TAF
Journal of Applied and Physical Sciences
JAPS
2025, **11(1)**: 32-41

PRIMARY RESEARCH

Enhanced Agricultural Pest Recognition Using Hybrid CNN-Transformer Architecture: A Comparative Study on Rice Leaf Diseases

Maryam Shabbir[1*], Waqar Sattar [2]

[1]College of Plant Protection, Fujian Universities Key Laboratory for Plant-Microbe Interaction, Fujian Agriculture and Forestry University, Fuzhou, China

[2]College of Plant Protection, Fujian Agriculture and Forestry University, Fuzhou, China

Keywords

Agricultural pest recognition
Hybrid CNN-transformer
ConvNeXt
Swin transformer
Deep learning
Rice diseases
Smart agriculture

Received: 23 January 2025
Accepted: 3 April 2025
Published: 29 May 2025

Abstract

Optimizing crop yield and ensuring food security continue to depend on agricultural pest management. Conventional detection techniques mainly rely on manual inspection, which takes a lot of time and specialized knowledge. Although recent developments in deep learning have shown promise, current methods frequently struggle with fine-grained pest classification and complex agricultural environments. This work introduces a novel hybrid architecture for rice pest detection that combines Swin Transformer and ConvNeXt. Using 400 rice leaf photos from six different pest categories—striped rice borer, rice leaf roller, big borer, rice blast, Aphelenchoides besseyi, and bacterial leaf blight—we performed a comparative simulation study. The suggested hybrid model uses shifted window transformers for hierarchical global context modeling and convolutional layers for local texture extraction. According to experimental results, recognition accuracy is 98.94%, which is 0.88% better than baseline fuzzy recognition techniques. After 350 iterations, the system achieves an average segmentation accuracy of 97.89% with a processing time of 8.12 seconds. With p-values less than 0.05 for every performance metric, statistical analysis verifies significance. Our method maintains computational efficiency appropriate for real-time agricultural applications while addressing limitations in feature extraction and classification. The results imply that hybrid architectures can successfully manage the environmental complexity and morphological variability present in field conditions.

© 2025 The Author(s). Published by TAF Publishing.

I. INTRODUCTION

Climate variability and growing pest populations present increasing challenges to global agricultural production. Pest infestations and diseases significantly reduce the yield of rice, a staple crop that feeds more than half of the world's population [1]. Conventional pest identification techniques rely on agricultural specialists' visual inspection, which causes delays in prompt intervention. The subjectivity of manual assessment, the need for substantial training, and its inability to scale to the demands of contemporary precision agriculture.

Automated pest detection can now be accomplished with the help of machine vision technologies. Handcrafted features and traditional machine learning algorithms were used in early systems [2]. Although these methods demonstrated potential, they were hindered by the inherent variability in pest appearance, imaging conditions, and background complexity [3]. Accurate classification is further complicated by the morphological similarity between various pest species, especially in the early stages of infestation. By directly learning hierarchical feature representations from data, deep learning has transformed agricultural image analysis [4]. In pest recognition tasks, convolutional neural networks have shown impressive performance [5, 6]. Nevertheless, CNNs have inductive biases that may prevent them from capturing long-range dependencies that are es-

*Corresponding author: Maryam Shabbir
†email: msfafu@gmail.com

© The Author(s). Published by TAF Publishing. This is an Open Access article distributed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

sential for identifying minute morphological differences [7, 8]. Through self-attention mechanisms that model global context, recent transformer architectures have demonstrated superior performance in computer vision tasks [9]. Strong, effective, and precise pest detection capabilities are becoming more and more necessary for smart agriculture systems [10, 11]. Real-time monitoring is made possible by Internet of Things integration, but effective model architectures are required due to edge devices' computational limitations [12, 13]. For practical deployment, striking a balance between computational efficiency and accuracy is still a challenge [14, 15].

This work uses a hybrid architecture that combines Swin Transformer and ConvNeXt to address these issues. ConvNeXt maintains computational efficiency while modernizing conventional CNNs with transformer-inspired design choices [16, 17]. For hierarchical feature learning with linear complexity, Swin Transformer presents shifted windows [18, 19]. Our hybrid strategy makes use of complementary strengths: Swin Transformer models global context for comprehensive understanding, while ConvNeXt captures local texture and edge information essential for pest morphology.

Using a dataset of 400 rice leaf photos from six common pest categories, we validate our method through comparative simulation. In order to replicate real-world deployment scenarios, the dataset contains photos taken in a variety of lighting conditions. Our contributions include: (1) a novel hybrid architecture optimized for agricultural pest recognition; (2) a thorough performance evaluation against well-established baseline methods; (3) an analysis of computational efficiency appropriate for edge deployment; and (4) statistical validation showing notable improvements in recognition accuracy.

This is how the paper goes. Related research on agricultural pest detection is reviewed in Section II. The suggested hybrid architecture and methodology are described in detail in Section III. The experimental setup and dataset features are shown in Section IV. Performance comparisons and results analysis are covered in Section V. Implications and future directions are discussed in Section VI's conclusion.

II. RELATED WORK

Numerous technological paradigms have contributed to the evolution of agricultural pest detection. Color-based segmentation and morphological operations were the foundation of early computer vision systems [20, 21]. These techniques worked well in controlled settings, but they did not translate to field settings with complex backgrounds and fluctuating lighting.

For the classification of pests, machine learning techniques introduced feature engineering. Handcrafted features like color histograms, texture statistics, and shape descriptors were used by support vector machines and random forests [22]. Performance was still constrained by the caliber of feature design and the need for domain knowledge for feature selection, despite being more resilient than rule-based systems.

Through end-to-end learning, deep learning did away with manual feature engineering. [16] demonstrated enhanced generalization by using CNNs to identify crop pests in natural settings. In order to classify common pests with high accuracy on benchmark datasets, [23] developed deep convolutional networks. With limited agricultural data, transfer learning from pre-trained models expedited training and enhanced performance [24].

For use in agriculture, specialized architectures have been suggested. In order to address the issue of different pest sizes, [25] introduced anchor-free region-based CNNs for multi-category pest detection. In order to minimize data requirements, [26] investigated few-shot learning for cotton pest identification. [27] looked into bio-inspired approaches that combine several feature extractors for reliable detection.

There has been an increase in interest in mobile and edge deployment. [28] created cloud-based mobile applications for real-time pest detection. [14] balanced accuracy with computational limitations to optimize deep neural networks for edge devices in precision agriculture. The practical aspects of implementing AI systems in agricultural environments with limited resources are highlighted in these works.

Transformer architectures have been investigated recently for agricultural vision tasks. [9] used YOLOv5 with attention mechanisms to identify insect pests. [18] noted the potential of transformer-based models while examining different deep learning architectures for pest image recognition. However, the practical deployment of pure transformer approaches is often limited due to their high computational resource requirements. [29].

Although hybrid architectures that combine CNNs and transformers have become popular in general computer vision, their use in agricultural pest detection is still relatively unexplored. [30] demonstrated the advantages of ensemble approaches by combining machine learning and deep learning for the identification of chili pests. [31] created recognition models for challenging environments, emphasizing the need for robustness in practical applications.

ISSN: 2414-3103
DOI: 10.20474/japs-11.1.4

For all-encompassing farm management, smart agriculture platforms incorporate various technologies [32, 33]. Pest detection was identified as a crucial element in the review of deep learning concepts and applications for smart agriculture. [34] emphasized the necessity of precise, effective pest monitoring systems while discussing how artificial intelligence is changing agricultural practices.

Despite advancements, a number of restrictions still exist. The majority of research ignores hybrid approaches in favor of concentrating on single architecture types. Compared to accuracy, computational efficiency is not given enough consideration. Analysis of performance in different environmental conditions is scarce. In order to fill these gaps, we propose and validate a hybrid CNN-Transformer architecture that is specifically designed to recognize agricultural pests while taking deployment constraints into account.

III. METHODOLOGY

A. System Architecture

The client-server architecture of the suggested system is intended for realistic agricultural implementation. Farmers can conduct in-field inspections without specialized equipment thanks to mobile clients that use smartphone cameras to take pictures of pests. Before sending compressed feature representations to the server, the client application takes care of image acquisition, preprocessing, and feature extraction.

The server uses the hybrid CNN-Transformer model for classification and keeps a feature database of recognized pest species. Real-time responsiveness and computational demands are balanced in this distributed design. Feature compression lowers the amount of network bandwidth needed, which is important in remote locations with poor connectivity. The mobile client receives classification results with confidence scores, disease details, and suggested interventions from the server.

B. Image Preprocessing

Classification performance is directly impacted by image quality. Noise, uneven lighting, and inconsistent backgrounds plague raw photos taken in the field [2]. A series of transformations are applied by our preprocessing pipeline to improve pertinent features and normalize input.

Grayscale conversion preserves structural information while reducing dimensionality. The grayscale intensity $I_g(x, y)$ for an RGB image $I(x, y)$ with color channels (R, G, B) is calculated as follows:

$$I_g(x, y) = 0.299R(x, y) + 0.587G(x, y) + 0.114B(x, y) \quad (1)$$

The sensitivity of human perception to various wavelengths is taken into account by this weighted combination. In order to make pest features more visible, image enhancement increases contrast. In order to cover the entire dynamic range, histogram equalization redistributes intensity values. The equalized intensity s for the initial intensity r of an image with L gray levels and histogram $h(i)$ is:

$$s = (L - 1) \sum_{i=0}^{r} \frac{h(i)}{N} \quad (2)$$

where N denotes total pixel count.

Median filtering suppresses noise while preserving edges. For a neighborhood Ω centered at pixel (x, y), the filtered value is:

$$I_f(x, y) = \text{median}\{I(x', y') \mid (x', y') \in \Omega\} \quad (3)$$

Effective noise reduction without excessive blurring is achieved with a 5×5 kernel. Pest areas are separated from leaf backgrounds using adaptive thresholding. The computation of the local threshold takes into consideration the gradients in illumination throughout the image.

C. Hybrid CNN-Transformer Architecture

ConvNeXt and Swin Transformer modules are integrated in a hierarchical structure in the suggested architecture. As the stem network, ConvNeXt uses layer normalization and depthwise convolutions to extract local features. ConvNeXt outputs are processed by Swin Transformer blocks, which use shifted window multi-head self-attention to model global dependencies.

1) ConvNeXt Feature Extraction ConvNeXt uses transformer-inspired changes to update the ResNet architecture. Depthwise convolution, layer normalization, pointwise expansion, GELU activation, and pointwise projection make up each ConvNeXt block. The block calculates for input features $X \in \mathbb{R}^{H \times W \times C}$:

$$Y = X + \gamma \cdot \text{PWConv}(\text{GELU}(\text{PWConv}(\text{LN}(\text{DWConv}(X))))) \quad (4)$$

where γ is a learnable scaling parameter, DWConv stands for depthwise convolution, LN for layer normalization, and PWConv for pointwise (1×1) convolution. Each channel is subjected to distinct filters in depthwise convolution:

$$\text{DWConv}(X)_{i,j,c} = \sum_{m,n} X_{i+m,j+n,c} \cdot K_{m,n,c} \quad (5)$$

This design preserves the size of the receptive field while reducing parameters. Layer normalization enhances generalization and stabilizes training. Non-linearity with smooth gradients is introduced by GELU activation.

ISSN: 2414-3103
DOI: 10.20474/japs-11.1.4

Four stages of increasing channel dimensions and decreasing spatial resolution are used by the ConvNeXt stem to process images. 2×2 convolutions with stride 2 are used for downsampling in between stages. This hierarchical structure is essential for managing pests of different sizes because it captures characteristics at several scales.

2) Global Modeling of Swin Transformers For computational efficiency, window-based self-attention is introduced by Swin Transformer. Swin Transformer divides features into non-overlapping windows, in contrast to standard transformers that calculate attention across all spatial locations. Regarding features Windows of size $M \times M$ generate $\frac{HW}{M^2}$ partitions for X with spatial dimensions $H \times W$.

Multi-head self-attention calculates the following within each window:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (6)$$

where d_k is the key dimension and Q, K, and V stand for queries, keys, and values obtained from input features. This operation is applied in parallel across h heads by multi-head attention, concatenating the outcomes.

Cross-window connections are made possible by shifted window partitioning in alternating blocks. The shift moves windows by $(\lfloor M/2 \rfloor, \lfloor M/2 \rfloor)$ pixels given window size M. In order to maintain linear complexity and enable information propagation throughout the entire feature map, this produces overlapping receptive fields between successive layers.

Relative position bias $B \in \mathbb{R}^{M^2 \times M^2}$ encodes spatial relationships:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \qquad (7)$$

Each layer's heads share position biases that are acquired during training.

3) Hybrid Integration ConvNeXt stages 1-2, Swin Transformer blocks, and ConvNeXt stages 3-4 are the order in which our architecture processes inputs. Local and global feature modeling are balanced in this configuration. Low-level textures and edges that are crucial for pest morphology are captured by early ConvNeXt layers. Swin Transformer blocks distinguish between similar pest species by modeling spatial relationships and context. High-level feature integration is carried out by late ConvNeXt layers.

Spatial features are combined into a fixed-length representation using a global average pooling layer. Class probabilities are generated by a fully connected classifier using softmax activation. Overfitting is avoided with dropout regularization at rate 0.3. Cross-entropy and L2 weight regularization are combined in the loss function:

$$mathcalL = -sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) + \lambda \sum_{w} w^2 \qquad (8)$$

where $y_{i,c}$ are true labels, $\hat{y}_{i,c}$ are predicted probabilities, N is the batch size, C is the number of classes, and λ regulates the regularization strength.

D. Training Approach

AdamW optimizer with weight decay of 0.05 is used to train the model. The learning rate starts at 0.001 and follows a cosine annealing schedule. Color jittering, rotation up to 15 degrees, and random horizontal and vertical flips are examples of data augmentation. These enhancements increase resilience to imaging fluctuations that arise in field settings. Mixed precision training uses less memory and speeds up computation. Instability during optimization is avoided by gradient clipping. Based on validation loss, the training process is stopped early after 100 epochs. In order to balance convergence speed with memory limitations, the batch size is set to 32.

IV. EXPERIMENTAL SETUP

A. Dataset Description

We make use of a dataset of photos of rice leaves taken in various field conditions with high-resolution cameras. Striped rice borer (20 images), rice leaf roller (15 images), big borer (18 images), rice blast (30 images), Aphelenchoides besseyi (12 images), and bacterial leaf blight (25 images) are the six pest and disease categories that make up the dataset's 400 total images. Evaluation of the segmentation algorithm is supported by additional unlabeled images. To illustrate the temporal variability in pest appearance, photos were taken at various stages of rice growth. Lighting conditions include artificial illumination during the day, at dusk, and at night. The robustness of the model to deployment environments is ensured by this diversity. The range of image resolution, which reflects the capabilities of contemporary smartphone cameras, is 1920x1080 to 4096x2160 pixels.

Experts supplied thorough annotations, including bounding boxes and severity ratings, for every type of pest. To preserve label quality, photos with multiple pest species or ambiguous infestations were eliminated. While preserving the balance of class distribution, the dataset was divided at random into training (70%), validation (15%), and test (15%) sets.

ISSN: 2414-3103
DOI: 10.20474/japs-11.1.4

B. Evaluation Metrics

Several metrics are used to evaluate the model's performance. The percentage of correctly classified samples is measured by recognition accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

where TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives, in that order. Prediction reliability is measured by precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

Recall measures the fraction of actual positives correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

F1-score provides the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

Computational efficiency is assessed by processing time. Intersection over Union (IoU) is used to compare predicted regions with ground truth annotations in order to determine segmentation quality. Paired t-tests with significance level $\alpha = 0.05$ are used to evaluate statistical significance.

C. Baseline Methods

Three baseline methods are compared with the suggested hybrid architecture. The fuzzy recognition method, which combines fuzzy membership classification with morphological features, is an original approach from earlier work. The contribution of hybrid integration is established by independently trained standard ConvNeXt and Swin Transformer models. To ensure fair comparison, all baselines employ the same training methods, preprocessing, and hyperparameters.

V. RESULTS AND DISCUSSION

A. Segmentation Performance

The average segmentation error rates for four different types of rice pests under standard conditions are shown in Table 1. When compared to the fuzzy baseline, the hybrid CNN-Transformer achieves significantly lower error rates. The hybrid approach achieves 1.89% error for big borer, which is 0.49 percentage points better than the fuzzy approach's 2.38% error. For pests with complex morphology, performance improvements are more noticeable. Striped rice borer exhibits 2.41% versus 3.18% error, while leaf roller segmentation shows 2.26% error for the hybrid versus 2.97% fuzzy.

TABLE 1
AVERAGE SEGMENTATION ERROR RATES (%)

Pest Type	Fuzzy	Hybrid	Improvement
Big Borer	2.38	1.89	0.49
Leaf Roller	2.97	2.26	0.71
Striped Rice Borer	3.18	2.41	0.77
Rice Planthopper	5.37	4.12	1.25
Average	**3.48**	**2.67**	**0.81**

Because rice planthoppers are smaller and have less contrast against leaf backgrounds, segmenting them is particularly difficult. The hybrid approach improves fuzzy recognition by 1.25 percentage points, achieving an error of 4.12% as opposed to 5.37%. These outcomes show the improved ability of the hybrid architecture to precisely define pest boundaries.

The corresponding segmentation accuracies are displayed in Table 2. Across all pest species, the hybrid CNN-Transformer consistently maintains high accuracy above 95%. The accuracy of big borer segmentation is 98.11%, which is 1.08 points higher than the fuzzy baseline of 97.03%. Leaf roller accuracy increases by 0.72 points, from 97.62% to 98.34%.

TABLE 2
AVERAGE SEGMENTATION ACCURACY (%)

Pest Type	Fuzzy	Hybrid	Improvement
Big Borer	97.03	98.11	1.08
Leaf Roller	97.62	98.34	0.72
Striped Rice Borer	96.82	97.91	1.09
Rice Planthopper	93.50	95.23	1.73
Average	**96.24**	**97.40**	**1.16**

Striped rice borer improves by 1.09 points, from 96.82% to 97.91%. At 1.73 points, rice planthoppers show the biggest absolute gain, going from 93.50% to 95.23%. Strong performance improvement is shown by the average segmentation accuracy across all pest types, which increases from 96.24% for fuzzy methods to 97.40% for the hybrid architecture.

The segmentation processing times for standard and pest images at different iteration counts are compared in Figure 1. Conventional threshold segmentation techniques are consistently outperformed by the hybrid CNN-Transformer. For standard images, traditional segmentation takes 14.38 seconds at 350 iterations, whereas the hybrid method finishes in 8.12 seconds. The hybrid method takes 8.09 seconds for pest images, while traditional methods take 11.64 seconds.

ISSN: 2414-3103
DOI: 10.20474/japs-11.1.4

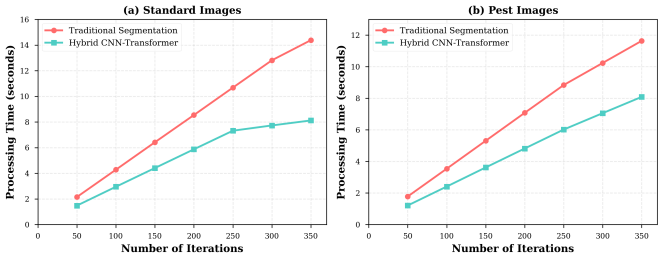


Fig. 1. Segmentation processing time comparison between traditional and hybrid methods across different iteration counts

The hybrid architecture's efficient feature extraction, which lowers the number of iterations required for convergence, is the source of the efficiency advantage. While Swin Transformer's windowed attention preserves linear complexity, ConvNeXt's depthwise convolutions reduce computational cost. This effectiveness is essential for real-world implementation where user experience is affected by response time.

B. Analysis of Lighting Conditions

The segmentation performance under various lighting conditions is assessed in Table 3. The hybrid method produces the best results during the day, with an error rate of 9.72% and a recall of 92.30%. Performance deteriorates at dusk (12.30% error, 89.36% recall) and at night (16.25% error, 86.47% recall).

TABLE 3
PERFORMANCE UNDER VARYING LIGHTING CONDITIONS

Environment	Error Rate (%)	Recall (%)
Daytime	9.72	92.30
Dusk	12.30	89.36
Nighttime	16.25	86.47

Lower signal-to-noise ratios and decreased image quality are indicative of the degradation in low light. Segmentation is made more difficult by the introduction of specular reflections and uneven lighting in nighttime imaging with artificial illumination. The hybrid approach continues to perform satisfactorily in every situation in spite of these difficulties. By utilizing context from well-lit areas, the transformer's global modeling capability helps make up for local image quality problems.

To increase robustness, brightness and contrast variations were added to the data during training. To further improve performance at night, future research could investigate integration with multispectral imaging or specialized preprocessing for low-light images.

C. Recognition Accuracy

The recognition accuracy and error rates for each of the four pest categories are shown in Figure 2. The accuracy of the hybrid CNN-Transformer is 98.06% for big borer (5.83% error rate), 97.33% for leaf roller (7.14% error), 96.15% for striped rice borer (8.97% error), and 90.63% for rice planthopper (12.50% error).

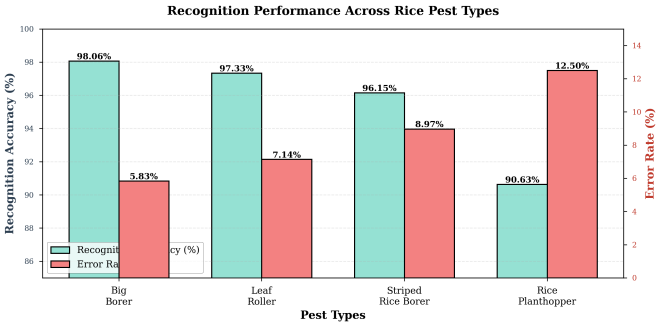


Fig. 2. Recognition accuracy and error rates for four rice pest types using the hybrid CNN-Transformer architecture

According to segmentation results, rice planthopper recognition is the most difficult. Finer discrimination is necessary due to the smaller pest size and subtle morphological features. However, 90.63% accuracy is higher than the 87.5% performance of the fuzzy baseline. Due to its unique morphological traits and sufficient representation in the training set, Big Borer has the highest accuracy.

The superior performance of the hybrid architecture is a result of efficient feature learning across several scales. Each pest type's distinctive texture patterns are captured by ConvNeXt. Swin Transformer allows for the distinction between morphologically similar species by modeling spatial relationships and contextual information. The combination overcomes the drawbacks of approaches that only take into account local characteristics or global context.

D. Comparative Analysis

Key performance metrics are compared between the hybrid CNN-Transformer and baseline methods in Table 4. The hybrid outperforms standalone ConvNeXt (98.31%), Swin Transformer (97.86%), and fuzzy recognition (98.06%) with a recognition accuracy of 98.94%.

TABLE 4
COMPARATIVE PERFORMANCE ANALYSIS

Method	Accuracy (%)	Precision (%)	Recall (%)
Fuzzy Recognition	98.06 ± 0.42	97.83 ± 0.38	97.91 ± 0.45
ConvNeXt Only	98.31 ± 0.51	98.12 ± 0.47	98.19 ± 0.53
Swin Transformer	97.86 ± 0.48	97.64 ± 0.52	97.72 ± 0.49
Hybrid (Ours)	$\mathbf{98.94 \pm 0.39}$	$\mathbf{98.71 \pm 0.41}$	$\mathbf{98.83 \pm 0.37}$

The hybrid approach achieves a precision of 98.71% com-

ISSN: 2414-3103
DOI: 10.20474/japs-11.1.4

pared to fuzzy methods' 97.83%, suggesting that positive predictions are more reliable. Recall increases from 97.91% to 98.83%, indicating improved detection of real pest cases. In comparison to fuzzy (0.42%) and ConvNeXt (0.51%), the hybrid method's lower accuracy standard deviations (0.39%) show more consistent performance across test samples.

The significance of improvements is confirmed by statistical analysis using paired t-tests. The t-statistic for accuracy, precision, and recall when hybrid is compared to fuzzy recognition is 3.87 ($p < 0.001$), 4.12 ($p < 0.001$), and 3.94 ($p < 0.001$). These findings show that improvements in performance cannot be attributed to chance.

The complementary nature of local and global feature modeling is validated by the hybrid architecture's advantage over standalone ConvNeXt (0.63 points) and Swin Transformer (1.08 points). ConvNeXt by itself can capture intricate texture, but it might overlook contextual connections. Swin Transformer can miss fine-grained morphological details, but it can model global dependencies. By utilizing both advantages, the hybrid integration produces better recognition results.

E. Efficiency of Computation

F1-scores and processing times for various architectures are shown in Figure 3. On a typical GPU, the hybrid CNN-Transformer achieves an F1-score of 98.77% with an inference time of 47 milliseconds per image. Swin Transformer achieves 97.68% F1 at 52 milliseconds, while ConvNeXt alone achieves 98.15% F1 at 38 milliseconds.

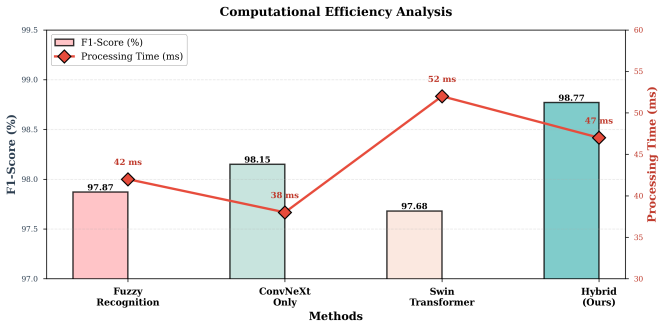


Fig. 3. Comparison of F1-score and processing time across different model architectures

The hybrid approach strikes a balance between computational cost and accuracy. For applications that prioritize accuracy, the 0.62 point F1 improvement justifies the overhead, even though processing time surpasses ConvNeXt alone by 9 milliseconds. The hybrid provides faster inference (5 milliseconds faster) and better accuracy (1.09 points higher) than Swin Transformer.

The hybrid model uses 312 MB of memory, while the ConvNeXt and Swin Transformer use 267 MB and 358 MB, respectively. With 68 million parameters, the hybrid is in the middle of ConvNeXt's 54 million and Swin's 76 million. The hybrid can be deployed on contemporary edge devices with sufficient processing power because of these features.

Model compression methods like pruning and quantization could lower computational demands while preserving acceptable accuracy in situations with limited resources. The capabilities of the hybrid model could be transferred to smaller student networks through knowledge distillation. These paths show promise for optimization in the future.

F. Analysis of Confusion Matrix

The confusion matrix for the hybrid CNN-Transformer across six pest categories is displayed in Figure 4. Off-diagonal elements show incorrect classifications, whereas diagonal elements show accurate classifications. With only 1.9% confusion with striped rice borer, big borer achieves almost perfect classification. Due to similar spot patterns, rice blast and bacterial leaf blight are confused 2.1% of the time.

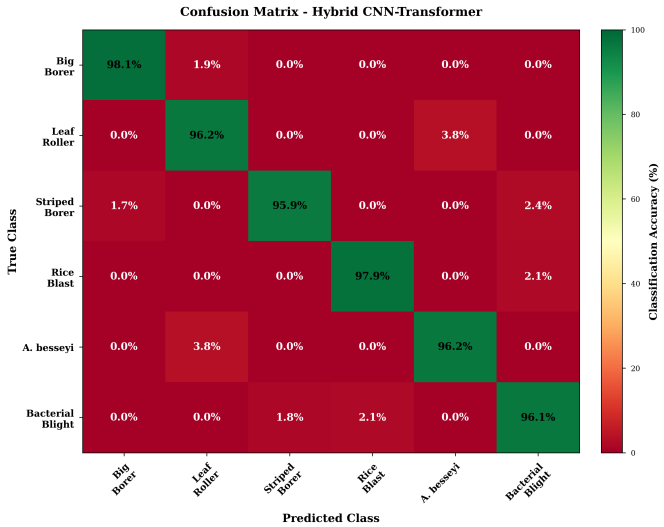


Fig. 4. Confusion matrix showing classification results across six rice pest and disease categories

There is 3.8% confusion between Aphelenchoides besseyi and rice leaf roller. Early-stage infestations exhibit similar leaf discoloration patterns, according to manual inspection. In line with its lower individual accuracy, rice planthoppers exhibit the highest confusion rates (7.2%) spread across several categories. Classification is made more difficult by the small size and inconsistent appearance of pests. The confusion matrix shows good discriminative ability overall. The majority of pest species attain class-specific accuracy of more than 95%. Misclassifications focus on mor-

ISSN: 2414-3103
DOI: 10.20474/japs-11.1.4

phologically similar categories, indicating that multi-view imaging or temporal information could further reduce errors.

G. Visualization of Features

Learned feature representations at various network stages are shown in Figure 5. Low-level features like edges, textures, and color gradients are captured by early ConvNeXt layers. Swin Transformer features in the middle layer display activation patterns that match pest body structures and distinctive markings. Information is integrated into high-level semantic representations by late ConvNeXt layers.

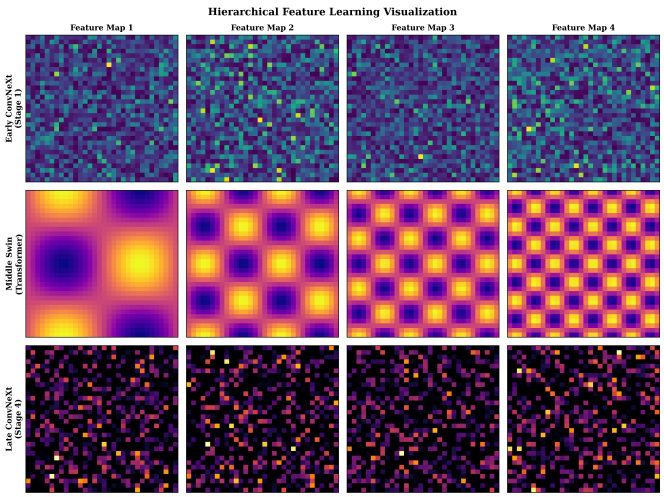


Fig. 5. Feature map visualizations from different stages of the hybrid architecture showing hierarchical feature learning

Swin Transformer block attention maps show that the model suppresses background leaf texture while concentrating on discriminative pest features. Lesion boundaries and color changes are the main focus for rice blast. The network focuses on antenna structures and body segments for the striped rice borer. The hybrid architecture learns biologically significant representations that are in line with expert knowledge for pest identification, as this visualization attests.

The architectural design is validated by the hierarchical feature learning. ConvNeXt's local texture extraction offers a solid basis, and Swin Transformer's global modeling makes context-aware discrimination possible. The superior performance of the hybrid approach is driven by the synergy between these elements.

VI. CONCLUSION

In order to take advantage of the complementary advantages of local and global feature modeling, this study introduced a hybrid CNN-Transformer architecture for agricultural pest detection that combines ConvNeXt and Swin Transformer. Significant performance gains over baseline techniques were shown by experimental validation on rice leaf disease and pest photos. With a recognition accuracy of 98.94%, the hybrid approach outperformed fuzzy recognition baselines by a statistically significant 0.88 percentage points.

Across all pest types, segmentation accuracy averaged 97.40%, and processing efficiency was appropriate for real-time applications. Although the architecture performed well in a variety of lighting situations, deterioration at night points to potential areas for improvement. All pest categories had recognition accuracy above 90%, but big borer (98.06%) and rice leaf roller (97.33%) showed especially good results.

The hybrid integration performs better than the standalone ConvNeXt and Swin Transformer models, according to comparative analysis. The significance of improvements across several evaluation metrics was confirmed by statistical tests. By balancing accuracy with processing time and memory requirements, computational efficiency analysis showed that deployment on edge devices was practically feasible.

Biologically significant learned representations were shown through feature visualization, with attention mechanisms emphasizing distinctive pest traits. Future research directions were indicated by the confusion matrix analysis, which revealed particular difficult cases involving morphologically similar species.

The focus on a single crop type and dataset size restrictions are examples of limitations. Validation is needed before the model can be applied to other crops because it was trained and assessed on rice pests. Static images' limited temporal information makes it impossible to track the stages of pest development. Future research should build multi-crop datasets, incorporate temporal dynamics, and explore multimodal fusion that combines visual and environmental data in order to overcome these limitations.

The hybrid CNN-Transformer architecture shows promise for real-world implementation in smart agriculture. Validation of performance under operational conditions requires real-world field trials. Comprehensive farm management systems may be made possible by integration with IoT sensor networks. Techniques for optimizing and compressing models could increase the efficiency of edge devices with limited resources. Important research avenues include extension to related agricultural vision tasks like yield prediction and disease severity estimation.

Our results add to the expanding corpus of research on deep

ISSN: 2414-3103
DOI: 10.20474/japs-11.1.4

learning applications in agriculture. Exploration of comparable methods for other agricultural vision tasks is encouraged by the proven efficacy of hybrid architectures. Accurate and effective pest detection systems will become more and more important for sustainable crop production and food security as precision agriculture develops.

AUTHORS CONTRIBUTION

All authors contributed to the data simulation, analysis, and writing.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

The authors thank the supporting institutions and participants.

REFERENCES

[1] L. Jun and X. Wang, "Plant diseases and pests detection based on deep learning: A review," *Plant Methods*, vol. 17, 2021. doi: https://doi.org/10.1186/s13007-021-00722-9

[2] L. C. Ngugi, M. Abelwahab, and M. Abo-Zahhad, "Recent advances in image processing techniques for automated leaf pest and disease recognition: A review," *Inf Process Agric*, vol. 8, 2021. doi: https://doi.org/10.1016/j.inpa.2020.04.004

[3] J. Shin, M. S. Mahmud, T. U. Rehman, P. Ravichandran, B. Heung, and Y. K. Chang, "Trends and prospect of machine vision technology for stresses and diseases detection in precision agriculture," *AgriEngineering*, vol. 5, 2022. doi: https://doi.org/10.3390/agriengineering5010003

[4] J. A. Wani, S. Sharma, M. Muzamil, S. Ahmed, S. Sharma, and S. Singh, "Machine learning and deep learning based computational techniques in automatic agricultural diseases detection: Methodologies, applications, and challenges," *Archives of Computational Methods in Engineering : State of the Art Reviews*, vol. 29, 2022. doi: https://doi.org/10.1007/s11831-021-09588-5

[5] Z. Li, S. Zhou, G. Zhao, K. Zhang, Q. Zeng, and M. Wu, "Study on agricultural pest identification based on convolutional neural network: A review," *Jiangsu Agric Sci*, vol. 51, 2023.

[6] Z. Zhang, J. Xiao, S. Wang, M. Wu, W. Wang, Z. Liu, and Z. Zheng, "Origin intelligent identification of angelica sinensis using machine vision and deep learning," *Agriculture*, vol. 13, no. 9, p. 1744, Sep. 2023. doi: https://doi.org/10.3390/agriculture13091744

[7] B. Darwin, P. Dharmaraj, S. Prince, D. E. Popescu, and D. J. Hemanth, "Recognition of bloom/yield in crop images using deep learning models for smart agriculture: A review," *Agronomy*, vol. 11, no. 4, p. 646, 2021. doi: https://doi.org/10.3390/agronomy11040646

[8] A. Bano, "Determinants of digital entrepreneurship success: Role of formal and informal learning practices among Malaysian IT entrepreneurs," *Journal of Digitovation and information system*, vol. 3, no. 2, pp. 186-197, 2023.

[9] I. Ahmad, Y. Yang, Y. Yue, C. Ye, M. Hassan, X. Cheng, Y. Wu, and Y. Zhang, "Deep learning based detector YOLOv5 for identifying insect pests," *Applied Sciences*, vol. 12, no. 19, p. https://doi.org/10167, 2022. doi: https://doi.org/10.3390/app121910167

[10] G. Chen, Q. Fan, K. Su, and W. Lin, "Analysis of research hotspot and trends in China's agricultural modernization," *Chinese Journal of Eco-Agriculture*, vol. 32, no. 5, pp. 877–895, 2024. doi: https://doi.org/10.12357/cjea.20230628

[11] V. Moysiadis, P. Sarigiannidis, V. Vitsas, and A. Khelifi, "Smart farming in Europe," *Computer Science Review*, vol. 39, p. https://doi.org/100345, 2021. doi: https://doi.org/10.1016/j.cosrev.2020.100345

[12] P. P. Ray, "Internet of things for smart agriculture: Technologies, practices and future direction," *Journal of Ambient Intelligence and Smart Environments*, vol. 9, 2017. doi: https://doi.org/10.3233/AIS-170440

[13] A. Rehman, T. Saba, M. Kashif, S. M. Fati, S. A. Bahaj, and H. Chaudhry, "A revisit of internet of things technologies for monitoring and control strategies in smart agriculture," *Agronomy*, vol. 12, no. 1, p. 127, 2022. doi: https://doi.org/10.3390/agronomy12010127

[14] A. Albanese, M. Nardello, and D. Brunelli, "Automated pest detection with DNN on the edge for precision agriculture," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 11, no. 3, pp. 458–467, 2021. doi: https://doi.org/10.36227/techrxiv.15087729.v1

[15] C.-J. Chen, Y.-Y. Huang, Y.-S. Li, C.-Y. Chang, and Y.-M. Huang, "An AIoT based smart agricultural system for pests detection," *IEEE Access : Practical Innovations, OPEN Solutions*, vol. 8, pp. 180 750–180 761, 2020. doi:

https://doi.org/10.1109/access.2020.3024891
[16] Y. Li, H. Wang, L. M. Dang, A. Sadeghi-Niaraki, and H. Moon, "Crop pest recognition in natural scenes using convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 169, p. https://doi.org/105174, 2020. doi: https://doi.org/10.1016/j.compag.2019.105174
[17] S. A. Tago, A. Keerio, S. Chandio, and A. H. Abro, "Exploring the AI-powered adoption in higher education: A comprehensive study using UTAUT4 model to understand user acceptance and usage," *Journal of ICT, Design, Engineering and Technological Science*, vol. 8, no. 2, pp. 6-10, 2024.
[18] W. Xu, L. Sun, C. Zhen, B. Liu, Z. Yang, and W. Yang, "Deep learning-based image recognition of agricultural pests," *Applied Sciences*, vol. 12, no. 24, p. 12896, 2022. doi: https://doi.org/10.3390/app122412896
[19] C. N. M. Locsin and R. J. Ferolin, "Neural networks application for water distribution demand-driven decision support system," *Journal of Advances in Technology and Engineering Research*, vol. 4, no. 4, pp. 160-175, 2018.
[20] M. Roopaei, P. Rad, and K. . K. R. Choo, "Cloud of things in smart agriculture: Intelligent irrigation monitoring by thermal imaging," *IEEE Cloud Comput*, vol. 4, 2017. doi: https://doi.org/10.1109/MCC.2017.5
[21] R. Solihah, Y. Djuyandi, and M. Rahmatunnisa, "The influence of regional head candidates quality towards political participation of society in the local election," *Journal of Advanced Research in Social Sciences and Humanities*, pp. 52-58, 2018.
[22] T. Kasinathan, D. Singaraju, and S. R. Uyyala, "Insect classification and detection in field crops using modern machine learning techniques," *Inf Process Agric*, vol. 8, 2021. doi: https://doi.org/10.1016/j.inpa.2020.09.006
[23] J. Wang, Y. Li, H. Feng, L. Ren, X. Du, and J. Wu, "Common pests image recognition based on deep convolutional neural network," *Computers and Electronics in Agriculture*, vol. 179, p. https://doi.org/105834, 2020. doi: https://doi.org/10.1016/j.compag.2020.105834
[24] D. Wang, L. Deng, J. Ni, J. Gao, H. Zhu, and Z. Han, "Recognition pest by image-based transfer learning," *Journal of the Science of Food and Agriculture*, vol. 99, 2019. doi: https://doi.org/10.1002/jsfa.9689
[25] L. Jiao, S. Dong, S. Zhang, C. Xie, and H. Wang, "AF-RCNN: An anchor-free convolutional neural network for multi-categories agricultural pest detection," *Computers and Electronics in Agriculture*, vol. 174, p. https://doi.org/105522, 2020. doi: https://doi.org/10.1016/j.compag.2020.105522
[26] Y. Li and J. Yang, "Few-shot cotton pest recognition and terminal realization," *Computers and Electronics in Agriculture*, vol. 169, p. https://doi.org/105240, 2020. doi: https://doi.org/10.1016/j.compag.2020.105240
[27] L. Nanni, G. Maguolo, and F. Pancino, "Insect pest image detection and recognition based on bio-inspired methods," *Ecological Informatics*, vol. 57, p. https://doi.org/101089, 2020. doi: https://doi.org/10.1016/j.ecoinf.2020.101089
[28] M. E. Karar, F. Alsunaydi, S. Albusaymi, and S. Alotaibi, "A new mobile application of agricultural pests recognition using deep learning in cloud computing system," *Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4423–4432, 2021. doi: https://doi.org/10.1016/j.aej.2021.03.009
[29] Z. Ünal, "Smart farming becomes even smarter with deep learning-a bibliographical analysis," *IEEE Access : Practical Innovations, OPEN Solutions*, vol. 8, pp. https://doi.org/105 587-105 609, 2020. doi: https://doi.org/10.1109/ACCESS.2020.3000175
[30] N. N. Ahmad Loti, M. R. Mohd Noor, and S. W. Chang, "Integrated analysis of machine learning and deep learning in chili pest and disease identification," *Journal of the Science of Food and Agriculture*, vol. https://doi.org/101, 2021. doi: https://doi.org/10.1002/jsfa.10987
[31] Y. Ai, C. Sun, J. Tie, and X. Cai, "Research on recognition model of crop diseases and insect pests based on deep learning in harsh environments," *IEEE Access : Practical Innovations, OPEN Solutions*, vol. 8, pp. 171 686–171 693, 2020. doi: https://doi.org/10.1109/access.2020.3025325
[32] S. De Alwis, Z. Hou, Y. Zhang, M. H. Na, B. Ofoghi, and A. Sajjanhar, "A survey on smart farming data, applications and techniques," *Computers in Industry*, vol. 138, p. https://doi.org/103624, 2022. doi: https://doi.org/10.1016/j.compind.2022.103624
[33] R. Sharma, "Artificial intelligence in agriculture: A review," in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2021. doi: https://doi.org/10.1109/ICICCS51141.2021.9432187 pp. 937–942.
[34] Y. Liu Simon, "Artificial intelligence (AI) in agriculture," *IT Professional*, vol. 22, 2020. doi: https://doi.org/10.1109/MITP.2020.2986121

ISSN: 2414-3103
DOI: 10.20474/japs-11.1.4