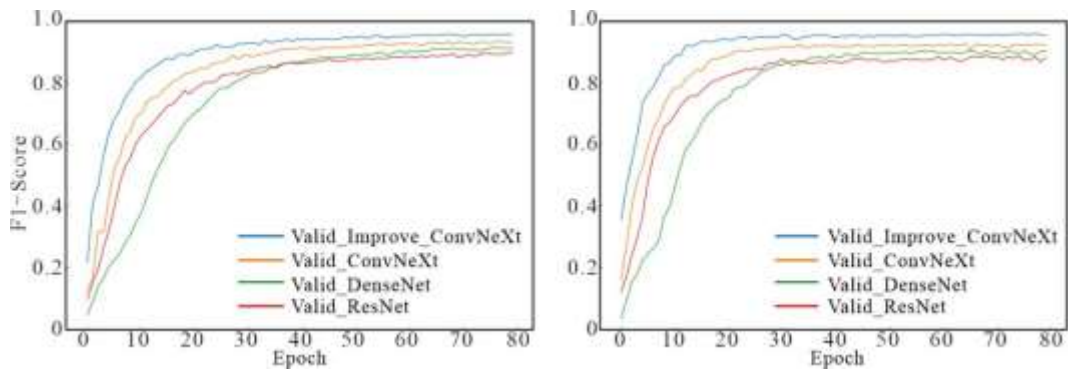


Fig.9 F1-score curves of different models on training and validation sets

To finalize the evaluation of each model's generalization ability, the rice disease classification research compared the performance of four models on an independent test set, with the detailed results shown in Table 4. The results indicate that the improved model, Improve_ConvNeXt, achieved the best performance on the test set, with a classification accuracy as high as 96.27%. This is significantly superior to the other models, representing an improvement of 4.85% over the standard ConvNeXt and an increase of over 8% compared to ResNet. These results provide strong evidence that Improve_ConvNeXt possesses stronger disease classification capabilities and higher reliability. Furthermore, Improve_ConvNeXt led across the board in precision, recall, and F1-score, indicating that the model effectively balances and reduces both misjudgment and missed detection rates while achieving high-precision classification. ConvNeXt followed in second place, with all metrics remaining stable above 90%, reflecting its efficient feature extraction capabilities as an advanced convolutional neural network. In contrast, DenseNet and ResNet exhibited relatively weaker classification performance. In conclusion, the evaluation results on the test set further validate the effectiveness of the improved model Improve_ConvNeXt. In the rice disease classification task, thanks to the introduction of the HAT attention mechanism and SCConv convolution module, the model's ability to perceive subtle and similar lesion features has been enhanced, thereby demonstrating robust performance and superior generalization ability. The improvement in all metrics compared to classic CNN baseline models fully proves that this

model can well satisfy the requirements for accurate rice disease classification in practical applications.

Table 4 Comparison of classification performance of different models on the test set

Model	Accuracy	Precision	Recall	F1-score
Model_ResNet	87.90%	87.54%	87.74%	87.59%
Model_DenseNet	89.17%	89.31%	89.12%	88.99%
Model_ConvNeXt	91.42%	91.43%	91.19%	91.30%
Improve_ConvNeXt	96.27%	95.84%	96.11%	95.95%

3.1.3 Confusion Matrix Analysis

To evaluate the specific classification performance of the improved Improve_ConvNeXt model across different categories in more detail, this section introduces the confusion matrix. This matrix clearly illustrates the correspondence between the model's classification results on the test set and the actual labels. In the confusion matrix, each row represents the true category of various rice diseases, and the row sum represents the actual number of image samples for that category in the test set. Each column represents the category predicted by the improved model, and the column sum represents the total number of image samples predicted as that category. Numerical values along the diagonal represent the number of correctly classified samples; for example, the value at position [1, 1] indicates the number of samples correctly classified as rice blast. Non-diagonal elements represent specific instances of misclassification. Generally, higher values on the diagonal and lower values off the diagonal indicate superior model performance. This analysis allows for the identification of easily confused categories and the primary circumstances of misclassification. As shown in Fig. 10, the improved rice disease classification model Improve_ConvNeXt achieved high accuracy on the test set, consistent with the evaluation results in Table 4.

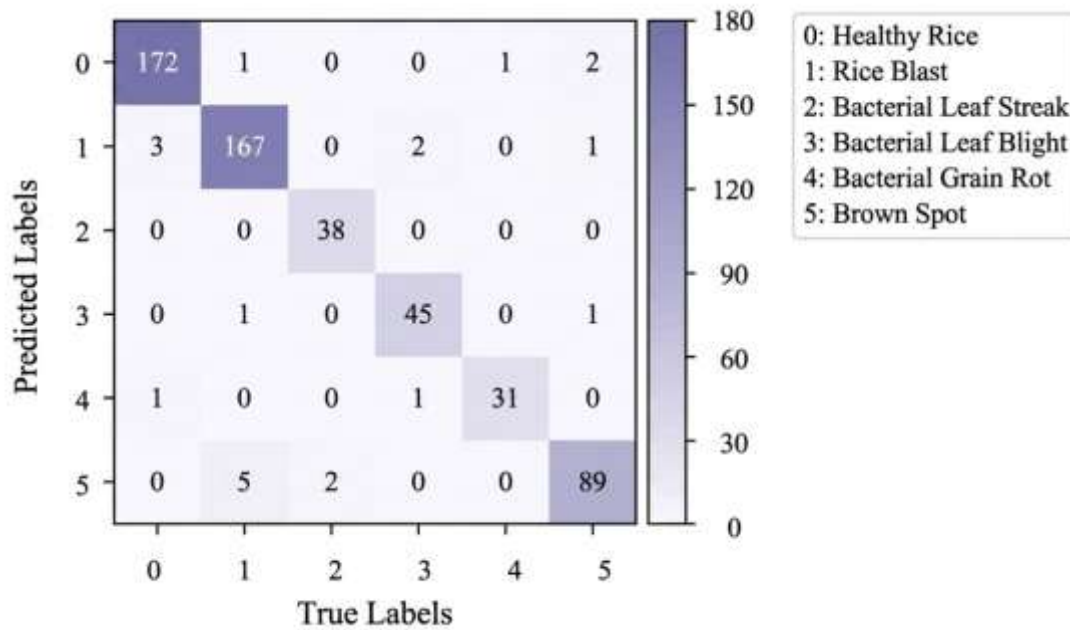


Fig. 10 Confusion Matrix of Improve_ConvNeXt

Specifically analyzing each category, the model demonstrates exceptional diagnostic ability for Bacterial Leaf Streak (Class 2). This is due to the highly prominent image features of this disease, where lesions primarily appear as water-soaked dark green or pale yellow streaks, making them easily distinguishable for the model. The classification performance on Healthy Rice (Class 0) is equally outstanding; as the category with the highest proportion in the test set, 172 samples were accurately identified, with only rare instances of misclassification as Rice Blast, Bacterial Grain Rot, or Brown Spot. Healthy rice leaves exhibit uniform color and regular texture features, showing significant differences from diseased plants. However, minor unavoidable misclassifications occurred, likely due to shadows or physical leaf damage in some images. Rice Blast (Class 1), the second most prevalent disease in the test set, also yielded ideal classification results, with its spindle-shaped lesions and brownish-yellow halos serving as key identification markers. Similarly, the model showed high performance in identifying Bacterial Leaf Blight (Class 3) with very few errors, as this disease typically spreads from the leaf tips and edges to form large dark brown areas. While the visual features and textures of these two diseases are distinct, occasional misclassifications still occur, possibly due to boundary blur interference during the early stages of infection. In contrast, the model's classification capability for Bacterial Grain Rot (Class 4) is slightly less robust. This disease occurs in the rice panicle, where lesions transition from water-soaked spots to faded or withered yellowing, causing grain desiccation. The complex structure of the panicle means that lesion morphology and color changes are less clear than those in leaf diseases, presenting a challenge for feature extraction. Brown Spot (Class 5) also poses difficulties because its lesions vary in size and resemble early-stage Rice Blast; both exhibit

brown spots surrounded by yellow halos, leading to some confusion for the model.

In summary, the Improve_ConvNeXt model exhibits excellent classification performance across most rice disease categories, particularly in identifying leaf diseases with distinct features. While accuracy slightly decreases for panicle-based diseases or those with ambiguous early-stage symptoms, the overall performance remains at a high level.◦

3.1.4 Interpretability Analysis

While the improved Improve_ConvNeXt model demonstrates superior performance in rice disease classification, understanding its decision-making process and enhancing model trustworthiness is equally crucial. To this end, Gradient-weighted Class Activation Mapping (Grad-CAM) was introduced for interpretability analysis. Grad-CAM is a technique used to explain the decisions of convolutional neural networks by calculating the gradients of a target class relative to the feature maps of a convolutional layer. This produces a heatmap that highlights the image regions the model focused on most when making a specific prediction, providing an intuitive visual explanation for the model's decision (Selvaraju et al., 2017). A key advantage of this method is that it requires no structural modifications or retraining of the existing network, offering a straightforward way to understand how the model interprets inputs.

Fig. 11 displays the Grad-CAM visualization results of the Improve_ConvNeXt model on typical image samples of healthy rice and the five disease types. Subplots (a) through (f) correspond to the original images and their respective heatmaps for Healthy Rice, Rice Blast, Bacterial Leaf Streak, Bacterial Leaf Blight, Bacterial Grain Rot, and Brown Spot, respectively. In these heatmaps, red regions indicate areas of high model attention, while blue regions represent areas with lower focus.

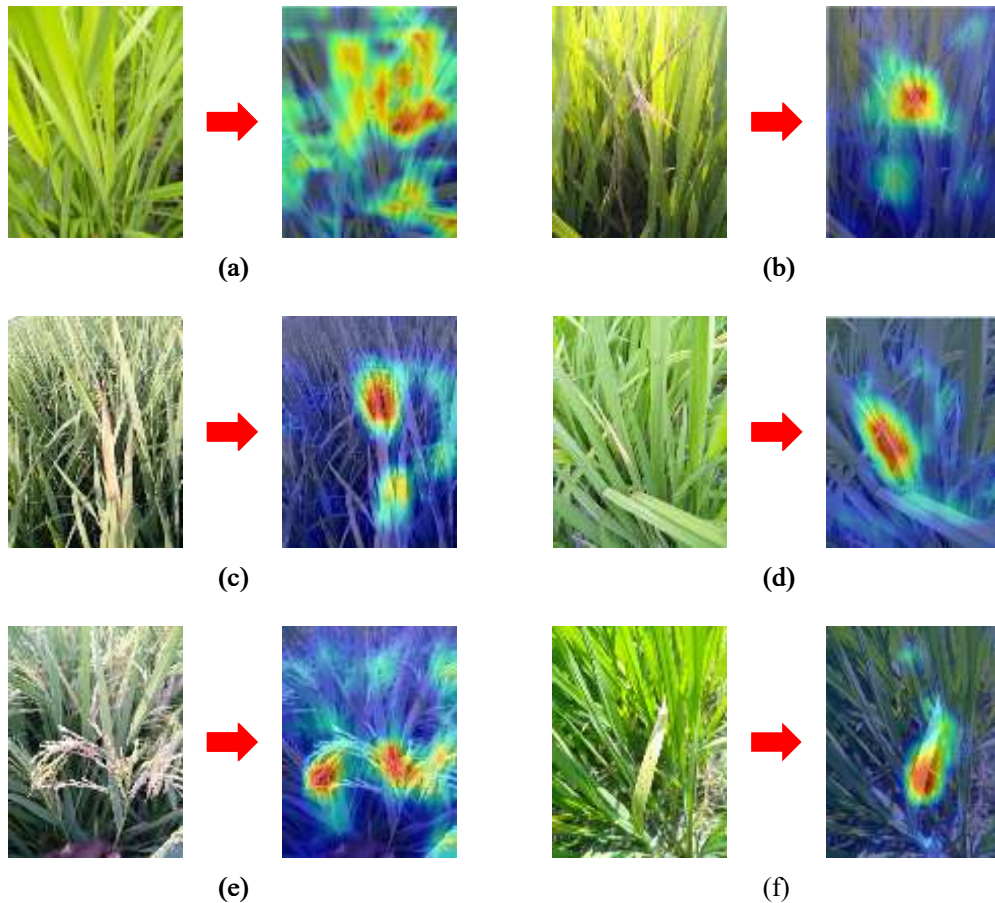


Fig.11 Grad-CAM visualization results of Improve_ConvNeXt

Analyzing the specific heatmaps for each category reveals that the model's focus on Healthy Rice samples is relatively dispersed, with no significant clustering. The attention is broadly distributed across leaf textures and edges, indicating that the model assesses health status by identifying global features such as overall leaf color and texture. For Rice Blast, Bacterial Leaf Streak, and Brown Spot, the heatmaps clearly focus on their respective typical lesion areas, demonstrating that the model successfully localized key pathological features. The Grad-CAM results for Bacterial Leaf Blight are equally impressive, with attention accurately covering the dark brown lesions spreading inward from the leaf tips. Regarding Bacterial Grain Rot, despite the complex structure of the panicle—where contour features and color degradation are less distinct than on leaves—the model still managed to concentrate its attention on the faded and withered areas of the grains. In conclusion, the Grad-CAM visualizations intuitively prove that the Improve_ConvNeXt model effectively learns and focuses on key characteristic regions associated with different rice diseases. The model maintains strong localization capabilities even when processing detailed features and complex structures. This further validates the positive impact of the HAT attention mechanism and SCConv module in enhancing feature extraction and classification performance, while significantly strengthening the interpretability of the model's

decision-making process.

4 Discussion

4.1 Analysis of Classification Performance and Advantages of the Improved Model

Improve-ConvNeXt

The Improve-ConvNeXt model constructed in this study achieved a test accuracy of 96.27% in the fine-grained rice disease classification task, which is significantly superior to the ResNet (He et al., 2016), DenseNet (Huang et al., 2017), and original ConvNeXt baseline models (Liu et al., 2022), showing a clear improvement in overall performance. Compared with the original ConvNeXt, the accuracy increased by 4.85 percentage points, and compared with the traditional ResNet model, it increased by more than 8 percentage points, indicating that the improved model possesses stronger feature extraction and discrimination capabilities in complex field scenarios (Liu and Wang, 2021). From the perspective of the training process, Improve-ConvNeXt converged faster in the early stages of training, stabilizing after approximately 30 rounds, and exhibited lower validation set loss with smaller fluctuations, demonstrating that the model possesses good generalizability while maintaining high precision. On a real-world dataset with class imbalance, the model's precision, recall, and F1-score all improved simultaneously without any single metric being disproportionately high, indicating that the model can handle different disease categories in a balanced manner, effectively reducing misjudgments and missed detections, and better fitting actual field monitoring needs. Compared with traditional convolutional networks, this model is more robust under interference conditions such as complex backgrounds, light variations, leaf occlusion, and tiny lesions. Benefiting from the hierarchical structure and large receptive field design of the ConvNeXt backbone (Liu et al., 2022), the model can simultaneously capture local lesion details and global leaf structures, effectively overcoming the interference of the field environment on classification results and offering greater advantages in distinguishing similar diseases and identifying early-stage diseases.

4.2 Analysis of the Contribution and Effectiveness of HAT Attention and SCConv Modules

The HAT hybrid attention mechanism (Chen et al., 2025) and SCConv spatial-channel reconstruction convolution (Li et al., 2023) introduced in this study are the core factors in improving model performance. By fusing window self-attention and channel attention, HAT enables the model to precisely focus on lesion areas in the spatial dimension and adaptively strengthen key feature channels in the channel dimension while weakening redundant information, significantly enhancing the feature capture capability for subtle and irregular lesions. Combined with the Grad-CAM visualization results (Selvaraju et al., 2017), it can be seen that after adding HAT, the model's heatmap is more concentrated on the lesion areas rather than the background or

healthy tissues, making the decision basis more reasonable and interpretable. The SCConv module reconstructs features from both spatial and channel dimensions; through the split-reconstruct and segment-transform-fuse strategies, it effectively reduces feature redundancy and improves computational efficiency and information density. The spatial reconstruction unit strengthens effective lesion areas, and the channel reconstruction unit optimizes feature channel weights, allowing the model to improve feature expression capability without significantly increasing the computational load (Li et al., 2023). The HAT_SCConv module formed by combining the two retains the global modeling advantages of Transformer-like attention while inheriting the efficiency of convolutional networks, achieving a balance between detail perception capability and computational efficiency, and jointly driving a significant increase in classification performance.

4.3 Analysis of Classification Error Sources and Misclassification Reasons

From the confusion matrix results, the overall classification of the model is reliable, but a small number of misclassifications still exist, mainly concentrated in two scenarios: brown spot vs. rice blast, and bacterial grain rot. Both brown spot and rice blast exhibit brown lesions accompanied by yellow halos in the early stages of onset, and their visual features are highly similar, easily leading to model confusion (Phadikar et al., 2013); bacterial grain rot occurs in the panicle, where the structure is complex, lesion changes are slow, and differences in color and texture are weaker than those in leaf diseases, resulting in lower feature distinguishability and thus a relatively higher probability of misclassification. In addition, objective interference brought by field data collection is also an important source of error: uneven light intensity, overlapping leaf occlusion, differences in shooting angles and distances, and interference from background weeds and soil all reduce the clarity of lesion features. Imbalanced sample distribution also affects classification performance; the sample sizes for healthy rice and rice blast are sufficient, while samples for bacterial leaf streak and bacterial grain rot are relatively scarce, leading to less robust decision boundaries for small-sample categories (Johnson and Khoshgoftaar, 2019). These errors indicate that in real agricultural scenarios, disease feature similarity, data collection conditions, and sample balance remain key factors affecting the stability of deep learning models.

4.4 Research Limitations, Practical Applications, and Future Outlook

This study still has certain limitations: first, the dataset only covers 6 categories, and the diversity of disease types, growth stages, regions, and varieties needs to be expanded; second, the research was conducted based on static images and has not implemented real-time video stream detection or edge-side deployment; third, disease information was completed only as an independent task and has not achieved multi-modal deep fusion with UAV remote sensing and environmental data; fourth, the model has not been lightweight-optimized for mobile and

embedded devices, limiting rapid field deployment. At the application level, this model can directly serve rice field disease monitoring, intelligent diagnosis, precision spraying, and phenotypic breeding, providing core algorithm support for UAV high-throughput phenotypic platforms, improving disease recognition efficiency and accuracy, and reducing manual observation costs. In the future, deepening can continue from three aspects: first, expanding disease datasets across multiple regions, varieties, and growth stages to improve model generalization capability; second, performing lightweight improvements on the model to adapt to mobile phones, edge computing boxes, and UAV onboard equipment (Howard et al., 2019); third, fusing disease classification results with remote sensing, meteorology, and yield data to construct a more precise yield estimation and loss assessment system (Weiss et al., 2020), providing a more complete solution for smart agriculture.

5 Conclusion

This study systematically investigated deep learning-based rice disease classification methods. First, a comprehensive dataset consisting of healthy rice and five types of rice diseases was constructed by integrating field-captured images with public datasets. Subsequently, this chapter proposed a disease classification model named Improve_ConvNeXt, which utilizes ConvNeXt as the backbone network and incorporates the HAT_SCConv module. This module integrates the HAT attention mechanism with SCConv convolution to enhance the model's capability to capture intricate disease features. Experimental results demonstrate that the performance of the Improve_ConvNeXt model is significantly superior to various baseline models. Furthermore, confusion matrix analysis reveals the model's high discriminative power across different disease categories, while Grad-CAM visualization provides an intuitive interpretability analysis for the classification decisions. In summary, the Improve_ConvNeXt model constructed in this chapter not only achieves a substantial increase in classification accuracy within complex field scenarios but also effectively meets the requirements for practical agricultural applications.

Acknowledge:

This work was supported by National Key R&D Program Project (2024YFF0808403).

Declaration of competing interest:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Bock, C.H., Poole, G.H., Parker, P.E., Gottwald, T.R., 2010. Plant disease severity estimated

visually, by digital photography and image analysis, and by hyperspectral imaging. Crit. Rev. Plant Sci. 29, 59–107.

Cai, J., Pan, R., Lin, J., Liu, J., Zhang, L., Wen, X., Chen, X., Zhang, X., 2023. Improved EfficientNet for corn disease identification. Front. Plant Sci. 14, 1224385.

Chen, X., Wang, X., Zhang, W., Kong, X., Qiao, Y., Zhou, J., Dong, C., 2025. Hat: Hybrid attention transformer for image restoration. IEEE Trans. Pattern Anal. Mach. Intell.

Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C., 2023. Activating more pixels in image super-resolution transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22367–22377.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., 2018. Recent advances in convolutional neural networks. Pattern Recognit. 77, 354–377.

GUAN, Z., Tang, J., YANG, B., ZHOU, Y., FAN, D., Yao, Q., 2010. Study on recognition method of rice disease based on image. Chin. J. Rice Sci. 24, 497.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., 2019. Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324.

Hu, W.-J., Fan, J., Du, Y.-X., Li, B.-S., Xiong, N., Bekkering, E., 2020. MDFC-ResNet: An Agricultural IoT System to Accurately Recognize Crop Diseases. IEEE ACCESS 8, 115287–115298. <https://doi.org/10.1109/ACCESS.2020.3001237>

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Huang, X., Chen, Hongxing, Chen, Hui, Fan, C., Tai, Y., Chen, X., Zhang, W., He, T., Gao, Z., 2022. Spatiotemporal heterogeneity of chlorophyll content and fluorescence response within rice (*Oryza sativa* L.) canopies under different cadmium stress. Agronomy 13, 121.

Ji, M., Zhang, K., Wu, Q., Deng, Z., 2020. Multi-label learning for crop leaf diseases recognition and severity estimation based on convolutional neural networks. SOFT Comput. 24, 15327–15340. <https://doi.org/10.1007/s00500-020-04866-z>

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6, 27.

Jung, M., Song, J.S., Shin, A.-Y., Choi, B., Go, S., Kwon, S.-Y., Park, J., Park, S.G., Kim, Y.-M.,

2023. Construction of deep learning-based disease detection model in plants. *Sci. Rep.* 13, 7331. <https://doi.org/10.1038/s41598-023-34549-2>

Li, H., Huang, L., Ruan, C., Huang, W., Wang, C., Zhao, J., 2024. A dual-branch neural network for crop disease recognition by integrating frequency domain and spatial domain information. *Comput. Electron. Agric.* 219, 108843. <https://doi.org/10.1016/j.compag.2024.108843>

Li, J., Wen, Y., He, L., 2023. Scconv: Spatial and channel reconstruction convolution for feature redundancy, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6153–6162.

Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J., 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* 33, 6999–7019.

Liu, J., Wang, X., 2021. Plant diseases and pests detection based on deep learning: a review. *Plant Methods* 17, 22.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11976–11986.

MA, P., ZHOU, A., YAO, Q., YANG, B., TANG, J., PAN, X., 2018. Influence of image features and sample sizes on rice pest identification. *Chin. J. Rice Sci.* 32, 405.

Mohidem, N.A., Hashim, N., Shamsudin, R., Che Man, H., 2022. Rice for food security: Revisiting its production, diversity, rice milling process and nutrient content. *Agriculture* 12, 741.

Pang, D., Wang, H., Ma, J., Liang, D., 2023. DCTN: a dense parallel network combining CNN and transformer for identifying plant disease in field. *SOFT Comput.* 27, 15549–15561. <https://doi.org/10.1007/s00500-023-09071-2>

Phadikar, S., Sil, J., Das, A.K., 2013. Rice diseases classification using feature selection and rule generation techniques. *Comput. Electron. Agric.* 90, 76–85.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 618–626.

Upadhyay, N., Gupta, N., 2024. Detecting fungi-affected multi-crop disease on heterogeneous region dataset using modified ResNeXt approach. *Environ. Monit. Assess.* 196, 610. <https://doi.org/10.1007/s10661-024-12790-0>

Weiss, M., Jacob, F., Duveiller, G., 2020. Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* 236, 111402.

Xu, W., Li, W., 2024. Deep residual neural networks with feature recalibration for crop image

661 disease recognition. CROP Prot. 176, 106488. <https://doi.org/10.1016/j.cropro.2023.106488>

662 Yuan, X., Xu, Q., Wang, T., Gao, L., Zhao, G., Jia, L., Wang, Y., 2026. RiceDetect-Net: a

663 lightweight real-time detection framework for rice diseases. BMC Plant Biol.

664 Zhai, Z., Martínez, J.F., Beltran, V., Martínez, N.L., 2020. Decision support systems for

665 agriculture 4.0: Survey and challenges. Comput. Electron. Agric. 170, 105256.