Hybrid ConvNeXt–ViT Feature Extraction for Rice Blast Disease Severity Classification

Original Scientific Paper

M K Shwetha *
Department of Computer Science and Engineering
PES College of Engineering, Mandya
Research Center: Vidyavardhaka College of Engineering, Mysuru
Visvesvaraya Technological University, Belagavi, Karnataka, India
shwethamk@pesce.ac.in

V Ravi Kumar
Department of Computer Science and Engineering
Vidyavardhaka College of Engineering, Mysuru
Visvesvaraya Technological University, Belagavi, Karnataka, India
ravikumarv@vvce.ac.in

*Corresponding author

Abstract *– Rice blast is one of the most destructive fungal diseases affecting rice cultivation, which causes large yield losses and creates a major risk to the world's food security. In addition to early detection, accurate severity-level assessment is crucial for effective disease management. A hybrid deep learning method based on ConvNeXt and Vision Transformer (ViT) models is proposed for assessing rice blast disease severity in the present work. A new well-organised dataset comprising 1,440 rice leaf images annotated into seven severity levels ranging from 1–10% to 61–70% based on the lesion area coverage is utilised in the model. The proposed hybrid ConvNeXt–ViT model has two parallel feature extraction branches, where ConvNeXt captures spatial and textural patterns, while ViT learns global contextual representations. Complementary feature representations from both branches are fused through a unified classification head to enable fine-grained severity prediction. A series of tests was carried out to evaluate the effect of different fusion strategies, pre-trained weights, patch sizes, and batch sizes on model performance. The proposed approach demonstrates consistent training behaviour and generalisation across different configurations. Comparative evaluation against state-of-the-art CNN models such as ResNet, EfficientNet, VGG, MobileNet, and Xception, as well as a transformer-based baseline shows that the proposed approach achieves superior performance. The ConvNeXt–ViT model achieves a classification accuracy of 92.63%, demonstrating its potential for automated rice blast severity assessment in agriculture applications.*

Keywords: *Rice blast disease, Severity classification, ConvNeXt, Vision Transformer (ViT), Feature fusion, Hybrid deep learning*

Received: February 25, 2026; Received in revised form: May 4, 2026; Accepted: May 31, 2026

1. INTRODUCTION

India's economy depends heavily on agriculture, which contributes around 18% to the national GDP and employs nearly two-thirds of the workforce. This sector helps to improve economic growth through agricultural exports [1]. As a staple food crop, rice farming supports rural livelihoods and contributes to agricultural GDP through production and processing activities [2]. However, rice productivity is challenged by multiple factors, including climate-related changes such as variations in rainfall, temperature, droughts, and floods. Additionally, biotic stresses such as pest infestations and diseases, particularly blast and brown spot, continue to reduce rice yield stability and threaten rice production [3].

Rice blast is one of the most destructive diseases in rice farming, making rice production very challenging. This disease can cause damage to rice plants at any stage of growth and is caused by the fungus Magnaporthe oryzae. The disease first appears as small lesions on the leaves, later spreads to the grains, panicle axes, nodes, necks, and leaf sheaths [4]. As the infection worsens, both the number and spread of lesions on the

leaf surface increase, reflecting higher levels of disease severity. This severity is commonly assessed on a scale of 0 to 9, representing the proportion of plant tissue damaged by lesions [5]. Severe blast outbreaks can result in yield losses ranging from 10% to as high as 50% under favorable environmental conditions [6, 7].

Proper categorization of blast disease severity is required to determine the level of infection and minimizing yield losses. Also, it simplifies decision-making for sustainable rice production and allows for more effective use of resources [8].

Traditionally, specialists use visual inspection to identify the severity of rice blast by looking at leaves for lesions and other symptoms. This method is laborious and time-consuming because it relies on experts' experience and judgment. Discrepancies between the evaluator's decisions and early-stage symptoms can lead to inconsistent or inaccurate assessments [9].

Most existing studies focus on the detection of rice leaf diseases, while relatively few have addressed the more challenging task of disease severity classification. A few studies concentrate on two to three severity levels, such as mild, moderate, and severe. For effective disease management, models capable of classifying rice blast disease into severity levels and capturing subtle differences between infection stages are needed. To address this, a newly segregated rice blast severity-level image dataset is created and used to train the proposed model. The proposed model presents a ConvNeXt–ViT model for rice blast disease severity classification. It comprises two parallel feature-extraction branches: ConvNeXt captures fine-grained spatial and textural information, and the Vision Transformer (ViT) learns global contextual dependencies. Features extracted from both branches are fused to generate the final severity-level prediction. This work presents the following major contributions:

1. A newly segregated rice blast severity-level image dataset has been developed, encompassing seven se-verity levels ranging from early-stage lesions to exten-sive infection.
2. A hybrid ConvNeXt–ViT framework is designed to combine complementary local spatial–textural features and global contextual representations for fine-grained disease severity classification.
3. A systematic investigation of multiple fusion strategies (Early, Intermediate, and Late Fusion) is conducted, demonstrating that the proposed Late Fusion approach achieves superior performance.
4. Extensive experimental evaluation, including compari-sons with CNN-based models and a transformer-based baseline (ViT), validates the effectiveness and robust-ness of the proposed approach under different training configurations.

2. RELATED WORKS

The Whale Optimization Algorithm and Adaptive Particle Swarm Optimization were combined in a hybrid bio-inspired optimization approach [10] to enhance rice disease classification using a CNN. The PlantVillage dataset was used to test the strategy, with a focus on segmentation, pre-processing, and optimal feature selection. ANN, SVM, and conventional CNN models were used to compare performance. According to the results, the hybrid approach attained a high accuracy of 97.5%. However, the reliance on the PlantVillage dataset, which is composed of controlled laboratory images, raises concerns about the model's generalizability to real-world field conditions where lighting, background, and image quality vary considerably. Furthermore, the study focuses on disease identification without addressing severity-level classification. The proposed ConvNeXt–ViT framework addresses these limitations by training on a field-representative dataset with seven fine-grained severity levels.

Wang [11] used a convolutional block attention module to develop a double-branch DCNN incorporating a Convolutional Block Attention Module (CBAM) to improve feature representation for rice disease classification. The dual-branch design allowed the model to capture complementary feature representations from different network pathways. The results show that their model outperforms the conventional models VGG-16, MobileNet-V2, EfficientNet-B1, ResNet-50, ResNet50 + CBAM, GoogLeNet, and Inception-V2 in terms of accuracy (98.73%). The model was evaluated solely on disease identification without addressing severity-level classification, which is a critical requirement for agronomic decision-making in real field conditions.

A hybrid model was developed for classifying blast disease severity into four levels based on the percentage of leaf area infected [12]. The dataset, comprising 1908 images from various sources (Mendeley, Kaggle, GitHub, and UCI), was used to train the system. Features were extracted using a CNN, and severity detection was performed with an SVM. The resulting model achieved 97% accuracy, demonstrating the viability of combining deep feature extraction with traditional classifiers for severity prediction. However, the four-level severity categorization remains coarse for practical field applications that demand more granular disease monitoring, and the hybrid CNN-SVM design may limit end-to-end optimization since the feature extraction and classification stages are not jointly trained. Furthermore, the aggregation of images from heterogeneous sources without standardization may introduce inconsistencies in image quality and annotation.

Ritharson et al. [13] introduced the DeepRice framework for the classification of rice leaf disease into five categories—healthy, blast, blight, tungro, and brown spot—with each disease further divided into mild and severe levels. A customized VGG16 model achieved a

high accuracy of 99.94%, demonstrating the effectiveness of CNN-based approaches. However, the binary severity classification is relatively coarse and may not fully capture the progressive nature of disease development. Additionally, relying on a single CNN model limits the ability to capture both local and global features. In contrast, the proposed ConvNeXt–ViT framework uses a dual-branch architecture to model both aspects and provides classification across seven fine-grained severity levels for more detailed assessment.

Roy and Kukreja [14] addressed the challenge of rice leaf disease classification and severity estimation using Vision Transformers (ViTs) — an area that had been largely unexplored before their work. Their dataset comprised 3,345 images spanning ten disease categories annotated at three severity levels: mild, moderate, and severe. The contribution lay in leveraging the self-attention mechanism of ViT to jointly model disease identity and progression stage within a single framework. The model achieved a weighted average F1-score of 77.94% for severity prediction. While this work demonstrated the feasibility of transformer-based models for rice disease analysis, its relatively modest F1-score and coarse three-level severity highlight the need for more discriminative architectures capable of fine-grained severity classification — a gap addressed by the proposed ConvNeXt–ViT model in this study.

A multimodal deep learning framework integrating visual and environmental modalities — such as temperature, humidity, and rainfall — alongside RGB leaf images was proposed in [15] for simultaneous rice leaf disease classification and severity prediction. A dual-output CNN architecture was employed to jointly predict disease category and severity level within a unified training pipeline, eliminating the need for separate models for each task. However, the dependency on environmental sensor data introduces practical limitations in resource-constrained or remote agricultural settings where such ancillary data may not be readily available, and the fusion of heterogeneous modalities adds architectural complexity.

3. PROPOSED METHODOLOGY

3.1. DATASET

A dataset of 1,440 rice blast images was collected from Mendeley Data [16]. These images were manually annotated into seven severity levels by an agriculture university professor specializing in plant pathology, following standard plant pathology practices. The severity levels were defined based on the estimated percentage of lesion area coverage (1–10%, 11–20%, 21–30%, 31–40%, 41–50%, 51–60%, and 61–70%), as shown in Fig. 1.

Fig. 1 also shows the number of images in each severity level. This newly segregated rice blast severity-level image dataset has been uploaded to the Mendeley Data repository [17].

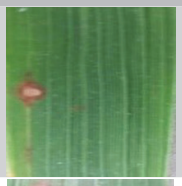
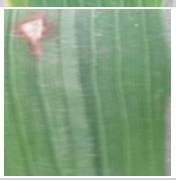
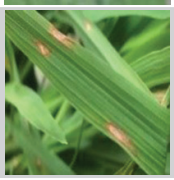
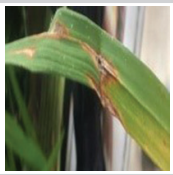
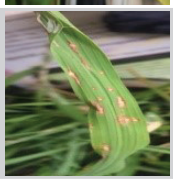
Sl. No.	Severity levels	Sample	No. of images
1	1-10%		72
2	11-20%		154
3	21-30%		235
4	31-40%		343
5	41-50%		316
6	51-60%		266
7	61-70%		81
	Total number of images		**1440**

Fig. 1. Rice blast disease severity levels Dataset

3.2. PROPOSED CONVNEXT–VIT MODEL

To meet the input requirements of the ConvNeXt and Vision Transformer backbones, all rice blast imager were scaled to 224 × 224 pixels and converted from BGR to RGB format. For the validation and testing sets, resizing and normalization were applied. Data augmentation, such as random rotations and horizontal and vertical flips, was applied to the training set to enhance generalization and prevent overfitting.

Fig. 2 illustrates the Hybrid ConvNeXt–ViT [18] architecture used for disease-specific severity prediction for Blast disease. The image is first pre-processed and then passed through two parallel feature-extraction

paths: ConvNeXt [19] and the Vision Transformer (ViT). ConvNeXt processes the input images through stacked convolutional stages to emphasize local details [20], while the ViT branch models long-range dependencies by analyzing relationships among image patches using transformer encoder layers.

The features learned from both branches are then combined and forwarded to a classification module consisting of a fully connected layer followed by a softmax function. Attention-based characteristics and localized convolutional information are successfully combined in this hybrid architecture [21].

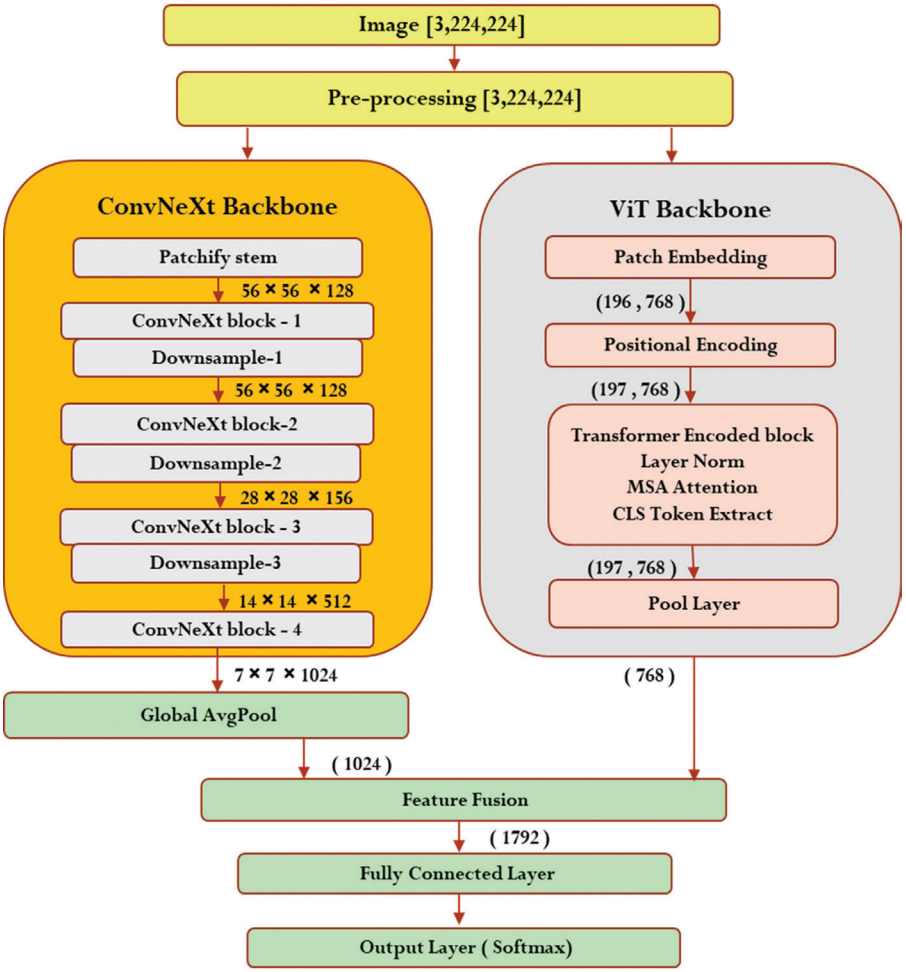


Fig. 2. Hybrid ConvNeXt–ViT architecture

The ConvNeXt backbone focuses on capturing fine-grained spatial features such as lesion boundaries, texture variations, spot geometry, and chlorotic regions, which are essential for distinguishing different severities of rice diseases [22].

The ConvNeXt backbone begins with a stem convolution, followed by multiple ConvNeXt blocks that progressively refine feature representations, and concludes with global pooling.

The Stem is also called patchify Convolution. The Stem of the ConvNeXt backbone is implemented as a 2D convolutional layer designed to convert the raw input image into a compact grid of patch embedding. It uses a convolution kernel of size applied with a stride of 4 and zero padding, meaning that the filters slide across the image in non-overlapping steps. Given an input image $X \in R^{3\times224\times224}$ this operation 4×4, partitions the leaf image into 4×4 pixel patches and transforms each patch into a learned feature representation $F_{stem}(X)$ [23] as expressed in Eq. (1).

$$F_{stem}(X) = Conv_{4\times4,s=4}(X) \tag{1}$$

For ConvNeXt-Base, the Stem generates 128 output channels, producing an output feature map of size 56×56×128. The output shape is denoted as tensor $T1$, and it is expressed in Eq. (2).

$$T_1 \in R^{56\times56\times128} \tag{2}$$

ConvNeXt transforms it through a sequence of hierarchical convolutional stages. The complete transformation is expressed in Eq. (3).

$$F_{conv} = B_4(D_3(B_3(D_2(B_2(D_1(B_1(T_1))))))) \tag{3}$$

Where B_i stands for the ConvNeXt block in the i^{th} stage, each of which is in charge of using GELU activation, normalization, and depth-wise convolutions to gradually improve the spatial and contextual representations. The component D_i represents the downsampling layer located between consecutive stages, which reduces the spatial dimensions while increasing

channel depth, thereby enabling hierarchical feature abstraction. Finally, F_{conv} corresponds to the final feature map, containing the highest-level semantic and texture descriptors extracted by the network.

$$B_i(T_i) = T_i + MLP(\sigma(DWConv(T_i))) \quad (4)$$

The ConvNeXt block at stage i, as defined in Eq. (4), is designed to refine the input feature map. The input tensor T_i, which contains spatial and semantic information extracted from earlier layers. In the first step, a depth-wise convolution, DWConv(T_i), is applied to process each channel independently using individual spatial kernels. This preserves channel-wise structure and captures disease visual cues such as lesion edges and texture variations. The resulting features σ (.) pass through a GELU activation [24] and a multi-layer perceptron (MLP) to enable interaction across channels. Finally, a residual connection adds the original tensor T_i back to the transformed output, which ensures that both original and refined information are retained.

$$D_i(T_i') = Conv_{2\times2,s=2}(T_i') \quad (5)$$

The down-sampling layer between the ith Stage and $(i+1)^{th}$ Stage of the ConvNeXt is represented by the operation $D_i(T'_i)$ shown in Eq. (5). It applies a 2×2 convolution with stride 2 to reduce the spatial dimensions of the feature map by half. As a result, the output of D_i becomes a more compact feature representation that contains richer high-level information. This down-sampled tensor then forms the input to the next hierarchical stage of ConvNeXt, enabling deeper layers to process disease regions with an enlarged receptive field and enhanced discrimination capacity.

$$z_{conv} \in GAP(F_{conv}) \quad (6)$$

Following the hierarchical progression of ConvNeXt blocks and down-sampling layers, the final extracted feature map obtained from Eq. (3) is F_{conv}, which is subsequently passed through a Global Average Pooling (GAP) layer.

The GAP operation shown in Eq. (6) compresses the spatial dimensions F_{conv} by computing the average value of each channel across all spatial locations. This transforms the feature map z_{conv} from a tensor of size 7×7×1024 into a compact vector of size 1024, shown in Eq. (7).

$$z_{conv} \in R^{1024} \quad (7)$$

The ViT model first converts the input rice leaf image x into a sequence of fixed-size patches, a process known as patch embedding. The image is divided into 'n' patches [25], and linearly projected using a learnable weight matrix E. This produces the patch embedding represented in Eq. (8).

$$x_i = [x_1E;\ x_2E;\ x_3E;\ldots\ldots\ldots x_nE;\ ] \quad (8)$$

Where E is a learnable weight matrix for patch projection, and 'x_iE' is the linear embedding of the i^{th} patch. Through this step, the 2D image is effectively transformed into a 1D token sequence compatible with the transformer encoder. Following patch embedding, a learnable classification token x_{lcls} and positional embedding E_{pos} are added to retain spatial information that would otherwise be lost during the flattening process [26]. The sequence of embedded patches Z_0 is denoted in Eq. (9).

$$Z_0 = [x_{1cls};x_1E;\ x_2E;\ x_3E;\ldots\ldots\ldots x_nE;\ ] + E_{pos} \quad (9)$$

After positional encoding, the patch sequence passes through multiple transformer encoder layers, each containing a Multi-Head Self-Attention (MHSA) mechanism followed by a Feed Forward Network (FFN) [27]. In the self-attention operation, the attention output is computed and shown in Eq. (10).

$$Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (10)$$

Where the query, key, and value matrices are denoted by Q, K, and V. The input token sequence Z_0 is linearly projected using learnable weight matrices to produce these matrices, $Q = Z_0\ W_Q$, $K = Z_0\ W_K$, and $V = Z_0\ W_V$, where W_Q, W_K, W_V are learnable projection matrices. Then ViT uses multiple attention heads [28] to capture diverse contextual relationships within the rice leaf image. The multi-head attention output is defined in Eq. (11).

$$MultiHead(Z_0) = [head_{1;\ldots\ldots\ldots}\ head_h]\ W^o \quad (11)$$

Where W^o is the output of the projection matrix and $head_i$ = Attention (Q_i, K_i, V_i). A residual connection is used to combine the output of the Multi-Head Self-Attention (MHSA) module with the original input Z_0 through ($Z_1 = Z_0$ + MultiHead(Z_0)). Each token is then passed independently through a Feed-Forward Network (FFN). Eq. (12) represents the FFN used within each Transformer encoder layer [29]. The FFN is applied independently to every token and is defined as

$$z_{vit} = FFN(Z_1) = GELU(Z_1W_1 + b_1)W_2 + b_2 \quad (12)$$

In Eq. (12), W_1 and W_2 are learnable weight matrices that project the token into a higher-dimensional intermediate space and then back to the model dimension. The terms b_1 and b_2 are learnable bias parameters associated with each linear transformation. The non-linear Gaussian Error Linear Unit (GELU) activation function introduces smooth gating behavior, enabling richer and more expressive feature transformations. This ViT branch outputs a high-level embedding that encapsulates long-range dependencies, subtle texture differences, and global disease-spread patterns. After several encoder layers, a pooler operation is applied to obtain the final 768-dimensional feature vector z_{vit} [30], and it is shown in Eq. (13).

$$z_{vit} \in R^{768} \quad (13)$$

Following the extraction of complementary feature representations from ConvNeXt and the ViT branch [31], the hybrid architecture fuses [32] these two modalities to create a unified descriptor for severity-level prediction. The 1024-dimensional ConvNeXt feature vector given in Eq. (6) and the 768-dimensional ViT fea-

ture vector presented in Eq. (13) are concatenated to generate a joint feature embedding z_{fused} [33] described in Eq. (14). The fused output has a total dimensional size of 1792.

$$z_{fused} = [z_{conv} ; z_{vit}] \in R^{1792} \quad (14)$$

This fused vector is then transformed through a series of nonlinear mappings to enhance representational capacity. The first fusion layer applies a linear transformation followed by a ReLU activation, producing the intermediate representation h_1 as defined in Eq. (15)

$$h_1 = ReLU(W_1 z_{fused} + b_1) \quad (15)$$

Where b_1 denotes the bias vector, and W_1 is the weight matrix of the first fusion layer. Together, they constitute the learnable parameters of the first nonlinear mapping applied to the fused feature vector. A subsequent nonlinear transformation is applied to refine the features further and introduce additional abstraction, resulting in h_2, as described in Eq. (16). In this transformation, the term b_2 serves as the bias vector of the second fusion layer.

$$h_2 = ReLU(W_1 h_1 + b_2) \quad (16)$$

To reduce overfitting and improve generalization, dropout regularization is applied to the second fusion layer. As a result, the fusion block produces a compact 512-dimensional representation illustrated in Eq. (17).

$$h_2 \in R^{512} \quad (17)$$

Following the feature fusion stage, the transformed representation h_2 is forwarded to the final classification module, where the model determines the disease-specific severity level. After the fused feature vector has gone through a fully connected layer, a softmax activation with a weight matrix W_c and a bias vector b_c is employed to transform its output $\hat{y}$ into class-probability values. This operation is expressed in Eq. (18). The resulting vector $\hat{y}$ encodes the predicted probability values over all possible severity classes for the identified disease, serving as the final output of the Hybrid ConvNeXt–ViT severity-level prediction pipeline.

$$\hat{y} = softmax(W_c h_2 + b_c) \quad (18)$$

For the total number of severity classes 's_c', the probability value corresponding to each class k, denoted as $\hat{y}_k$, is obtained using the formulation shown in Eq. (19). This equation gives the probability associated with each class based on the output logits across all classes 's_c'. The predicted severity level $\hat{c}$, corresponds to the class with the maximum probability is expressed in Eq. (20).

$$\hat{y}_k = e_k^z / (\sum_{j=1}^{sc} e_j^z) \quad (19)$$

$$\hat{c} = \underset{k}{\text{argmax}}\, \hat{y}_k \quad (20)$$

3.3. PERFORMANCE METRICS

As stated in Eq. (21), accuracy is used to gauge how accurate the model predictions are overall.

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \quad (21)$$

Eq. (22) is used to compute precision, which evaluates the reliability of positive predictions.

$$Precision = TP/(TP+FP) \quad (22)$$

Recall represents the proportion of correctly identified disease severity samples, as given in Eq. (23).

$$Recall = TP/(TP+FN) \quad (23)$$

Eq. (24) express F1-score.

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (24)$$

Where TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative, respectively.

4. RESULTS

Python was used in a Jupyter Notebook environment to carry out all experiments. An NVIDIA A100 GPU was utilised to speed up both training and inference processes throughout the model's implementation on Google Colab.

The suggested ConvNeXt–ViT hybrid model is consistent with the performance findings for blast disease severity classification, which are presented in Table 1. Every evaluation measure shows improvement as the number of training epochs rises. The model performs moderately throughout the first ten epochs, with metric values near 0.80, indicating that it has started selecting important blast-related information.

As the epochs increase to 20 and 30, the model shows improvement with Accuracy rising to 0.8525 and 0.8802, respectively. The same rising trend was seen in other parameters, which showed improved classification consistency and feature extraction. All evaluation measures surpassed 0.90 by 40 epochs, demonstrating a high degree of blast disease symptom detection. The accuracy reached at 0.9263 after 50 epochs, indicating the best performance. The bar chart in Fig. 3 depicts an increase in every performance metric as the number of epochs increases, supporting the numerical results recorded in Table 1.

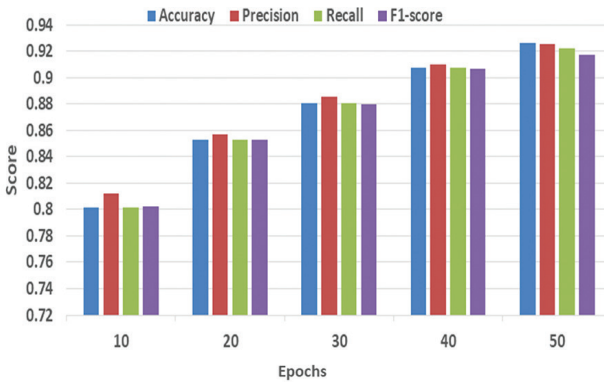


Fig. 3. Comparison of Results of the Hybrid Model ConvNeXt+ViT for Blast disease

Table 1. Results of the Hybrid ConvNeXt+ViT model for Blast disease severity

Epochs	Accuracy	Precision	Recall	F1-score
10	0.8018	0.8121	0.8018	0.8026
20	0.8525	0.8568	0.8525	0.8532
30	0.8802	0.8858	0.8802	0.8798
40	0.9078	0.9097	0.9078	0.9068
50	0.9263	0.9253	0.9223	0.9175

The Blast disease model's accuracy and loss curves of training and validation sets are shown in Fig. 4. Both accuracy curves rise quickly during the initial epochs, indicating fast learning of key Blast features, and later stabilize above 0.9, reflecting strong generalization. The training and validation loss drop sharply at the beginning and continue to decline smoothly, with both curves closely aligned. This tight convergence indicates stable learning without overfitting.

Fig. 5 presents the confusion matrix for Blast severity classification. The confusion matrix provides detailed class-wise performance insights for the proposed model. It can be observed that the majority of predictions lie along the diagonal, indicating accurate classification across all severity levels. Misclassifications are minimal and primarily occur between adjacent severity classes, such as 21–30% and 31–40%, or 41–50% and 51–60%, which is expected due to their visual similarity. Importantly, there are no significant misclassifications across distant classes, demonstrating the model's ability to preserve severity order. These results confirm that the proposed model performs consistently across different classes despite class imbalance.

Table 2 presents a comparative evaluation of individual backbone models and hybrid fusion strategies. To ensure comparison with transformer-based approaches, a Vision Transformer (ViT) model was implemented as a baseline. The ViT achieved an accuracy of 0.7972, slightly outperforming the ConvNeXt model (0.7742), indicating the effectiveness of transformer-based global feature learning. However, the performance of the standalone ViT remains significantly lower than that of the proposed hybrid model.

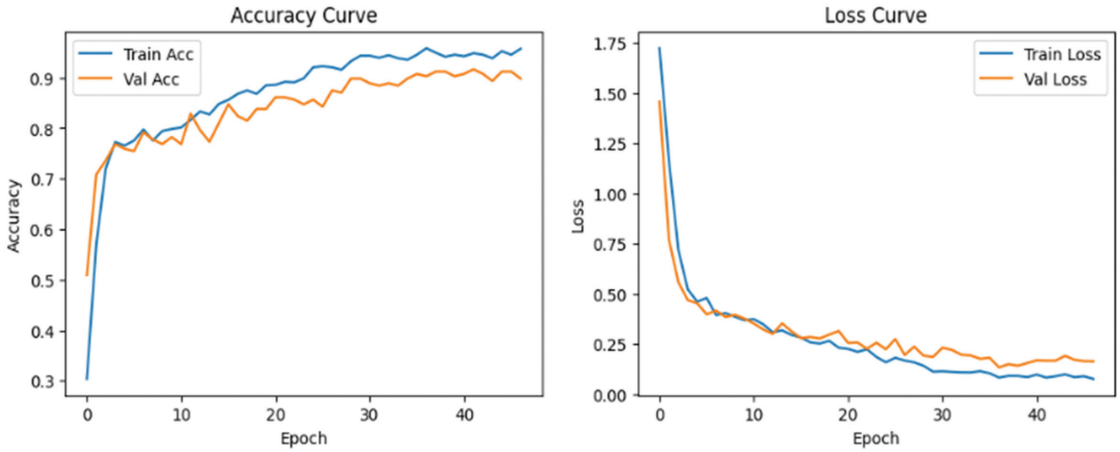


Fig. 4. Training and validation performance curves for the blast disease model

Table 2. Performance evaluation of individual and combined ConvNeXt–ViT models under varying fusion strategies, pre-trained weights, patch sizes, and batch sizes

DL model	Pre-trained weights	Patch size	Batch size	Epochs	Accuracy	Precision	Recall	F1-score
ConvNext	Yes	8	32	40	0.7742	0.7798	0.7742	0.7737
ViT	Yes	8	32	50	0.7972	0.7997	0.7972	0.7962
ConvNeXtViT_EarlyFusion	Yes	8	32	27	0.8295	0.8258	0.8430	0.8313
ConvNeXtViT_IntermediateFusion	Yes	8	32	42	0.8618	0.8953	0.8230	0.8230
ConvNextViT_LateFusion (Proposed)	**Yes**	**8**	**32**	**50**	**0.9263**	**0.9253**	**0.9263**	**0.9175**
ConvNextViT_LateFusion	No	8	32	40	0.9078	0.9097	0.9078	0.9068
ConvNextViT_LateFusion	No	8	16	25	0.8525	0.8470	0.8495	0.8455
ConvNextViT_LateFusion	No	16	32	48	0.8986	0.8936	0.8854	0.8876

To take advantage of the complementary features of both architectures, several hybrid ConvNeXt–ViT fusion strategies were tested. Early Fusion, which combines features at the initial stage, improved the accuracy to 0.8295, showing the enhanced discriminative capability. Intermediate Fusion, which integrates features at a later stage, achieved an even higher accuracy of 0.8618, indicating more confident predictions of severity levels.

The proposed ConvNeXt–ViT Late Fusion model with pre-trained weights, using a patch size of 8 and a batch size of 32, achieves the highest performance, attaining an accuracy of 0.9263, with corresponding precision, recall, and F1-score of 0.9253, 0.9263, and 0.9175, respectively. In comparison, training the same model without pre-trained weights leads to a drop in accuracy to 0.9032, highlighting the importance of transfer learning in improving generalization. Further analysis shows that the model's performance is influenced by key hyper parameters. For instance, reducing the batch size to 16 lowers the accuracy to 0.8525, while increasing the patch size to 16 results in an accuracy of 0.8986.

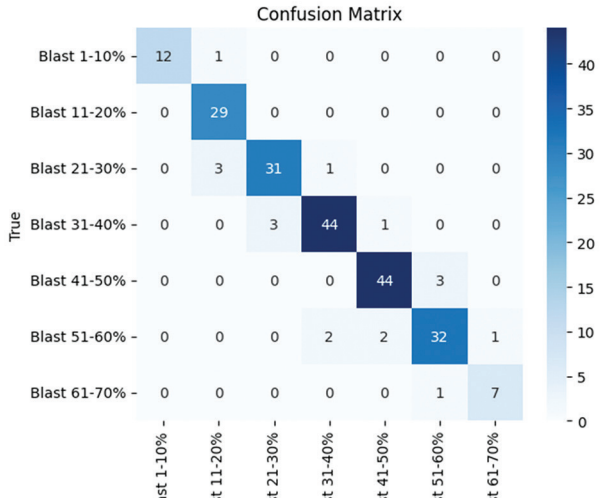


Fig. 5. Confusion matrix for Blast severity classification

Overall, the findings indicate that hybrid models perform better than individual models. And the proposed ConvNeXt–ViT Late Fusion model with pre-trained weights gives the accurate classification of blast disease severity among the fusion strategies. These results showcase the advantages of integrating transformer-based and convolutional representations for the analysis of disease severity.

Furthermore, the significant improvement over the standalone ViT baseline demonstrates that combining convolutional local feature extraction with transformer-based global context modelling leads to accurate classification.

Table 3 summarizes the performance of the proposed ConvNeXt–ViT model in comparison with several classical CNN-based architectures for rice blast disease severity classification. ResNet-34 records the lowest accuracy of 76.50% among the baseline models, stating that severity patterns are not well presented. With accuracies of 85.25% and 86.64%, EfficientNet and MobileNetV2 performed better, demonstrating their capacity to extract more useful features. VGG16 further improved the classification accuracy to 87.10%, while Xception achieved the highest accuracy among the classical CNN models at 88.94%. Comparatively, the suggested ConvNeXt–ViT model outperformed all traditional CNN-based techniques with 0.9263 accuracy, 0.9253 precision, 0.9263 recall, and an F1-score of 0.9175.

Table 3. Performance Comparison with Classical CNN Methods

DL model	Accuracy	Precision	Recall	F1-score
ResNet-34	0.7650	0.7661	0.7650	0.7646
EfficientNet	0.8525	0.8571	0.8525	0.8512
MobileNetV2	0.8664	0.8752	0.8664	0.8649
VGG16	0.8710	0.8752	0.8710	0.8705
Xception	0.8894	0.8917	0.8894	0.8885
Proposed ConvNext-ViT	0.9263	0.9253	0.9263	0.9175

5. CONCLUSIONS AND FUTURE WORK

A hybrid ConvNeXt–ViT deep learning model for rice blast disease severity classification was introduced in this study. A severity-level annotated dataset comprising seven classes was utilized to address the limitations of existing severity categorization approaches. Experimental results demonstrate that the Late Fusion strategy consistently outperforms both Early and Intermediate Fusion approaches, confirming that independent feature learning before integration yields more discriminative representations. The incorporation of pre-trained weights proved instrumental in enhancing model generalization, particularly under the constraint of limited annotated agricultural data. The proposed ConvNeXt–ViT Late Fusion model achieved a classification accuracy of 92.63%, with a precision of 0.9253, recall of 0.9263, and F1-score of 0.9175, outperforming all evaluated conventional deep learning architectures. Overall, the proposed framework establishes a reliable and accurate solution for automated rice blast severity classification. The integration of convolutional and transformer-based learning paradigms within a unified architecture presents a broadly applicable strategy for fine-grained agricultural disease classification.

Future enhancements of this work will aim to support multi-disease severity detection by including other major rice diseases in the proposed model. Expanding the dataset with field-acquired images under diverse environmental conditions will further improve model robustness and real-world applicability. In addition, future work will focus on validating the model on external datasets, as they become available, to further assess its generalization capability across different data sources.

ACKNOWLEDGMENTS

The authors thank Dr Murali Mohan and Dr K.S. Somashekar from the College of Agriculture, University of Agricultural Sciences, GKVK, Bangalore, for their invaluable guidance, insights, and support throughout this research. Dr Murali Mohan also assisted in categorizing the rice blast disease images into seven distinct severity levels.

DATA AVAILABILITY

Datasets analysed in this study are publicly available at Mendeley Data, https://data.mendeley.com/drafts/r845f89rsy

DECLARATION OF THE USE OF AI TOOLS

During the preparation of this work, the authors used ChatGPT and Quillbot in order to prepare the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

[1] T. Agarwal, S. P. Sreelakshmi, "A VECM analysis on the impact of agricultural exports, financial openness and trade openness on India's ecological footprint", International Journal of Innovative Science and Research Technology, Vol. 9, No. 3, 2024, pp. 1747-1777. doi: 10.38124/ijisrt/IJISRT24MAR1081

[2] D. Mondal, S. R. Akhtar, R. Mukhopadhyay, "A review on important diseases of rice in West Bengal", International Journal of Research in Agronomy, Vol. 7, No. 4S, 2024, pp. 67-70. doi: 10.33545/2618060X.2024.v7.i4Sa.533

[3] S. Sindhuja, D. P. Malik, "Growth and instability analysis of rice production in India: A zone wise study", Archives of Current Research International, Vol. 25, No. 9, 2025, pp. 94-109. doi: 10.9734/acri/2025/v25i91479

[4] Q. Wang et al. "Key transcription factors required for outburst of rice blast disease Magnaporthe oryza", Phytopathology Research, Vol. 6, No. 1, 2024, pp. 5, doi: 10.1186/s42483-024-00225-0

[5] N. Mandal et al. "Assessing rice blast disease severity through hyperspectral remote sensing", Journal of Agrometeorology, Vol. 24, No. 3, 2022, pp. 241-248. doi: 10.54386/jam.v24i3.1587

[6] A. E. Asibi, Q. Chai, J. A. Coulter, "Rice blast: A disease with implications for global food security", Agronomy, Vol. 9, No. 8, 2019, p. 451. doi: 10.3390/agronomy9080451

[7] Chaudhary and U. Kumar, "Automated detection and classification of rice crop diseases using advanced image processing and machine learning techniques", Traitement du Signal, Vol. 41, No. 2, 2024, pp. 739-752. doi: 10.18280/ts.410216

[8] R. Li et al. "Predicting rice diseases using advanced technologies at different scales: Present status and future perspectives", aBIOTECH, Vol. 4, No. 4, 2023, pp. 359-371. doi: 10.1007/s42994-023-00126-4

[9] S. R. Shah, S. Qadri, H. Bibi, S. M. W. Shah, M. I. Sharif, F. Marinello, "Comparing Inception V3, VGG16, VGG19, CNN, and ResNet50: A case study on early detection of a rice disease", Agronomy, Vol. 13, No. 6, 2023, p. 1633. doi: 10.3390/agronomy13061633

[10] S. Vijayan, C. L. Chowdhary, "Hybrid feature optimized CNN for rice crop disease prediction", Scientific Reports, Vol. 15, No. 1, 2025, p. 7904. doi: 10.1038/s41598-025-92646-w

[11] X. Bi, H. Wang, "Double-branch deep convolutional neural network-based rice leaf disease recognition and classification", Journal of Agricultural Engineering, Vol. 55, No. 1, 2024. doi: 10.4081/jae.2023.1544

[12] S. Lamba, V. Kukreja, A. Baliyan, S. Rani, S. H. Ahmed, "A novel hybrid severity prediction model for blast paddy disease using machine learning", Sustainability, Vol. 15, No. 2, 2023, p. 1502. doi: 10.3390/su15021502

[13] P. I. Ritharson, K. Raimond, X. A. Mary, J. E. Robert, J. Andrew, "DeepRice: A deep learning and deep feature based classification of rice leaf disease subtypes", Artificial Intelligence in Agriculture, Vol. 11, 2024, pp. 34-49. doi: 10.1016/j.aiia.2023.11.001

[14] P. S. Roy, V. Kukreja, "Vision transformers for rice leaf disease detection and severity estimation: A precision agriculture approach", Journal of the Saudi Society of Agricultural Sciences, Vol. 24, No. 3, 2025, p. 3. doi: 10.1007/s44447-025-00007-w

[15] P. S. Roy, V. Kukreja, "Integrating Visual and Environmental Modalities for Rice Leaf Diseases Classification and its Severity Prediction with Dual-Output CNN", Journal of Transformative Technologies and Sustainable Development, Vol. 10, No. 1, 2026, p. 1. doi: 10.1007/s41314-025-00089-y

[16] P. K. Sethy, N. K. Barpanda, A. K. Rath, S. K. Behera, "Deep feature based rice leaf disease identification using support vector machine", Computers and

Electronics in Agriculture, Vol. 175, 2020, p. 105527. doi: 10.1016/j.compag.2020.105527

[17] M. K. Shwetha, V. Ravi Kumar, "Rice blast disease severity levels image dataset", Mendeley Data, Vol. 1, 2025. doi: 10.17632/r845f89rsy.1

[18] X. Huang, D. Xu, Y. Chen, Q. Zhang, P. Feng, Y. Ma, Q. Dong, F. Yu, "A strongly generalized apple leaf disease classification model based on the fusion of ConvNeXt and transformer", Information Processing in Agriculture, Vol. 12, No. 4, 2025, pp. 466-477. doi: 10.1016/j.inpa.2025.03.001

[19] X. Zhang, H. Dong, J. Yang, Z. Lu, L. Gong, L. Zhang, "A multi-disease detection method for paddy rice based on enhancing detection transformer with ConvNeXt-DAM-FFNet refinement", Journal of Phytopathology, Vol. 173, No. 4, 2025, p. e70106. doi: 10.1111/jph.70106

[20] P. Yashaswini, S. Bhavani, N. Yamuna, P. A. Kumar, M. A. Kiran, "Paddy leaf disease detection using fine-tuned ConvNeXt-Tiny model", Proceedings of the 5th International Conference on Intelligent Technologies, Hubballi, India, 20-22 June 2025, pp. 1-7. doi: 10.1109/CONIT65521.2025.11167785

[21] S. Tyagi, S. Reddy, R. Anand, A. Sabharwal, "Enhancing rice crop health: A lightweight CNN-based disease detection system with mobile application integration". Multimedia Tools and Applications, Vol. 83, 2024, pp. 48799-48829. doi: 10.1007/s11042-023-17449-5

[22] Y. Chen, N. Zhang, A. Wang, Z. Liu, J. Yue, "Automatic classification of tomato leaf diseases based on Mt-ConvNeXt model", Proceedings of the 8th International Conference on Electrical, Mechanical and Computer Engineering, Xi'an, China, 2024, pp. 2021-2025. doi: 10.1109/ICEMCE64157.2024.10862217

[23] J. Yan, Y. Mo, Y. Yu, S. Dou, R. Yang, "Citrus disease classification model based on improved ConvNeXt", IEEE Access, Vol. 2024, pp. 152498-152510. doi: 10.1109/ACCESS.2024.3417310

[24] E. Ozbilge, E. Ozbilge, "Robust plant disease recognition using a neutrosophic-enhanced, RBF-based stacked ensemble of ConvNeXt and classical CNN model", International Journal of Neutrosophic Science, Vol. 26, No. 2, 2025, pp. 132-152. doi: 10.54216/IJNS.260210

[25] M. De Silva, D. Brown, "Multispectral plant disease detection with vision transformer-convolutional neural network hybrid approaches", Sensors, Vol. 23, No. 20, 2023, p. 8531. doi: 10.3390/s23208531

[26] M. A. Haque et al. "An enhanced vision transformer network for efficient and accurate crop disease detection", Expert Systems with Applications, Vol. 21, 2025, p. 127743. doi: 10.1016/j.eswa.2025.127743

[27] P. Gole, P. Bedi, S. Marwaha, M. A. Haque, C. K. Deb, "A lightweight vision transformer network for identification of plant diseases", Frontiers in Plant Science, Vol. 14, 2023. doi: 10.3389/fpls.2023.1221557

[28] S. Kusuma, K. R. Jothi, "Early betel leaf disease detection using vision transformer and deep learning algorithms", International Journal of Information Technology, Vol. 16, No. 1, 2023, pp. 169-180. doi: 10.1007/s41870-023-01647-3

[29] J. N. Saeed, M. K. Hussein, "A multi-ViTs-based approach for automatic rice leaf disease classification", Iraqi Journal of Science, Vol. 66, No. 9, 2025, pp. 3938-3950. doi: 10.24996/ijs.2025.66.9.33

[30] N. S. Kollati et al. "Tomato plant leaf disease detection using multiple hybrid vision transformers", Proceedings of the 6th International Conference on Inventive Research in Computing Applications, Coimbatore, India, 25-27 June 2025, pp. 2000-2004. doi: 10.1109/ICIRCA65293.2025.11089791

[31] F. S. Alamri, T. Sadad, A. S. Almasoud, R. A. Aurangzeb, A. Khan, "Mango disease detection using fused vision transformer with ConvNeXt architecture", Computers, Materials & Continua, Vol. 83, No. 1, 2025, pp. 1023-1039. doi: 10.32604/cmc.2025.061890

[32] H. Ghosh et al. "Advanced neural network architectures for tomato leaf disease diagnosis in precision agriculture", Discover Sustainability, Vol. 6, No. 1, 2025, p. 312. doi: 10.1007/s43621-025-01149-1

[33] A. Malik et al. "Design and evaluation of a hybrid technique for detecting sunflower leaf disease using deep learning approach", Journal of Food Quality, Vol. 2022, No. 1, 2022, p. 9211700. doi: 10.1155/2022/9211700