CATNet: Attention-enhanced Robust CNN Architecture with Transformer for Rice Crop Disease Detection

Sanam Salman Kazi [1,*], Bhakti Palkar [1], Vishwayogita Savalkar [2], and Supriya Khaitan [3]

[1] Department of Computer Engineering, K. J. Somaiya School of Engineering, Somaiya Vidyavihar University, Mumbai, India

[2] Department of Computer Science and Engineering (Data Science), Anjuman-I-Islam's Kalsekar Technical Campus School of Engineering and Technology, Navi-Mumbai, India

[3] Department of Engineering and Technology, Bharati Vidyapeeth Deemed to be University, Navi-Mumbai, India

Email: sanam.kazi19@gmail.com (S.S.K.); bhaktiraul@somaiya.edu (B.P.); vishwayogita.ks@gmail.com (V.S.); supriyakhaitan21@gmail.com (S.K.)

*Corresponding author

***Abstract*—To ensure an adequate supply of food while preventing loss of yield, timely identification of rice leaf diseases is vital. Traditional deep learning models often exhibit poor discrimination among visually challenging diseases, low generalization for images with natural background, and fail to handle datasets with class imbalance. Convolutional Neural Networks (CNN)-based models perform poorly in a variety of field settings due to their significant emphasis on local feature extraction and difficulty capturing long-range spatial patterns. Attention mechanisms that are vital for emphasizing fine-grained lesion points are often ignored, which lowers localization accuracy and makes them more susceptible to background noise. This paper proposes a dual-branch attention-enhanced robust CNN model with transformer architecture named as CATNet to address these problems by combining transformer-based global context modelling with CNN-driven local texture learning. To enhance lesion-focused feature refinement and suppress unimportant background information, the model uses an attention fusion technique. On public as well as self-generated comprehensive real-world datasets, CATNet achieves 98% test accuracy with 99% category-wise recognition, effectively handling class imbalance and complex backgrounds where many state-of-the-art models fall short. The key design principle is to maintain it lightweight with only 3.9 M parameters and utilize the least amount of memory of 2287 MB as compared to state-of-the-art models, thereby rendering it ideal for real-world, resource-constrained precision farming. CATNet beats popular models like EfficientNetB0, DenseNet169, and MobileNetV3Small in terms of tradeoff between accuracy as well as efficiency, reaching 98% test accuracy with just 332 MFLOPs, thereby rendering it as a lightweight, efficient, and robust model for detecting rice diseases in real time systems.**

***Keywords*—deep learning, rice disease detection, transformer, depth-wise convolution, convolutional neural network, attention mechanism**

Manuscript received August 31, 2025; revised October 9, 2025; accepted March 10, 2026; published September 10, 2026.

I. Introduction

The nation's economic growth is heavily influenced by the agriculture sector. The Gross Domestic Product (GDP) of the agricultural sector has decreased dramatically in recent years, even though farming is the bedrock of a nation's economic development. Many diseases that impact crop production are at fault for this [1]. Farmers are short of the technical expertise sufficient for coping with diseases that disrupt rice crops [2]. Production of rice crops can be severely affected by a number of disease outbreaks. Among the most prevalent ailments in the community are leaf blast [3], brown spot [4], stem rot [5], leaf smut [6], and bacterial leaf blight [7]. To detect numerous types of diseases, a computer-generated disease detection technique is required. There are several phases in the automated disease detection system. The first step is acquiring information is gathering pictures from various sources. Post data collection and pre-processing are carried out in order to guarantee that the information obtained is of exceptional quality. The pre-processing phase incorporates data augmentation, segmentation, image enhancement and noise reduction.

Machine Learning (ML) is a fairly straightforward strategy, but the degree to which it serves relies extensively on the strength of the feature recognition strategies [8]. Because deep neural network models acquire information from data itself, they do not require manual extraction of characteristics, which has given rise to their widespread use in image classification [9].

Using the different classification frameworks, such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Naive Bayes, Pallathadka *et al.* [10] developed an ML-driven technique for automated paddy leaf disease recognition. The initial processing was performed employing histogram normalization, and feature extraction was done using Principal Component

Analysis (PCA). An enhanced CNN model has been designed for identifying paddy diseases [11]. The paper's deployed model endured the extraction and classification of features through a VGG-19 architecture to promote learning and boosts model generalization and lowers down on training time. It was observed that it yet suffers from domain mismatch as well as demanding more resources for tuning. Preprocessing accelerates training and promotes generalization by enriching data, but it can also increase computing complexity. Both are vital, but face a trade-off between computational efficiency and generalization. However, the vast majority of existing techniques largely depend on huge datasets and transfer learning, which could hinder accessibility and efficacy for lightweight, real-time deployment.

Recent studies report CNN models perform better with RGB images [12] and are mostly trained on clean laboratory or Kaggle-style datasets, which leads to models that have issues generalizing under real-field variability [13]. The majority of research still uses traditional convolutional networks for agricultural disease detection, despite Transformers' potential to capture global contextual information in computer vision [14]. Additionally, attention mechanisms which are essential for identifying discriminative lesion regions are frequently understudied or considered optional, which hinders the model's capacity to manage the background clutter, occlusion, and lighting variations that are frequently found in field settings [15]. Recent hybrid CNN-Transformer techniques are designed to integrate the local spatial reasoning with local texture extraction to improve generalizability, and improve interpretability and robustness, but still, it is observed that these architectures are underreported in rice disease detection [16].

Current advances in plant disease classification have explored hybrid designs that combine transformers with group-wise convolutions with the goal to optimize the use of both fast local feature extraction and global contextual attention. Research integrating transformer-based attention mechanisms such as Convolutional Block Attention Module (CBAM) with group-wise convolutions has proven that hybrid architectures could enhance feature extraction at a reduced computational cost than conventional deep networks. Influenced by these findings, this study offers a unique dual-path design for rice disease classification: transformer blocks with CBAM attention are employed in one path to capture global dependencies, while group-wise convolutions are utilized in the other to efficiently gather local characteristics. The approach attempts to develop a lightweight and efficient computational model suitable for actual-world agricultural applications by delivering high classification performance without transfer learning.

The proposed CATNet model results in an immense impact in the field of intelligent agricultural disease detection through integrating transformer-based reasoning with convolutions, which rely on attention, excelling conventional CNNs and hybrid models with respect to classification performance and generalization ability for all classes of rice disease, with a special focus on under-explored rice crop disease sheath rot.

The following are the author's key contributions:

(1) Improved feature recognition by designing the novel, lightweight, dual-path CATNet model, which boosts spatial and contextual knowledge and efficiently discriminates against visually identical rice leaf diseases having intra-class variations and inter-class similarity.
(2) The synergistic interaction between the CBAM and transformer modules improves the generalization capability of the proposed CATNet to accurately identify low-frequency and fine-grained disease patterns.
(3) Benchmark evaluation using a publicly available dataset demonstrates high generalization and consistent performance across public data sources, thereby achieving 97% accuracy.
(4) The proposed CATNet model outperformed EfficientNetB0 as well as other prominent lightweight architectures such as DenseNet169 and MobileNetV3Small by exhibiting its powerful computational performance, achieving 98% test accuracy with only 332 M FLOPs, making it a lightweight model suitable for real-time and mobile deployment.
(5) This study confirms the robustness of the proposed model by means of a five-fold cross-validation performed on a self-generated field dataset that includes various climatic conditions. The model achieved a mean accuracy of 95.24%, confirming the model's reliability in the face of real-world variations. The evaluation across various data sets uncovered an accuracy of 97%, underscoring the unmatched generalization capability of the proposed architecture above the training space.

The organization of the paper is as follows: Section II identifies the techniques and methods for rice crop disease detection. The architecture of the proposed CATNet framework is explained in Section III. The results of training the proposed CATNet a using self-generated dataset and a publicly available dataset are discussed, and cross validation results are represented proving the effectiveness of proposed model for rice disease detection in Section IV. Section V highlights the outcomes and achievements of this study, along with future plans for research.

II. Literature Review

The recent research on rice disease detection has explored a range of deep learning and machine learning algorithms leveraging various data sets and models. Despite the model's considerable computational complexity, Kaur *et al.* [17] scored 93.3% accuracy, an Area Under the Curve (AUC) of 0.989, and an F1-Score of 92.8%, deploying an ensemble of SqueezeNet, Visual Geometry Group (VGG)-16, and Inception-V3 on 8572 images from Mendeley and Kaggle.

Singh *et al.* [18] obtained 98.7% accuracy for rice blast detection by applying a variety of CNN architectures, notably VGG-16, Inception-v1, LeNet, AlexNet, and Xception, on 6300 real-time images. An ABi-LSTM model that was trained on Kaggle data achieved 98.86% accuracy and 94.24% F1-Score was put forward by Dubey *et al*. [19], albeit it was merely useful for three diseases. Although an advanced transfer learning combination was not employed, Haridasan *et al*. [20] obtained 91.45% accuracy through the combination of CNN and SVM on 2,000 real-time images. Using 2000 Kaggle photos, Liu *et al*. [21] developed a CNN having five layers' combinations of convolution and pooling that reached 98.64% accuracy, great precision, and recall; however, the dataset size and inadequate preprocessing were limitations. Though generalizability and overfitting were issues, Aggarwal *et al*. [22] applied EfficientNet-B3 and EfficientNet-V2B3 segmentation algorithms on 551 images, achieving 94% accuracy with good overall metrics. Nalini *et al.* [23] proposed a DNN-CSA model that was successfully trained on 121 real-time images with 96.96% accuracy, but it suffered from overfitting and computational challenges. Deng *et al*. [24] employed an ensemble of SE-ResNet-50, DenseNet-121, and ResNeSt-50 on an enormous 33,026-image real-time dataset. They scored 91% accuracy with impressive recall and F1-Score, but recognition speed was adversely affected by model complexity. Lu *et al*. [25] detected only sheath blight using a Backpropagation Neural Network on 230 real-time images and attained 85.8% accuracy. Likewise, Bhattacharya *et al.* [26] attained 94% accuracy using a standard CNN with two convolutional and two max-pooling layers on 2000 rice field images, with the shallow architecture.

A deep multi-task Transfer Learning (TL) model for detecting rice leaf diseases, which include leaf blast, brown spot, and bacterial blight [27]. Image enhancement techniques were utilized to boost the quantity of a small dataset of 40 photos from the University of California, Irvine (UCI) Machine Learning Repository. Using a multi-task TL methodology, the study notably improved the VGG-16 architecture and examined how it performed against ResNet-50 and DenseNet-121. The improved accuracy of the upgraded VGG-16 demonstrates effective multi-task learning and feature representation when working with limited data to identify and classify eleven different kinds of rice leaf diseases, including brown spot and leaf blast. Mandwariya and Jotwani [28] deployed a pre-trained Deep CNN applying transfer learning and optimization methods such as SGDM, ADAM, and RMSprop. The XceptionNet model achieved the greatest accuracy of 93.3 percent on a number of investigated architectures, including ResNet50, DenseNet, VGG19, SqueezeNet, CNN, and XceptionNet. It proved the efficiency of pre-trained DCNNs with improvement methods for exact disease detection. However, the study is limited for application in a resource-constraint environment by its high computational complexity and dependence on huge, high-quality datasets.

Saddami *et al*. [29] applied three mobile-compatible CNN architectures for classifying rice leaf disease: ShuffleNet, MobileNetV2, and EfficientNet-B0. Since the models are efficient and need less processing and memory capacity, these models are suitable for mobile devices. Each of the fully linked layers of each model has been separated by a dropout layer for improved performance. Early stopping served to avert overfitting during the training phase. The models' performance had been assessed based on how effectively they diagnosed rice leaf diseases; EfficientNet-B0 had the highest accuracy, at 99.8%. This study fails to compare the latest lightweight models as well as hybrid and ensemble models.

Xu and Yang [30] extracted local and global data by proposing a hybrid model called multi scale convolutional transformer network for plant disease image classification, which merged convolutional modules with a visual transformer backbone. The model attained an accuracy of 99.95% on the Plant Village dataset. Despite the model's high accuracy, the multi-scale convolution and transformer components make it a little complex. Since clean laboratory images are used, the cross-validation was mandatory, but found to be missing.

Convolution network enlightened transformer with regional crop disease classification, utilized transformer blocks and group-shuffle depth-wise blocks, another kind of group convolution, for combining both regional and global representation for tomato-leaf disease detection [31]. Reliability in an actual agricultural environment is still doubtful because the impacts of different illumination, occlusions, image quality, and field conditions are not thoroughly evaluated. These approaches indicate that incorporating both local patterns and long-range dependencies, such as hybrid models, may significantly boost accuracy and feature representation. But despite their outstanding results, these models' practical applicability in real-world agricultural settings has been restricted by their extensive computational demands, dependency on huge and high-quality datasets, the possibility of overfitting, and challenges with real-time deployment on low-resource devices. Furthermore, a great deal of studies focuses on certain types of typical diseases, finding it hard to generalize to unusual plant diseases such as sheath rot.

Recent advancements in rice disease detection have primarily used convolutional neural networks and transfer learning models, attaining classification accuracies between 85% and 99% for significant rice crop diseases. Regardless of the robust performance of deep CNN-based and ensemble frameworks on benchmark datasets, their efficacy is frequently constrained in actual-life situations due to notable intra-class variability and considerable inter-class similarity because of fluctuations in disease severity, illumination, leaf orientation, and background complexity. Conventional CNNs, due to their narrow receptive fields, express limited ability to simulate long-range contextual dependencies and complex distinctive cues, resulting in effectiveness drops in cluttered settings. The inclusion of attention mechanisms within transformer frameworks offers a powerful strategy by offering adaptive feature weighting and global context modelling, thus enhancing the network's potential toward focusing on disease-relevant regions while suppressing background noise.

Recent studies have gradually leveraged transformer-based topologies and attention mechanisms to address the shortcomings of CNN-based models for rice disease detection. As opposed to traditional CNNs that depend on localized feature extraction, transformers provide global context for modeling via self-attention, boosting the distinctiveness of visually identical disease patterns. The rice transformer model exhibited this ability by merging image and sensor data, reaching an accuracy of 97.38% across 4200 samples [32]. An enhanced BEiT-based system employed attention mapping for better feature interpretability, achieving a precision of 0.92 and an F1-Score of 0.91 [33]. Comparative studies have shown that vision transformers surpass CNNs in paddy disease classification by effectively addressing intra-class variation and inter-class similarity [34]. In addition, deformable attention algorithms have been used to capture multi-scale disease characteristics among intricate backgrounds, recording an F1-Score of 94.3% while enhancing computational efficiency by integrating the Residual module with transformers in the study [35]. Regardless of these gains, the majority of transformer-based methodologies continue to be computationally intensive and reliant on extensive annotated datasets, underscoring the want for improved attention-transformer frameworks that can attain robust generalization in real-world scenarios. Notwithstanding these benefits, the practical use of unified attention-transformer frameworks for rice disease classification is still constrained, especially in study focusing on real-field variability and class ambiguity. This gap drives the current study to create an attention-guided transformer architecture designed to enhance generalization and provide dependable performance in real-world agricultural settings.

III. Materials and Methods

A. Dataset

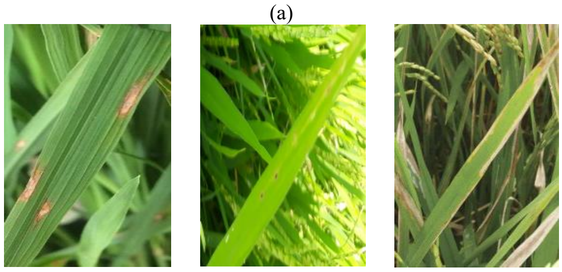
The comprehensive dataset with field and online images is created for robust training and improved generalization. Table I represents the collection of 9396 total images from real paddy fields as well as from online public rice disease datasets.

The images were captured from research fields in Maharashtra and labelled with the help of a plant pathologist at the agricultural university at Dapoli using a Redmi 12 pro mobile camera. The different time intervals in the morning, around 7.00 am to 9.00 am, and the evening, from 4.00 pm to 5.00 pm, were selected to capture diversity in image condition. Fig. 1 shows the difference in quality of images where the field collected images are present natural field setting, whereas online images have a synthetic background. All images are resized to 100 × 100. The rice crop disease images from the internet are combined to increase the dataset size as well as the variability. The field collected images are high in quality and in real condition, with natural lighting and a cluttered background. The training and testing ratio is 80:20. Cross-validation with a 5-fold is further implemented to prove the reliability of the proposed CATNet architecture.

TABLE I. Comprehensive Rice Disease Dataset

Diseases	Image count	Field Collected Images	Online images
Sheath rot	1224	1224	0
Rice Blast	1827	932	895
Bacterial Blight	2240	1463	777
Tungro	1308	NIL	1308
Brown Spots	1301	NIL	1301
Healthy	1496	350	1146

(a)



(b)

Fig. 1. Quality comparison between: (a) Field collected, (b) Online public images.

B. Proposed CATNet Architecture

The proposed CATNet architecture in Fig. 2 consists of two dual CNNs with concurrent feature extraction branches that are combined for classification. Branch-1 efficiently pulls local spatial information employing lightweight depth wise and pointwise convolutions whereas Branch-2 integrates standard convolutions with a Convolutional Block Attention Module (CBAM) and a transformer block to effectively capture prominent features and complex relationships. For the model to maximize the benefit of complementary information, features from both branches are concatenated. A convolutional layer further enhances the combined representation, and max pooling is utilized to cut down spatial dimensions. To be able to arrive at a softmax layer for multi-class prediction. The generated feature maps are flattened and further processed through two completely linked layers with dropout for regularization. By incorporating attention-based enhancement, global contextual modelling, and local feature extraction into a single framework, it increases efficiency and accuracy.

C. *Group Convolutional Block*

To successfully retrieve local features while preserving a low computational complexity, branch-1 incorporates depth wise and pointwise convolutions as shown in Fig. 2. The 1 × 1 pointwise convolutions merge data from multiple channels together, while depth-wise convolutions treat each channel independently, capturing spatial patterns. Training is reliable due to a progressive decrease in spatial dimensions. Fundamental local traits in the input stream are captured by this branch's feature representation, which is both lightweight and efficient.

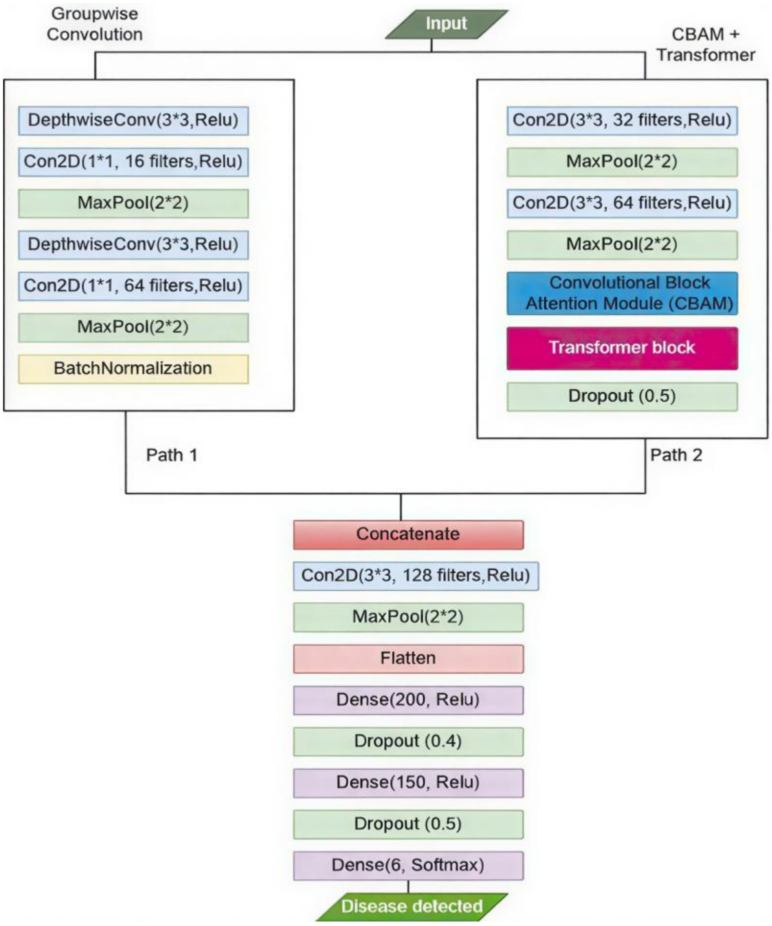


Fig. 2. Proposed CATNet architecture with transformer and attention mechanism.

Depth wise convolution convolves each input channel independently with its kernel $K^{(d)}$, given by Y_d output feature as in Eq. (1).

$$Y_d(i,j) = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} X_d(i+m, j+n) K^{(d)}(m,n) \qquad (1)$$

where X_d is the input feature for channel c and (i, j) is the spatial position.

Point-wise convolutions combine depth wise outputs to generate new feature maps in Eq. (2).

$$y^{c}_{i,j}, P'_0 = \sum_{c=1}^{cin} K^{(P)}_{1,1,c,0}\, y^{(d)}_{i,j,c} \qquad (2)$$

D. *Convolutional Block Attention Module Block*

The CBAM applies channel and spatial attention processes in a sequential manner to improve feature representations, as shown in Fig. 3. In the channel attention stage, channel-wise attention weights are generated by utilizing global average and max pooling across the spatial dimensions, followed by shared dense layers and a sigmoid activation. These weights are then multiplied by the input feature maps to highlight informative channels. A spatial attention map is created in the spatial attention stage by calculating average and max pooling along the channel axis, concatenating them, and then passing them through a 7 × 7 convolution with sigmoid activation. The channel-refined features are then multiplied by the spatial attention map. The discriminative power of the learned features is increased by the network's ability to selectively focus on both significant channels and spatial regions thanks to this dual attention mechanism. Given input feature map F, where F belongs to $R^{H\times W\times C}$. Two types of channel descriptors are extracted using Global average pooling and max average pooling as given by Eqs. (3) and (4). Later, they both pass through a shared Multilayer Perceptron (MLP) with a reduction ratio r, sigmoid, and ReLU activation indicated in Eq. (5).

$$f_{avg} = GAP(F) \tag{3}$$

$$f_{\max} = GMP(F) \tag{4}$$

$$M_c = \sigma\left(MLP\left(f_{avg}\right) + MLP\left(f_{\max}\right)\right) \tag{5}$$

The input is scaled channel-wise as in Eq. (6) and then spatially as in Eq. (7).

$$F' = M_c \otimes F \tag{6}$$

$$F'' = M_s \otimes F' \tag{7}$$

E. Transformer Block

In order to add global contextual information and long-range dependencies within the feature maps, the Transformer block is employed as depicted in Fig. 4. The input first undergoes transformation into a sequence that may be used for attention operations. The model can acquire associations across all spatial positions by means of the inclusion of a multi-head self-attention layer. Regularization of the dropout and residual layers then serves to stabilize training. The attentive features are further improved using a spatial feed-forward network with two dense layers and dropout. The output is then adjusted to get back to its initial spatial dimensions. Incorporating this block into the network boosts the local feature acquisition of convolutional layers by modelling complicated global relationships between features.

Layer (type)	Output Shape	Param #	Connected to
input_layer_1 (InputLayer)	(None, 100, 100, 3)	0	-
conv2d_7 (Conv2D)	(None, 100, 100, 32)	896	Input_layer_1[0]
max_pooling2d_6 (MaxPooling2D)	(None, 50, 50, 32)	0	conv2d_7[0][0]
conv2d_8 (Conv2D)	(None, 50, 50, 64)	19,496	max_pooling2d_6[...
max_pooling2d_7 (MaxPooling2D)	(None, 25, 25, 64)	0	conv2d_8[0][0]
global_average_poo... (GlobalAveragePool...	(None, 64)	0	max_pooling2d_7[...
global_max_pooling... (GlobalMaxPooling2...	(None, 64)	0	max_pooling2d_7[...
dense_2 (Dense)	(None, 8)	520	global_average_p... [0][0],max_pooli...
dense_3 (Dense)	(None, 64)	576	dense_2[0][0], dense_2[1][0]
add_1 (Add)	(None, 64)	0	dense_2[0][0], dense_2[1][0]
activation_1 (Activation)	(None, 64)	0	add_1[0][0]
reshape_1 (Reshape)	(None, 1, 1, 64)	0	activation_1[0][...
multiply_2 (Multiply)	(None, 25, 25, 64)	0	max_pooling2d_7[... reshape_1[0][0]
lambda_2 (Lambda)	(None, 25, 25, 1)	0	multiply_2[0][0]
lambda_3 (Lambda)	(None, 25, 25, 1)	0	multiply_2[0][0]
concatenate_1 (Concatenate)	(None, 25, 25, 2)	0	lambda_2[0][0], lambda_3[0][0]
conv2d_9 (Conv2D)	(None, 25, 25, 1)	90	concatenate_1[0][...
multiply_1 (Multiply)	(None, 25, 25, 64)	0	multiply_2[0][0], conv2d_9[0][0]
reshape_2 (Reshape)	(None, 625, 64)	0	multiply_3[0][0]

Layer (type)	Output Shape	Param #	Connected to
multi_head_attenti... (MultiHeadAttention)	(None, 625, 64)	33,216	reshape_2[0][0], reshape_2[0][0]
dropout_1 (Dropout)	(None, 625, 64)	0	multi_head_atten...
add_2 (Add)	(None, 625, 64)	0	reshape_2[0][0], dropout_2[0][0]
layer_normalization_ (LayerNormalization)	(None, 625, 64)	128	add_2[0][0]
depthwise_conv2d_2 (DepthwiseConv2D)	(None, 100, 100, 3)	20	input_layer_1[0]...
dense_4 (Dense)	(None, 625, 64)	4,160	layer_normalizat...
conv2d_5 (Conv2D)	(None, 100, 100, 16)	64	depthwise_conv2d...
dense_5 (Dense)	(None, 625, 64)	4,160	dense_4[0][0]
max_pooling2d_4 (MaxPooling2D)	(None, 50, 50, 16)	0	conv2d_5[0][0]
dropout_2 (Dropout)	(None, 625, 64)	0	dense_5[0][0]
depthwise_conv2d_3 (DepthwiseConv2D)	(None, 50, 50, 16)	160	max_pooling2d_4[...
add_3 (Add)	(None, 625, 64)	0	layer_normalizat... dropout_2[0][0]
conv2d_6 (Conv2D)	(None, 50, 50, 64)	1,000	depthwise_conv2d...
layer_normalization... (LayerNormalization)	(None, 625, 64)	128	add_3[0][0]
max_pooling2d_5 (MaxPooling2D)	(None, 25, 25, 64)	0	conv2d_6[0][0]
reshape_3 (Reshape)	(None, 25, 25, 64)	0	layer_normalizat...
batch_normalizatio... (BatchNormalization)	(None, 25, 25, 64)	256	max_pooling2d_5[...
dropout_3 (Dropout)	(None, 25, 25, 64)	0	reshape_3[0][0]
concatenate_2 (Concatenate)	(None, 25, 25, 128)	0	batch_normalizat... dropout_3[0][0]
conv2d_10 (Conv2D)	(None, 25, 25, 128)	147,584	concatenate_2[0]...
max_pooling2d_8 (MaxPooling2D)	(None, 12, 12, 128)	0	conv2d_10[0][0]
flatten (Flatten)	(None, 16012)	0	max_pooling2d_8[...
dense_6 (Dense)	(None, 200)	3,606,600	flatten[0][0]
dropout_4 (Dropout)	(None, 200)	0	dense_6[0][0]
dense_7 (Dense)	(None, 150)	30,150	dropout_4[0][0]
dropout_5 (Dropout)	(None, 150)	0	dense_7[0][0]
dense_8 (Dense)	(None, 6)	906	dropout_5[0][0]

Fig. 3. Proposed CATNet model architecture.

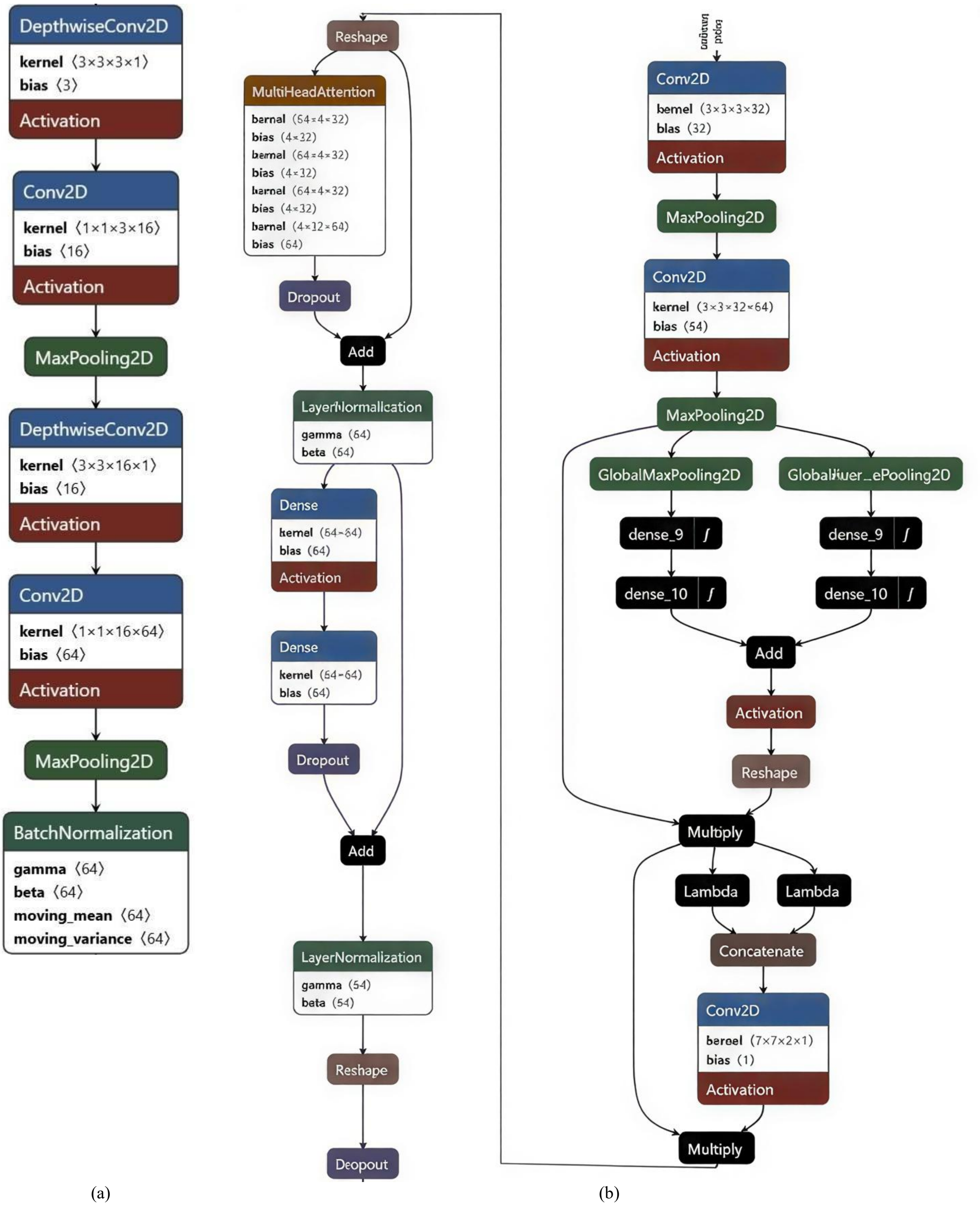


Fig. 4. Parallel path integration: (a) path-1 with Group-wise convolution, (b) Path-2 with CBAM and Transformer with Multi-head attention.

IV. Result and Discussion

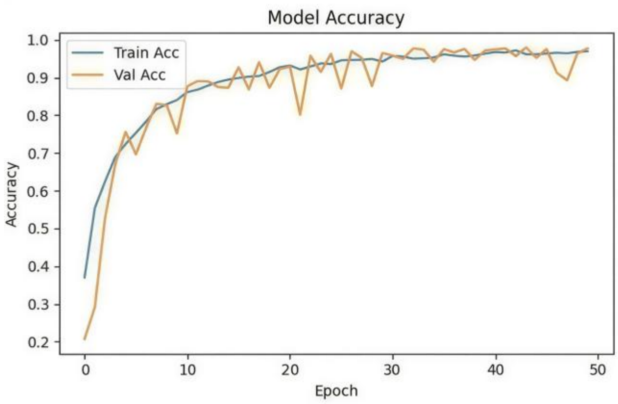
GPU acceleration was applied to improve the efficiency of model training. The tests were conducted employing the CUDA enabled deep learning environment on Kaggle, utilizing the computing capacity of an NVIDIA Tesla P100 GPU. The integration of TensorFlow with Keras allows optimal exploitation of GPU resources, drastically lowering training time and enabling faster optimization of the proposed deep convolutional architecture. The experiments have been executed using a T4 GPU from Kaggle. The internet speed of 100 mbps is used along with a system having 8 G.B. ram size and an Intel Core i7 processor. The code is implemented using TensorFlow, Keras, and Python libraries.

The Adam optimiser was employed to train the proposed VCNet with a starting rate of learning of 0.001. To guarantee steady gradient estimation and computational efficiency, the model was trained for 100 epochs with a mini-batch size of 32. For multiclass classification, categorical cross-entropy loss was used.

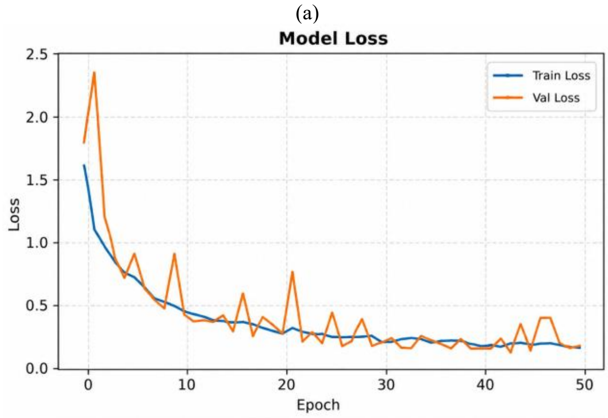
Online data augmentation techniques such as random rotation, horizontal and vertical flipping, zooming, spatial shifting, and brightness change were used during training to enhance generalisation under natural field conditions. Throughout the training stage, validation performance was continually monitored in order to ensure gradual convergence and prevent overfitting.

A. Result Analysis for Proposed CATNet

Validation accuracy and loss curves in Fig. 5 confirm the model's strong generalization capability and consistent convergence. The accuracy curves exhibit rapid gains in the early epochs, which is followed by gradual stabilization until the model achieves peak performance. The model effectively minimizes classification error over time, as indicated by the loss curves, which exhibit a rapid decrease at the initial phase of training and progressively smooth out.



(a)



(b)

Fig. 5. Proposed CATNet trained on self-generated comprehensive dataset: (a) Training and validation accuracy, (b) Training and validation.

Training and validation trajectories for accuracy and loss largely correlate, reflecting fair learning across datasets and verifying an absence of overfitting. In general, these results show that the proposed design and efficient optimization method work well and can be attributed to deliver good results for classifying rice diseases.

A complete visualization of the model's effectiveness in classification of rice diseases can be viewed in the in Fig. 6. The elevated overall accuracy is demonstrated by the strong diagonal dominance, which indicates the majority of samples are correctly categorized into the correct groups. There are very few misclassifications, mostly between visually identical classes like rice blast and bacterial blight, where there may be minor feature pattern overlaps. The model's resilience and discriminative capacity in identifying between various disease symptoms and healthy samples is demonstrated by an absence of notable off-diagonal values for the majority of classes. These results further validate the degree to which the proposed algorithm learns unique visual traits, resulting in precise and reliable disease detection.

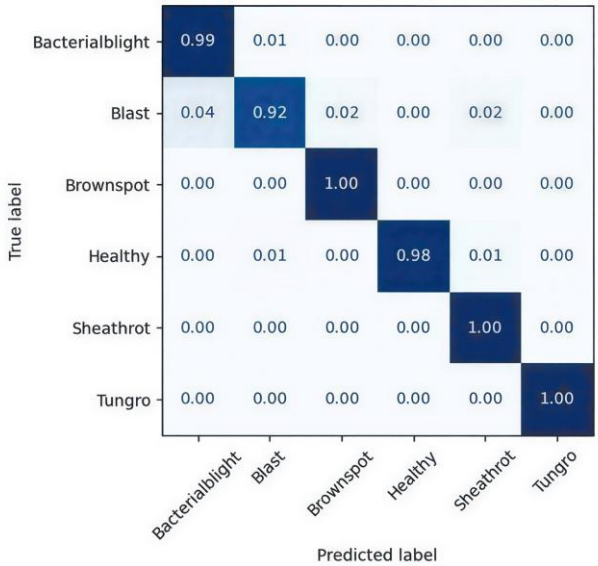


Fig. 6. Confusion matrix for CATNET.

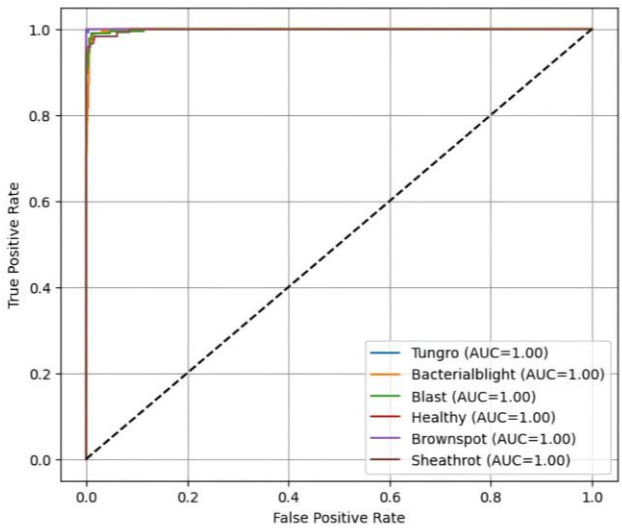


Fig. 7. ROC curve for CATNET.

ROC-AUC curve represents the difference between classes very well and extract features very effectively as shown in Fig. 7. The proposed model possesses an Area

Under the Curve (AUC) of 1.00 for each classification. This suggests that it can tell. The curves are quite near the top-left quadrant, which indicates that the true positive rate and the false positive rate occupy a good equilibrium. The findings prove the robustness of the proposed architecture in addressing inter-class similarity and intra-class variability in realistic agricultural settings.

The proposed CATNet demonstrated remarkable generalization in Table II with training accuracy of 99%, validation accuracy of 98.4%, and testing accuracy of 98%. Effective tuning and no overfitting can be seen by the corresponding training and validation losses of 0.96 and 0.93. Additionally, the model attained balanced precision, recall, and F1-Score values of 98%, proving its consistent performance and resilience in accurately determining rice disease.

The capacity of the model to effectively distinguish between various rice diseases and healthy samples is proven by its consistently high precision, recall, and F1-Scores. Fig. 8 highlights the exceptional prediction efficacy of the proposed deep learning-based algorithm for classifying rice crop diseases across all designated categories. It performed almost flawlessly for brown spot, tungro and healthy classes while accurately recognizing challenging disease types such rice blast and sheath rot. The outcomes confirm the model's extraction and classification of features framework's resilience and validate its capacity for recognizing minute visual variations in rice leaf disease.

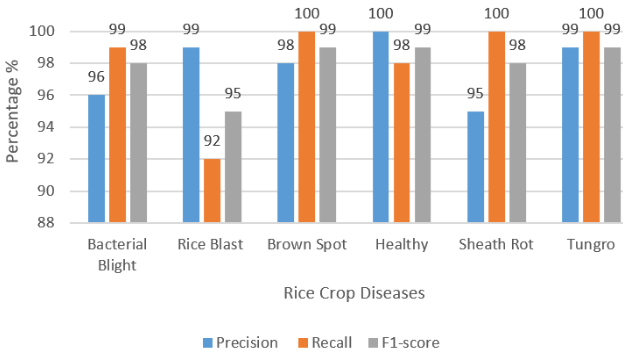


Fig. 8. Class-wise performance of proposed CATNet.

TABLE II. PERFORMANCE ANALYSIS OF PROPOSED CATNET

Training Accuracy	Validation Accuracy	Testing Accuracy	Training Loss	Validation Loss	Precision	Recall	F1-Score
99%	98.4%	98%	0.96	0.93	0.98	0.98	0.98

B. *Evaluation Using Cross-Dataset Testing*

To obtain a fair comparison, consistency and validity of results proposed CATNet is tested on a publicly accessible dataset. It shows that this model's capability isn't biased toward a private or curated dataset. To verify the strength of the proposed framework, we used the Mendeley rice diseases dataset, which is free and open to the public [36]. It has 5932 photos of four different types of rice diseases. Of these, 1185 images have been utilized for testing the model on a public dataset to prove its generalization power. All the images were resized to 100 × 100. The dataset offers a variety of samples from four disease kinds of bice blast, bacterial blight, tungro and brown spot which enables us to evaluate the suggested model's effectiveness and robustness on unseen data. Sheath rot is not available online in public datasets.

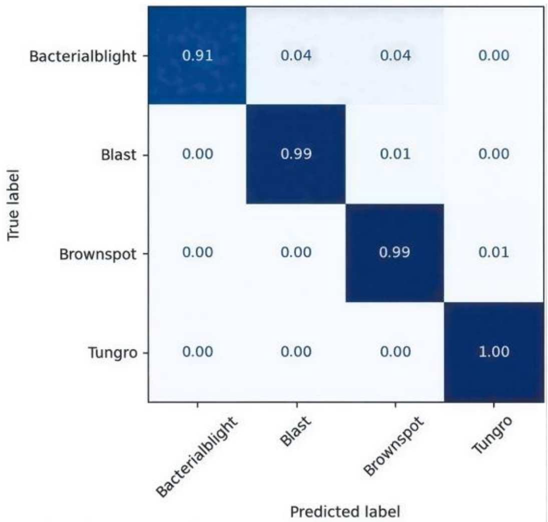


Fig. 9. Confusion matrix on cross dataset testing.

Fig. 10. Test result on public dataset.

The proposed framework exhibited strong proficiency in classification performance of 97% proved by the confusion matrix in Fig. 9. CATNet generalizes well beyond the training data, as evidenced by the significant diagonal dominance in the confusion matrix found on the publicly accessible rice disease dataset. While Bacterial blight displays some ambiguity with visually similar classes, which is expected in real-field situations, Blast, Brown spot, and tungro show high correct classification rates with very few misclassifications. The results presented here verify that, when tested on an accepted public benchmark, CATNet achieves accurate, unbiased, and reproducible performance. Fig. 10 represents the images from the test data predicted correctly by the CATNet model.

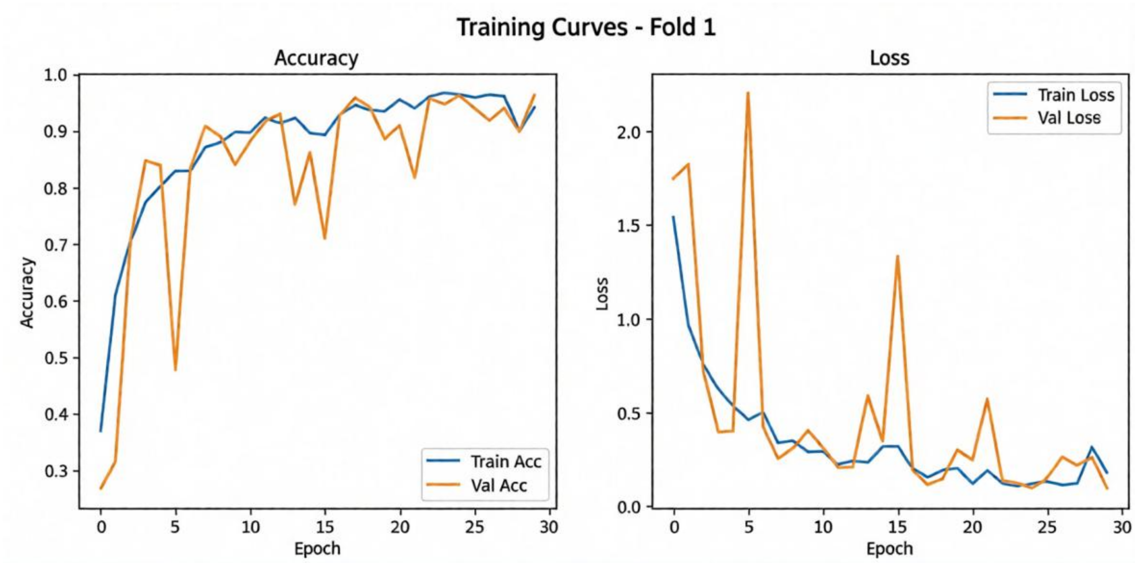
C. *5-Fold Cross-Validation and T-Test*

The stability and generalization abilities of CATNet were evaluated using a stratified 5-fold cross-validation technique shown in Fig. 11. To make sure reproducibility and maintain class distribution over folds, the dataset was randomly shuffled using a fixed seed. Each fold was trained for 30 epochs using the same hyperparameter values, with 80% of the data used for training and 20% for validation in each iteration. The accuracy and loss for each fold are represented in Table III. With an average accuracy of 95.24% ± 2.64% and an average loss of 0.1639 ± 0.1100, the model exhibited low inter-fold volatility and consistent convergence.

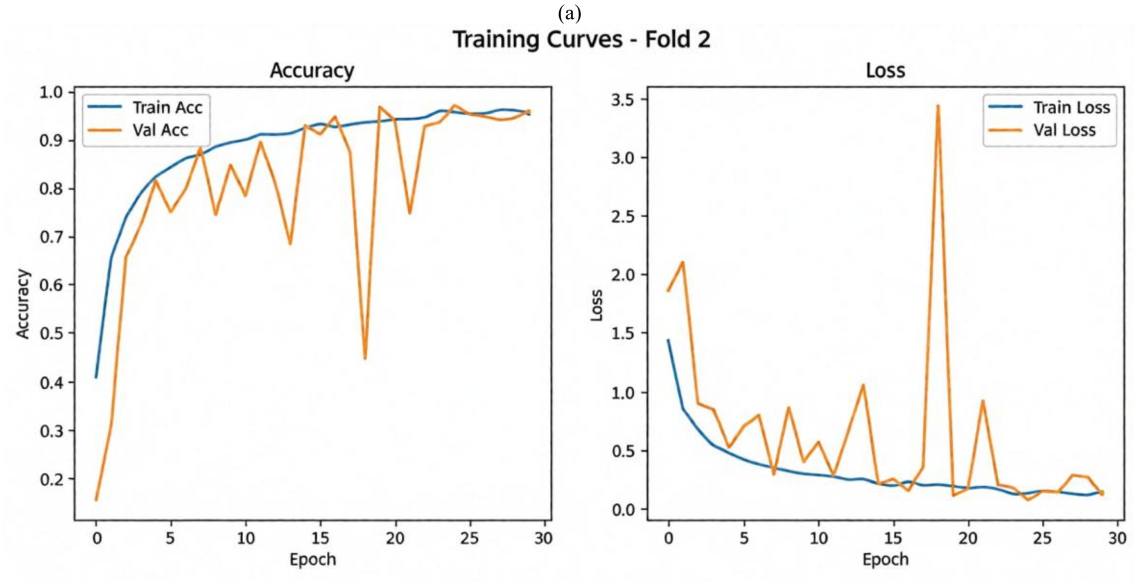
TABLE III. CROSS VALIDATION RESULTS

Metrics	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy (%)	96.28	96.06	97.98	95.69	90.21
Loss	0.1	0.12	0.08	0.13	0.38

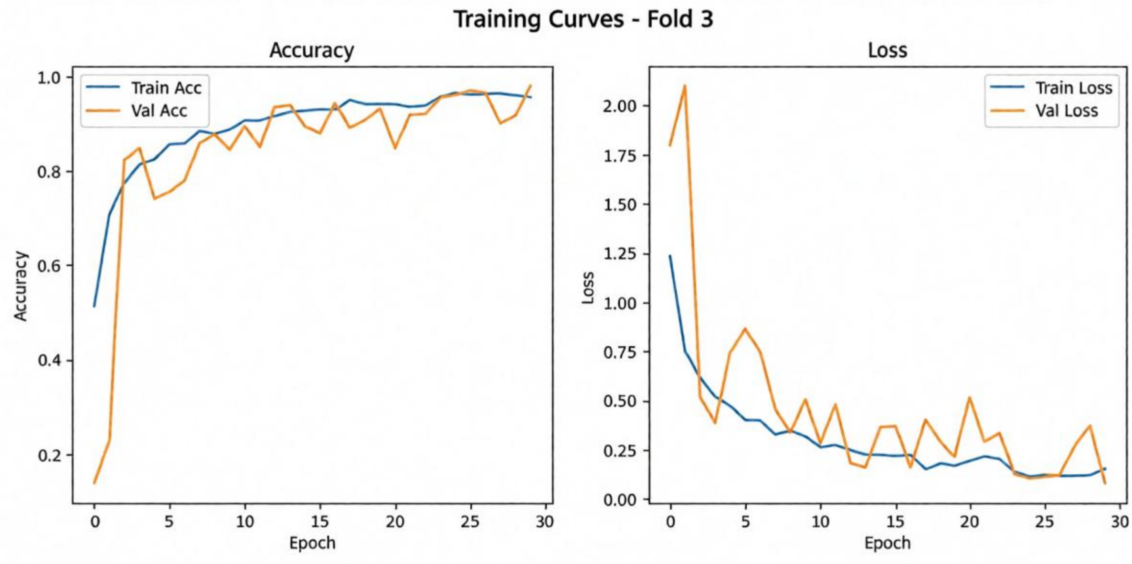
The same cross-validation splits, preprocessing method, optimizer setup, and training schedule have been employed to retrain EfficientNet-B0 for objective benchmarking. With a 95% confidence interval of [91.57%, 98.91%], CATNet achieved 95.24% ± 2.96%, beating EfficientNet-B0 (92% average accuracy). The statistical significance has been verified by a paired t-test as shown in Table IV with mean improvement +3.64%, $t = 2.79$ and $p = 0.048 < 0.05$ revealing that the performance rise is consistent and not the result of chance.



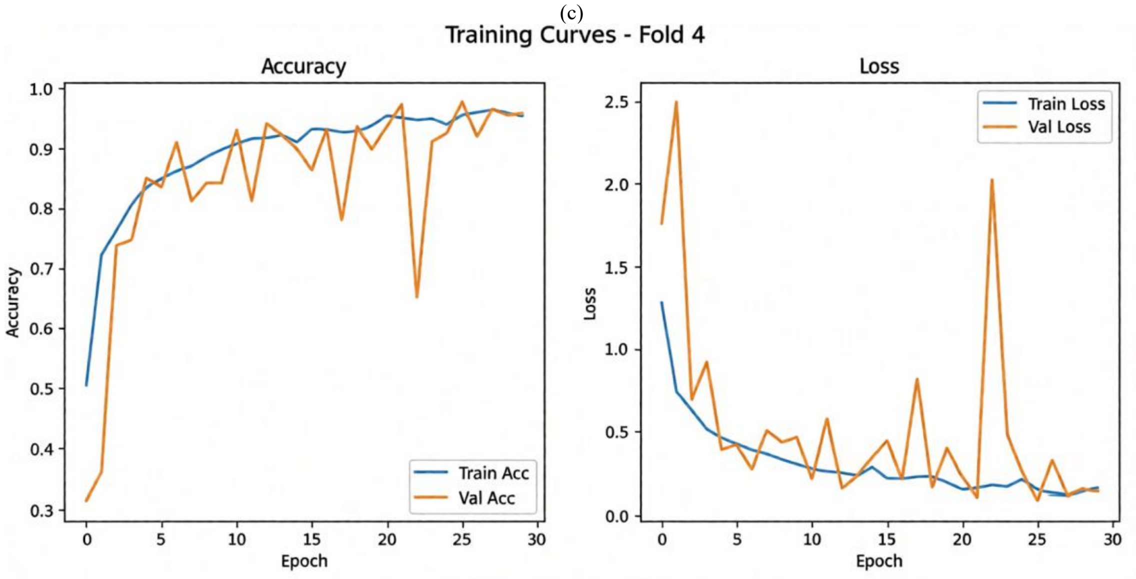
(a)



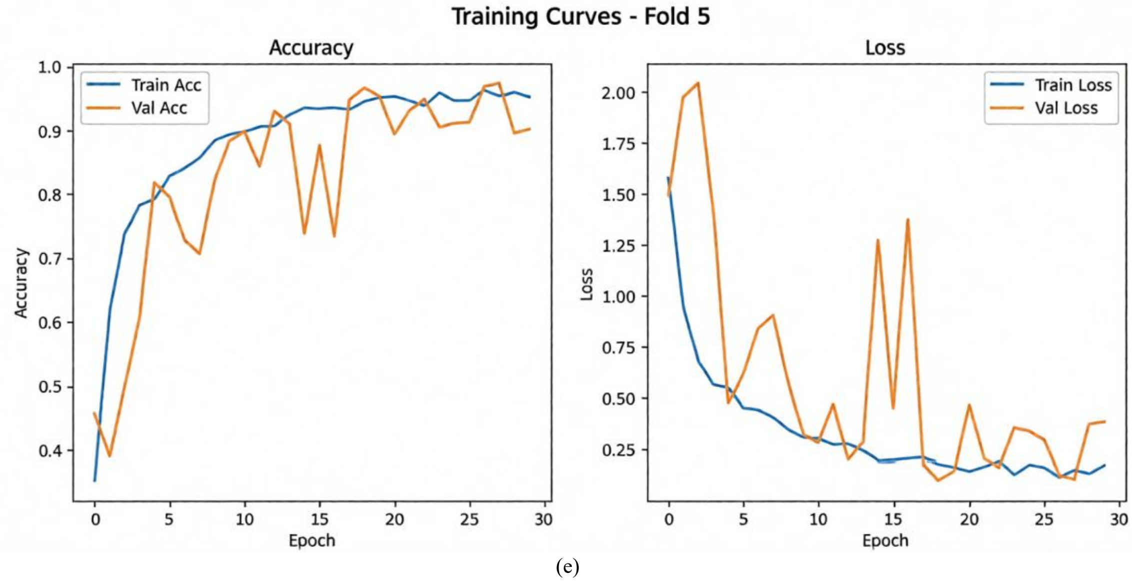
(b)



(c)



(d)



(e)

Fig. 11. 5-fold cross-validation: (a) Training curves-fold 1, (b) Training curves-fold 2, (c) Training curves-fold 3, (d) Training curves-fold 4, (e) Training curves-fold 5.

D. Comparative Evaluation with Popular Existing Models

To ensure a fair and impartial comparison with the proposed CATNet, all baseline models were implemented and retrained under identical experimental settings. Particularly, using the same self-generated rice disease dataset, identical train-validation-test splits of 80–20, consistent preprocessing procedures, and uniform augmentation strategies were used to train MobileNetV3-Small, DenseNet-169, and EfficientNet-B0. The Adam optimizer was employed to train each baseline model for 50 epochs with a batch size of 32 and a learning rate of 0.001. With the exception of the final classification layer, which has been modified in accordance with the number of rice disease classifications, default configurations were maintained and no architecture-specific hyperparameter advantages were added. It is observed in Table V that the proposed CATNet drastically reduces the computational and Random Access Memory (RAM) demands despite attaining 98% test accuracy, matching MobileNetV3-Small. 332 MFLOPs, 3.9 M parameters, a 12 ms inference time, and an exceptionally low RAM usage of 2287 MB which is less than half of MobileNetV3-Small and significantly less than DenseNet-169 and EfficientNet-B0. One path employs group convolutions to decrease parameter redundancy, whereas the other uses lightweight transformer encoders and CBAM attention modules to capture global context without expanding feature maps. CATNet is suited for real-time and resource-limited applications as it merges transformer-enhanced feature modelling, attention-based refinement and powerful convolutions that preserve high accuracy without getting high memory and compute efficiency.

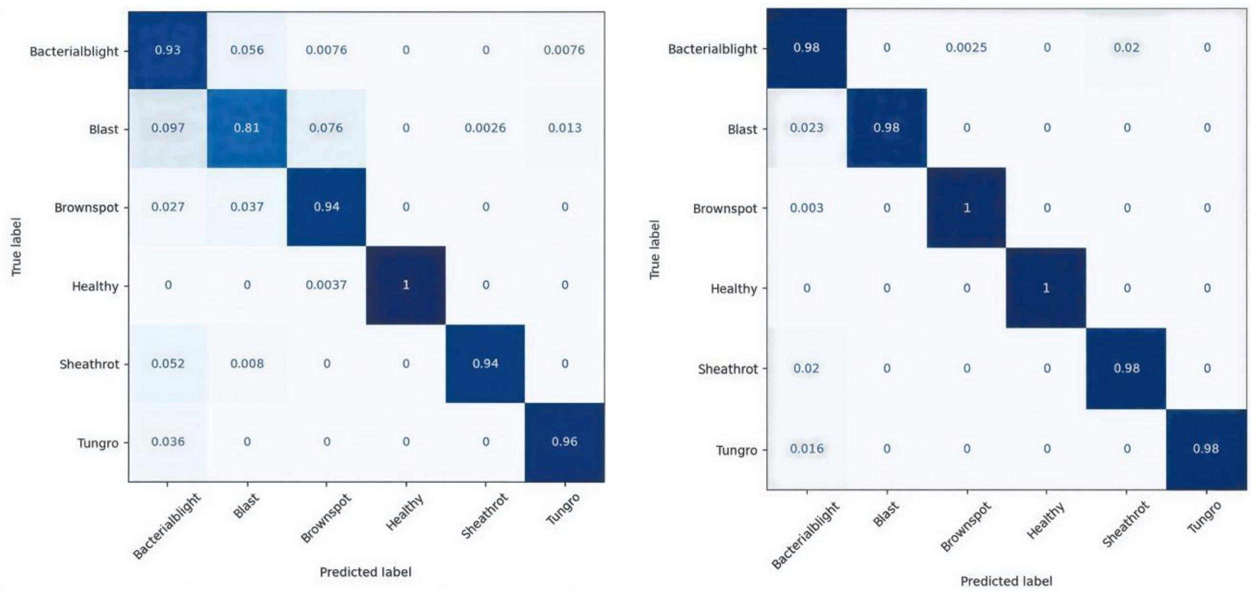
The confusion matrix depicted in Fig. 12 indicates that the proposed CATNet model transcends contemporary models by attaining a greater number of accurate predictions across all disease types, while utilizing fewer Flops and showcasing efficient inference time, despite its inclusion of transformers in the architecture. Additionally, CATNet identifies bacterial blight with an impressive 99% accuracy. CATNet predicts a small proportion of rice blast samples incorrectly since its symptoms are so identical to those of bacterial blight. Overall, the results we obtain indicate that CATNet outperformed present approaches and shows a good potential for generalization across all kinds of rice crop diseases, maintaining the highest prediction accuracy for sheath rot and balanced generalization for other classes thereby resolving the conflict between minimum claims of computational resources and performance.

TABLE IV. T-TEST

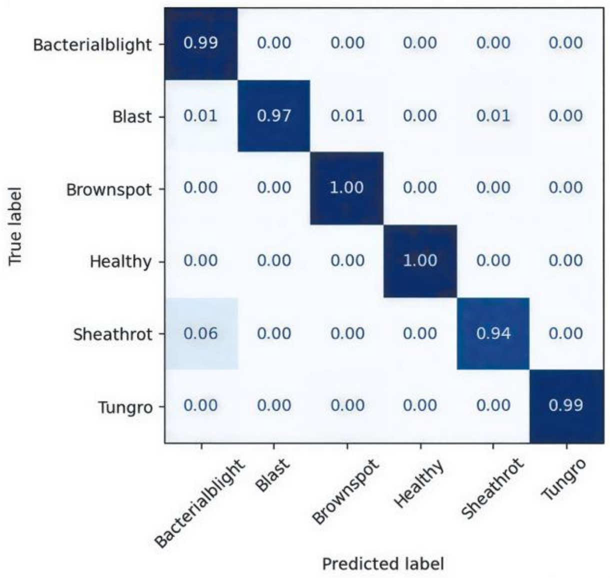
Model	Mean Accuracy (%)	Std. Deviation	95% Confidence Interval	Mean Difference (%)	t-value	*p*-value	Significance
Proposed Model	95.24	2.96	[91.57–98.91]	3.64	2.79	0.048	Significant ($p < 0.05$)
EfficientNetB0 (Base)	92	–	–	–	–	–	–

TABLE V. COMPARISON WITH EXISTING LIGHTWEIGHT MODELS

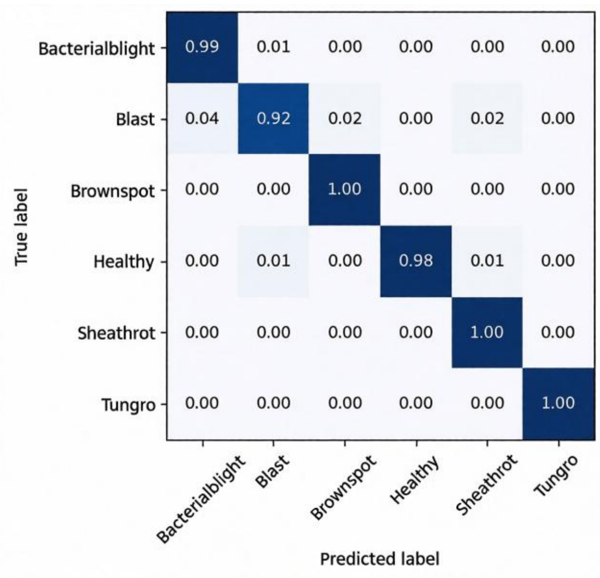
Model	Test accuracy (%)	Flops	Parameters (Millions)	Inference Time	RAM Usage (MB)
EfficientNetB0	92	0.21 G	4 M	8.77ms	4568
DenseNet169	94	3.1 G	7 M	8.25ms	5546
MobileNetV3Small	98	32 M	0.9 M	2.85ms	4247
Proposed CATNet	98	332 M	3.9 M	12ms	2287



(a) (b)



(c)



(d)

Fig. 12. Confusion matrix for comparative analysis: (a) EfficientNetB0, (b) DenseNet169, (c) MobileNetV3 Small, (d) Proposed CATNet.

V. Conclusion

The deep learning models developed in recent studies face challenges with generalization in real field environments as most of them are trained on controlled and balanced datasets. Earlier approaches like preprocessing of images using various techniques demand more computational cost making it more difficult to implement in resource-limited situations. The proposed CATNet model presented in this study is a novel and lightweight hybrid deep learning framework for robust and precise rice disease detection, achieving excellent generalization with real field dataset. The framework can simultaneously extract global contextual cues and specific spatial features due to its dual-path architecture comprising of a CBAM, a Transformer branch and Groupwise Depth wise Convolution. The attention-based convolutional stream improves local feature sensitivity while the transformer path enhances long-range relational understanding which together provides an optimal approach for learning multiple symptoms of disease effectively. This work's novel developments appear in the synergistic fusion of transformer-based contextual modelling and attention-based spatial learning, which contributes to improved generalization, effective inference, and robustness to data imbalance.

Proposed CATNet exhibits superior results with both public datasets and the self-created comprehensive dataset, reaching an ideal balance between accuracy, efficiency, and feasible deplorability. For vital diseases including brown spot, Sheathrot, and Tungro, it achieves 98% test accuracy with nearly perfect class separability with 99%–100%, effectively handling class imbalance where state-of-the-art models frequently fall short. Even though it uses a transformer and an attention module, CATNet is yet lightweight with only 3.9 M parameters and uses the least amount of 2287 MB RAM of all the models that have been studied. It has proved excellent in generalizing field images with complicated backgrounds. This trade-off provides greater resilience and robustness, validating CATNet as a practical, memory-efficient, and flexible solution for precise farming. CATNet delivers an enormous improvement in detecting paddy crop diseases and laying opens intelligent, interpretable, and scalable crop disease detection systems.

Data and Code Availability

The dataset used in this study was collected from the research fields of the agricultural university with permission. The implementation details of the proposed model are described in the manuscript. The source code supporting the findings of this study can be made available from the corresponding author upon reasonable request for academic purposes.

Ethical Approval Statement

The dataset used in this research was collected from the experimental rice fields of the agricultural university at Dapoli, Maharashtra, India, with prior permission from the respective authorities. The data consist solely of images of rice leaves affected by different diseases captured directly from the field environment. Since the study did not involve human participants, animals, or sensitive personal data, formal ethical approval was not required. The data collection process complied with institutional guidelines and posed no ethical concerns.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Sanam Kazi: Conceptualization, Methodology, Software, Investigation, Writing—Original Draft, Writing—Review & Editing, Visualization; Bhakti

Palkar: Validation, Supervision, Project Administration; Vishwayogita Savalkar: Formal Analysis, Resources; Supriya Khaitan: Investigation, Resources. All authors had approved the final version.

REFERENCES

[1] Y. Wang, U. S. R. Dhamodharan, N. Sarwar, F. A. Almalki, and Q. H. Naith, "A hybrid approach for rice crop disease detection in agricultural IoT system," *Discover Sustainability*, vol. 5, no. 1, 99, 2024. doi: 10.1007/s43621-024-00285-4

[2] H. Yang, X. Deng, H. Shen, Q. Lei, S. Zhang, and N. Liu, "Disease detection and identification of rice leaf based on improved detection transformer," *Agriculture*, vol. 13, no. 7, 1361, 2023. doi: 10.3390/agriculture13071361

[3] K. Mahadevan, A. Punitha, and J. Suresh, "Automatic recognition of rice plant leaf diseases detection using deep neural network with improved threshold neural network," *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, vol. 8, 100534, 2024. doi: 10.1016/j.prime.2024.100534

[4] R. Chhabra, R. Sharma, M. S. Hunjan *et al*., "Microstructural and metabolic variations induced by Bipolaris oryzae inciting brown spot disease of rice," *Cereal Research Communications*, vol. 51, no. 4, pp. 953–968, 2023. doi: 10.1007/s42976-023-00351-z

[5] M. J. U. Chowdhury, Z. I. Mou, R. Afrin, and S. Kibria, "Plant leaf disease detection and classification using deep learning: A review and a proposed system on Bangladesh's perspective," arXiv Print, arXiv:2501.03305, 2025.

[6] P. K. Mannepalli, A. Pathre, G. Chhabra, P. A. Ujjainkar, and S. Wanjari, "Diagnosis of bacterial leaf blight, leaf smut, and brown spot in rice leafs using VGG16," *Procedia Computer Science*, vol. 235, pp. 193–200, 2024. doi: 10.1016/j.procs.2024.04.022

[7] D. R. A. Sanya, S. F. Syed-Ab-Rahman, A. Jia *et al*., "A review of approaches to control bacterial leaf blight in rice," *World Journal of Microbiology and Biotechnology*, vol. 38, no. 7, 113, 2022. doi: 10.1007/s11274-022-03298-1

[8] S. Ramesh and R. Hebbar, "Plant disease detection using machine learning," in *Proc. Int. Conf. Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, 2018, pp. 41–45. doi: 10.1109/ICDI3C.2018.00017

[9] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016. doi: 10.1016/j.neucom.2015.09.116

[10] H. Pallathadka, P. Ravipati, G. S. Sajja *et al*., "Application of machine learning techniques in rice leaf disease detection," *Materials Today: Proceedings*, vol. 51, pp. 2277–2280, 2022. doi: 10.1016/j.matpr.2021.11.398

[11] G. Latif, S. E. Abdelhamid, R. E. Mallouhy, J. Alghazo, and Z. A. Kazimi, "Deep learning utilization in agriculture: Detection of rice plant diseases using an improved CNN model," *Plants*, vol. 11, no. 17, 2230, 2022. doi: 10.3390/plants1117223

[12] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 163, 104859, 2019.

[13] M. Shafay, T. Hassan, M. Owais, I. Hussain, S. G. Khawaja, L. Seneviratne, and N. Werghi, "Recent advances in plant disease detection: challenges and opportunities," *Plant Methods*, vol. 21, no. 1, 140, 2025. doi: 10.1186/s13007-025-01450-0

[14] M. Srivastava, V. Sisaudia, and J. Meena, "AgriTL-ViT: A vision transformer model with attention techniques for classification of plant leaf disease," *Expert Systems with Applications*, vol. 294, 128793, 2025. doi: 10.1016/j.eswa.2025.128793

[15] S. Kumar, S. S. Aravinth, S. Kollem, M. Kumar, Q. N. Naveed, A. Mubarakali, A. Bhattacherjee, and A. F. Emma, "Spatial attention-guided pre-trained networks for accurate identification of crop diseases," *Scientific Reports*, vol. 15, no. 1, 23213, 2025.

[16] X. Guo, Q. Feng, and F. Guo, "CMTNet: A hybrid CNN-transformer network for UAV-based hyperspectral crop classification in precision agriculture," *Scientific Reports*, vol. 15, no. 1, 12383, 2025. doi: 10.1038/s41598-025-97052-w

[17] A. Kaur, K. Guleria, and N. K. Trivedi, "A deep learning-based model for biotic rice leaf disease detection," *Multimedia Tools and Applications*, vol. 83, no. 36, pp. 83583–83609, 2024.

[18] A. Singh, J. Kaur, K. Singh, and M. L. Singh, "Deep transfer learning-based automated detection of blast disease in paddy crop," *Signal, Image and Video Processing*, vol. 18, no. 1, pp. 569–577, 2024. doi: 10.1007/s11760-023-02735-4

[19] R. K. Dubey and D. K. Choubey, "An efficient adaptive feature selection with deep learning model-based paddy plant leaf disease classification," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 22639–22661, 2024. doi: 10.1007/s11042-023-16247-3

[20] A. Haridasan, J. Thomas, and E. D. Raj, "Deep learning system for paddy plant disease detection and classification," *Environmental Monitoring and Assessment*, vol. 195, no. 1, 120, 2023. doi: 10.1007/s10661-022-10656-x

[21] H. Liu, Y. Cui, J. Wang, and H. Yu, "Analysis and research on rice disease identification method based on deep learning," *Sustainability*, vol. 15, no. 12, 9321, 2023. doi: 10.3390/su15129321

[22] M. Aggarwal, V. Khullar, N. Goyal *et al*., "Pre-trained deep neural network-based features selection supported machine learning for rice leaf disease classification," *Agriculture*, vol. 13, no. 5, 936, 2023. doi: 10.3390/agriculture13050936

[23] S. Nalini, N. Krishnaraj, T. Jayasankar *et al*., "Paddy leaf disease detection using an optimized deep neural network," *Computers, Materials & Continua*, vol. 68, no. 1, pp. 1117–1128, 2021. doi: 10.32604/cmc.2021.012431

[24] R. Deng, M. Tao, H. Xing *et al*., "Automatic diagnosis of rice diseases using deep learning," *Frontiers in Plant Science*, vol. 12, 701038, 2021. doi: 10.3389/fpls.2021.701038

[25] Y. Lu, Z. Li, X. Zhao *et al*., "Recognition of rice sheath blight based on a backpropagation neural network," *Electronics*, vol. 10, no. 23, 2907, 2021. doi: 10.3390/electronics10232907

[26] S. Bhattacharya, A. Mukherjee, and S. Phadikar, "A deep learning approach for the classification of rice leaf diseases," in *Proc. Intelligence Enabled Research: DoSIER 2019*, 2020, pp. 61–69. doi: 10.1007/978-981-15-2021-1_8

[27] Z. Jiang, Z. Dong, W. Jiang, and Y. Yang, "Recognition of rice leaf diseases and wheat leaf diseases based on multi-task deep transfer learning," *Computers and Electronics in Agriculture*, vol. 186, 106184, 2021. doi: 10.1016/j.compag.2021.106184

[28] D. Mandwariya and V. Jotwani, "A deep learning approach for automated rice disease detection and classification," *Journal of Advanced Zoology*, vol. 45, no. 3, 316, 2024.

[29] K. Saddami, Y. Nurdin, M. Zahramita, and M. S. Safiruz, "Advancing green AI: Efficient and accurate lightweight CNNs for rice leaf disease identification," arXiv Print, arXiv:2408.01752, 2024. doi: 10.48550/arXiv.2408.01752

[30] C. Xu and T. Yang, "A symmetric multi-scale convolutional transformer network for plant disease image classification," *Symmetry*, vol. 17, no. 8, 1232, 2025. doi: 10.3390/sym17081232

[31] Y. Wang, Y. Chen, and D. Wang, "Convolution network enlightened transformer for regional crop disease classification," *Electronics*, vol. 11, no. 19, 3174, 2022.

[32] R. Patil and S. Kumar, "Rice transformer: A novel integrated management system for controlling rice diseases," *IEEE Access*, vol. 10, pp. 87698–87714, 2022.

[33] S. T. Ahmed *et al*., "Enhancing precision in rice leaf disease detection: A transformer model approach with attention mapping," in *Proc. 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS)*, 2024, pp. 1–6.

[34] W. F. L. Chin and W. S. Suhaili, "Paddy disease detection through vision transformers," in *Proc. IOP Conference Series: Earth and Environmental Science*, 2024, vol. 1426, 012003.

[35] Y. Zhang, H. Yu, and Z. Zhai, "ResViT-Rice: A deep learning model combining residual module and transformer encoder for accurate detection of rice diseases," *Agriculture*, vol. 13, no. 6, 1264, 2023. doi: 10.3390/agriculture13061264

[36] P. K. Sethy. (2020). Rice leaf disease image samples. *Mendeley Data*. [Online]. Available: https://data.mendeley.com/datasets/fwcj7stb8r/1