Journal of Integrative Agriculture 2026, 25(2): 756–768

Available online at ScienceDirect

Journal of Integrative Agriculture



RESEARCH ARTICLE

E2ETCA: End-to-end training of CNN and attention ensembles for rice diseasediagnosis

Md. Zasim Uddin[1#], Md. Nadim Mahamood[1], Ausrukona Ray[1], Md. Ileas Pramanik[1], Fady Alnajjar[2#], Md Atiqur Rahman Ahad[3]

[1] *Department of Computer Science and Engineering, Begum Rokeya University, Rangpur 5404 , Bangladesh*
[2] *Department of Computer Science and Software Engineering, United Arab Emirates University, Alain 15551, UAE*
[3] *Department of Computer Science and Digital Technologies, University of East London, London E16 2RD, UK*

Highlights

- We introduce an end-to-end ensemble framework that fuses features from Inception V3, DenseNet-201, and a vision transformer to improve robustness and generalizability in rice disease diagnosis.

Abstract

Rice is one of the most important staple crops globally. Rice plant diseases can severely reduce crop yields and, in extreme cases, lead to total production loss. Early diagnosis enables timely intervention, mitigates disease severity, supports effective treatment strategies, and reduces reliance on excessive pesticide use. Traditional machine learning approaches have been applied for automated rice disease diagnosis; however, these methods depend heavily on manual image preprocessing and handcrafted feature extraction, which are labor-intensive and time-consuming and often require domain expertise. Recently, end-to-end deep learning (DL) models have been introduced for this task, but they often lack robustness and generalizability across diverse datasets. To address these limitations, we propose a novel end-to-end training framework for convolutional neural network (CNN) and attention-based model ensembles (E2ETCA). This framework integrates features from two state-of-the-art (SOTA) CNN models, Inception V3 and DenseNet-201, and an attention-based vision transformer (ViT) model. The fused features are passed through an additional fully connected layer with softmax activation for final classification. The entire process is trained end-to-end, enhancing its suitability for real-world deployment. Furthermore, we extract and analyze the learned features using a support vector machine (SVM), a traditional machine learning classifier, to provide comparative insights. We evaluate the proposed E2ETCA framework on three publicly available datasets, the Mendeley Rice Leaf Disease Image Samples dataset, the Kaggle Rice Diseases Image dataset, the Bangladesh Rice Research Institute dataset, and a combined version of all three. Using standard evaluation metrics (accuracy, precision, recall, and F1-score), our framework demonstrates superior performance compared to existing SOTA methods in rice disease diagnosis, with potential applicability to other agricultural disease detection tasks.

Keywords: rice disease diagnosis, ensemble method, CNN-based model, end-to-end model, Inception model, DenseNet model, vision transformer model, attention-based model, support vector machine

1. Introduction

Global population growth has intensified the demand for food security. Over 800 million people lack sufficient access to food, and approximately 1.3 billion live on less than one USD per day, making it difficult to afford adequate nutrition (Strange and Scott 2005). Plant diseases significantly impair agricultural productivity, negatively affecting economies and undermining efforts to meet rising food demands. According to the Food and Agriculture Organization (FAO), plant pests and diseases cause annual losses of USD 220 billion and reduce global crop yields by 20–40% (Agrios 2005).

Rice is a critical staple crop, with consumption steadily

Received 30 October 2023; Received in revised form 25 December 2023; Accepted 16 January 2024; Available online 27 March 2024
[#]Correspondence Md. Zasim Uddin, E-mail: zasim@brur.ac.bd; Fady Alnajjar, E-mail: fady.alnajjar@uaeu.ac.ae

© 2026 CAAS. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer review under responsibility of Editorial Board of *Journal of Integrative Agriculture*.
doi: 10.1016/j.jia.2024.03.075

increasing worldwide. However, rice diseases reduce production by 10–15% annually (Peng *et al.* 2009), and under severe conditions, yield losses can reach 40–50%, or even result in complete crop failure (Jiang *et al.* 2020). These diseases compromise both yield and grain quality, making early identification and control essential for sustainable rice cultivation (Deng *et al.* 2021).

Early diagnosis of rice diseases is crucial to minimize damage and restore optimal production levels. Manual inspection, though commonly used, relies on visual assessment and expert knowledge. For large-scale farms, this approach is costly, time-consuming, and impractical due to the need for continuous monitoring (Sethy *et al.* 2020; Zeng *et al.* 2023). Farmers often apply pesticides without precise knowledge of the specific disease or optimal timing, leading to ineffective treatments and environmental contamination due to overuse (Andrianto *et al.* 2020). Excessive chemical application also contributes to ecological pollution (Durmus *et al.* 2017). Therefore, developing an automated, continuous monitoring system is essential for accurate and timely disease diagnosis (Sharma M *et al.* 2022a).

Modern automated approaches to rice disease diagnosis fall into two categories: traditional machine learning (ML)-based methods and deep learning (DL)-based techniques. Traditional ML methods have been widely adopted due to their lower computational requirements and faster processing compared to manual diagnosis (Udayananda *et al.* 2022). For instance, Islam *et al.* (2018) utilized a naive Bayes classifier based on red, green, and blue (RGB) pixel values from infected leaf regions. Ahmed *et al.* (2019) segmented diseased areas, removed background noise, and applied K-nearest neighbor (K-NN), decision trees, naive Bayes, and logistic regression for classification. Despite promising results, these methods suffer from limited generalizability and dependence on manually engineered features, which may fail to capture complex visual patterns in disease manifestations.

In contrast, DL-based approaches offer greater potential for scalable and accurate disease classification. Lu *et al.* (2017) implemented a convolutional neural network (CNN) to classify ten common rice diseases. Ghosal and Sarkar (2020) employed a pre-trained VGG-16 CNN to distinguish between healthy plants and those affected by three diseases. Sharma M *et al.* (2022a) evaluated six pre-trained DL models (MobileNet, VGG-16, Inception V3, ResNet-50, Xception, and InceptionResNet V2) trained on ImageNet, focusing on bacterial blight, rice blast, and brown spot. While these studies demonstrate the effectiveness of DL, most rely on single end-to-end models that are prone to overfitting, limiting their adaptability and generalization across different datasets (Bejani and Ghatee 2021).

Ensemble learning has recently gained attention due to its enhanced performance and robustness (Talukder and Sarkar 2023). By combining multiple base models, ensemble methods create a more powerful composite model capable of outperforming individual components. The diversity among constituent models also helps mitigate overfitting (Mohammed and Kora 2023). Several studies have applied ensemble models to rice disease diagnosis. For example, Ahad *et al.* (2023) combined DenseNet-121, EfficientNetB7, and Xception to compare individual and ensemble performance on a nine-class dataset. Sharma M *et al.* (2022b) constructed binary, ternary, and quaternary ensembles using InceptionResNetV2, Xception, DenseNet-201, and VGG-19, averaging prediction probabilities. Deng *et al.* (2021) integrated DenseNet-121, SE-ResNet-50, and ResNeSt-50, averaging submodel outputs for final predictions. Turkoglu *et al.* (2022) proposed an ensemble model based on six state-of-the-art (SOTA) CNNs using majority voting to improve accuracy. However, these approaches typically employ post-hoc fusion strategies such as weighted averaging or voting, and the overall training process is not end-to-end, limiting joint optimization and real-time applicability.

End-to-end ensemble models unify data preprocessing, feature extraction, and classification within a single trainable framework, enabling seamless optimization and improved performance on complex real-world problems (Caruana *et al.* 2004; Serbetci and Akgul 2020). In this work, we introduce an end-to-end ensemble framework (E2ETCA) for rice disease diagnosis that integrates two CNN-based models and an attention-based vision transformer. Features from these models are concatenated and processed through a shared fully connected layer with softmax activation for final classification. The entire framework is trained jointly, ensuring coherent feature learning and optimal integration. We evaluate E2ETCA on three public datasets and their combination, demonstrating SOTA performance. Additionally, we extract high-level features from the trained model and feed them into a support vector machine (SVM) for non-end-to-end analysis, allowing comparative evaluation. We also assess each component model individually, both in end-to-end and SVM-based configurations, to understand their relative contributions.

2. Related work

2.1. Traditional ML-based approaches

Traditional ML-based approaches to rice disease diagnosis automate detection by relying on handcrafted features extracted from leaf images. Phadikar *et al.* (2012) used morphological changes and radial hue distribution from lesion boundaries as input features, achieving classification accuracies of 79.5 and 68.1% with naive Bayes and SVM, respectively, on a dataset of 1,000 images representing two diseases. Islam *et al.* (2018) accelerated diagnosis by extracting low-level features such as RGB values from localized diseased regions and applied a naive Bayes classifier to identify three rice diseases.

Joshi and Jadhav (2016) used shape and color features of diseased leaf portions, applying a minimum distance classifier (MDC) and K-NN for four disease classes. On a dataset of 115 leaf images, they achieved 87.0 and 89.2% accuracy with MDC and K-NN, respectively. Kishore and Kannan (2022) extracted color, shape, and texture features, employing a discrete wavelet transform to isolate affected regions and remove healthy green tissue. Using an adaptive boosting SVM, they achieved a maximum accuracy of 98.8% for three disease classes. Bhartiya *et al.* (2022) utilized shape descriptors such as area, roundness, and area-to-lesion ratio

to differentiate disease types, attaining 81.8% accuracy with a quadratic SVM.

Despite their utility, traditional ML-based methods are constrained by the need for manual feature engineering, which is labor-intensive and requires domain-specific knowledge. Moreover, handcrafted features may inadequately represent the intricate visual characteristics of plant diseases, limiting model generalization and practical deployment in real-world agricultural settings.

2.2. DL-based approaches

CNNs have become a dominant paradigm in computer vision, enabling end-to-end learning for tasks such as biometrics (Uddin *et al.* 2018, 2019), medical imaging (Anwar *et al.* 2018), and person re-identification (Serbetci and Akgul 2020). These models have also been extensively applied to rice disease detection. Jiang *et al.* (2020) combined a CNN with an SVM for classification, while Ramesh and Vydeki (2020) and Nalini *et al.* (2021) used K-means clustering for disease segmentation and diagnosis. Transfer learning with pre-trained CNNs has further advanced the field. Temniranrat *et al.* (2021) compared Faster R-CNN, RetinaNet, YOLO V3, and Mask R-CNN on a dataset of 8,767 images across six disease classes, finding YOLO V3 to be the most accurate (79.2%). Studies by Gautam *et al.* (2022) and Simhadri and Kondaveeti (2023) evaluated six pre-trained models, including VGG-16, VGG-19, ResNet, Inception, MobileNet, and SqueezeNet, and found Inception to perform best. Sudhesh *et al.* (2023) implemented ten transfer-learned CNNs with dynamic mode decomposition (DMD)-based attention mechanisms for image preprocessing. Generative adversarial networks (GANs) have also been explored; for example, Stephen *et al.* (2023) used 2D and 3D CNNs for feature extraction and a GAN optimized *via* an improved backtracking search algorithm for classification.

Vision transformers (ViT) (Dosovitskiy *et al.* 2020) leverage self-attention mechanisms to model long-range dependencies in images by dividing them into fixed-size patches and processing them as sequences. ViTs have shown strong performance in image classification and plant disease diagnosis. Borhani *et al.* (2022) applied a ViT-based method using image patches, achieving 91.7% accuracy on a small dataset of 120 images across three disease classes. Hybrid architectures combining CNNs and ViTs have also emerged. Zhang *et al.* (2023) proposed ResViT-Rice, embedding a ViT block with self-attention into ResNet. Evaluated on a dataset of 1,548 images (two diseases and healthy), their model achieved 99.1% accuracy.

Ensemble models enhance predictive performance by integrating multiple learners, improving generalization, and reducing overfitting (Opitz and Maclin 1999; Serbetci and Akgul 2020). Recent works have applied ensembles to rice disease diagnosis. For example, Ahad *et al.* (2023) combined DenseNet-121, EfficientNet B7, and Xception, achieving 97.6% accuracy on a nine-class dataset, a 17.0% improvement over a single Seresnext101 model. Putra *et al.* (2022) developed two fusion-based ensembles, MobileNet+DenseNet and DenseNet+XceptionNet, achieving 40.0 and 60.0% higher accuracy than individual models on a three-disease dataset. Deng *et al.* (2021) averaged predictions from DenseNet-121, SE-ResNet-50, and ResNet-50, reaching 91.0% accuracy on a large dataset of 33,026 images across six diseases. Turkoglu *et al.* (2022) combined six SOTA CNNs and evaluated combinations using SVM, achieving improved accuracy.

However, these ensemble methods rely on late fusion strategies such as weighted averaging or majority voting, and the training process is not end-to-end. In contrast, an end-to-end ensemble allows joint optimization of all components, leading to better feature alignment, reduced overfitting, and faster inference, critical for real-time applications. This paper proposes E2ETCA, an end-to-end framework that fuses features from two CNN-based models and an attention-based ViT model within a unified, trainable architecture.

3. Proposed framework

3.1. Overview of framework

Conventional ensemble methods typically train base models independently, either in parallel or sequentially, which is computationally prohibitive in deep learning contexts. Additionally, such approaches require post-training fusion strategies like voting or averaging. To overcome these limitations, we propose an end-to-end training framework, E2ETCA, that jointly trains multiple base learners within a unified architecture. The ensemble comprises two CNN-based models and one attention-based ViT. An overview of the framework is illustrated in Fig. 1. Specifically, we integrate Inception V3 (Szegedy *et al.* 2016) and DenseNet-201 (Huang *et al.* 2017), two well-established CNN architectures, along with ViT-L/32, a patch-based attention model (Dosovitskiy *et al.* 2020). Each model processes the input image and generates feature maps, which are flattened and combined via pointwise addition. The fused features are then passed through a fully connected embedding layer, followed by a softmax-equipped classification layer to compute cross-entropy loss.

More specifically, given a set of rice disease training images $X=\{x_i\}$ with height H, width W, and channel C, and consisting of N images corresponding to K distinct classes with labels $Y=\{y_i\}$, let the feature extraction network be represented by a function $h(x):\mathrm{R}^{H\times W\times C}\rightarrow\mathrm{R}^{D}$ that maps a given input image x to the D-dimensional feature embedding space. The function $h(x)$ extracts the features from the base models and combines them as follows:

$$F_{\text{final}}=F_i \oplus F_d \oplus F_v \quad (1)$$

where F_i, F_d, and F_v are the features from Inception V3, DenseNet-201, and the ViT network, respectively, and F_{final} is the combined feature pointwise. Then, a one-layer classification network outputs K-dimensional label probabilities for the D-dimensional feature vector as $f(h(x)):\mathrm{R}^{D}\rightarrow\mathrm{R}^{K}$.

The cross-entropy loss L_{CEL} is used to optimize the model:

$$L_{\text{CEL}}=-\sum_{K=1}^{K} y_i^k \log(f_i^k) \quad (2)$$

where f_i^k and y_i^k denote the estimated and actual probabilities of the ith image with kth rice disease class.

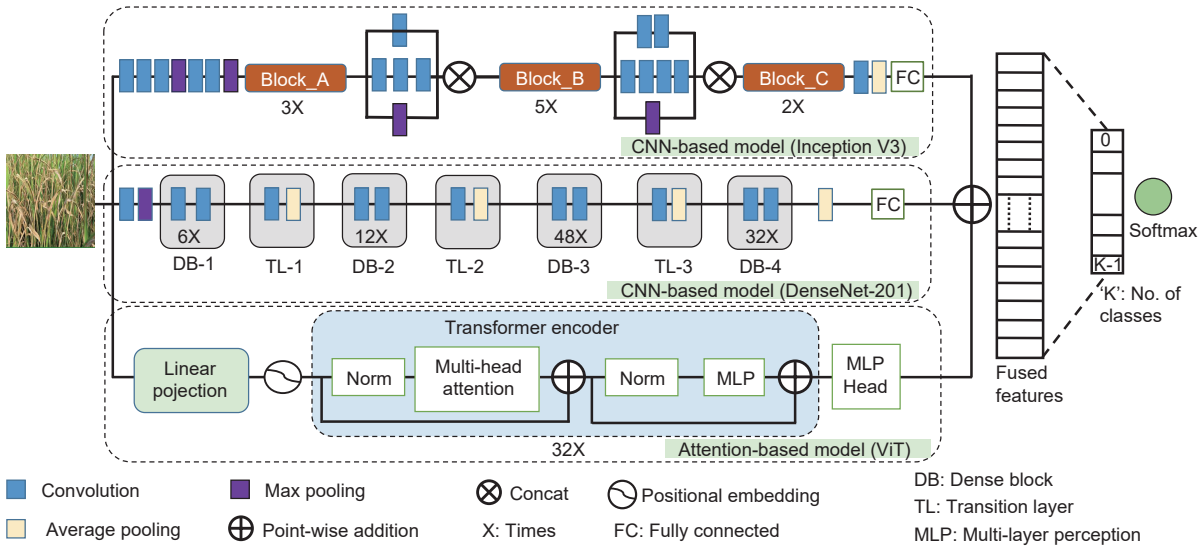


Fig. 1 Overview of the proposed E2ETCA framework, which comprises three modules: two CNN-based models and an attention-based model. The features extracted by these modules are integrated and fed into classification layers, consisting of a fully connected (FC) layer and a softmax function, to classify rice diseases. The entire framework is trained and evaluated in an end-to-end manner.

3.2. CNN-based model

We employ two SOTA CNN architectures: Inception V3 (Szegedy *et al.* 2016) and DenseNet-201.

Inception V3 (Szegedy *et al.* 2016) consists of three distinct Inception modules, Block_A, Block_B, and Block_C, as shown in Fig. 2. Block_A performs spatial aggregation by replacing large convolutions with smaller ones, reducing computational load. It also uses 1×1 convolutions to decrease channel depth and maintain efficiency without sacrificing representational capacity. Block_B extends Block_A with additional convolutional layers. Block_C incorporates factorized convolutions and asymmetric filters, enhancing the model's ability to learn multi-scale features. This hierarchical design enables rich, high-dimensional feature representations. Further architectural details are available in Szegedy *et al.* (2015, 2016).

An input image of rice disease, $x \in R^{H\times W\times C}$, first passes through a hybrid function $h_{hvbrid}(x)$: $R^{H\times W\times C}\rightarrow R^{P_{i-1}\times Q_{i-1}}$ to generate a feature map that consists of several convolutions and pooling with different sizes and maps the feature dimension as follows:

$$In_{i-1}=h_{hybrid}(x) \quad (3)$$

This feature map then passes through several Inception blocks, $h_{block}(In_{i-1}):R^{P_{i-1}\times Q_{i-1}}\rightarrow R^{P_i\times Q_i}$ from the (i-1)th to ith layer,

$$In_i=h_{block}(In_{i-1}) \quad (4)$$

Following multiple stages of hybrid processing, Inception block transformations, and concatenation of features extracted *via* parallel convolutional and pooling paths, the resulting feature vector is fed into a fully connected layer to generate the final output features:

$$F_i=FC(In_{n-1}) \quad (5)$$

where In_{n-1} is the feature map generated from the (n-1)th layer of the network, and $FC(\cdot)$ converts the feature map of dimensions $P_{n-1}\times Q_{n-1}$ into a D-dimensional feature vector F_i, which is used for feature fusion as described in eq. (1).

In this study, we employ DenseNet-201 (Huang *et al.* 2017) as the backbone feature extractor. DenseNet-201 is a densely connected convolutional neural network consisting of 201 layers and is known for its computation and memory consumption (Sanghvi *et al.* 2023). An input image of rice disease x passes through convolution and pooling layers and then to a dense block, in a process that can be expressed as $x \in R^{H\times W\times C}\rightarrow E_1 \in R^{P_1\times Q_1}$.

Each layer of a dense block receives input from all the preceding layers in a feedforward manner. The output E_m from the mth layer of the block is calculated as:

$$E_m=h_{dense}(cat\,[E_1,E_2,\ldots\ldots\ldots,E_{m-1}]) \quad (6)$$

where $h_{dense}(\cdot)$ represents the composite function in this layer, and cat is the concatenation function processing between each feature layer inside it. Between successive dense blocks, the network incorporates transition layers designed to reduce the spatial resolution and channel dimensionality of feature maps. Each transition layer consists of a convolutional operation with 1×1 filters followed by 2×2 average pooling. After traversing all dense blocks and their corresponding transition layers, the resulting feature map is subjected to an average pooling. This pooled output is then passed through a fully connected layer to generate the D-dimensional feature vector F_d, which is used for fusion in eq. (1).

3.3. Attention-based vision transformer model

Vision transformers (Dosovitskiy *et al.* 2020) model both global and local visual contexts by dividing an image into a sequence of fixed-size patches. Positional embeddings are added to preserve spatial information, forming an input sequence that is subsequently processed by a transformer encoder. Given an input image $x \in R^{H\times W\times C}$, it is reshaped into a sequence of flattened patches $xp \in R^{N\times p2C}$, where each patch has dimensions (p,p) corresponding to its height and

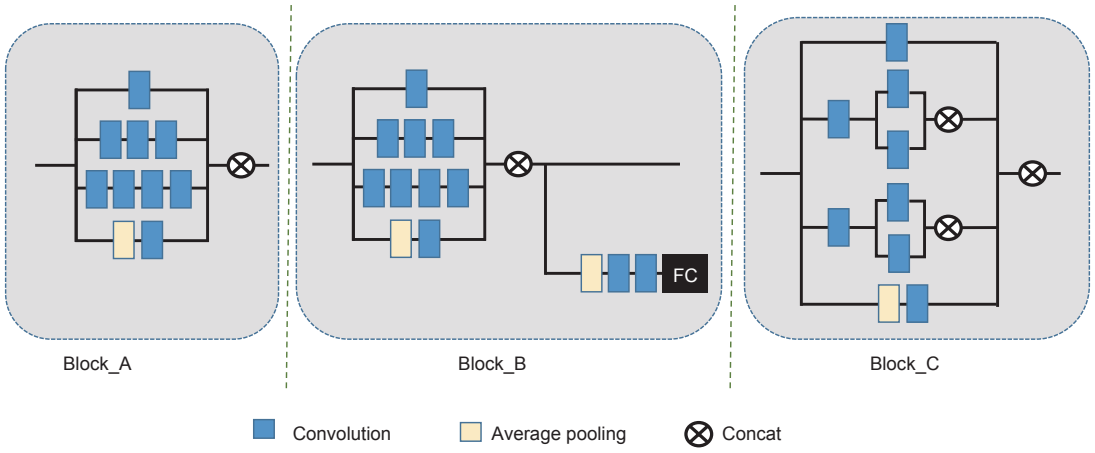


Fig. 2 The three Inception V3 blocks, Block_A, Block_B, and Block_C, each comprising convolutional layers of varying sizes with multiple parallel routes. The outputs from all routes are concatenated to produce the final feature representation.

width, and N represents the total number of patches, which serves as the effective input sequence length for the transformer:

$$x_p=\text{Patch}(x) \quad (7)$$

Then, each patch is flattened and mapped to a T_d-dimensional latent vector with a trainable linear projection $\alpha \in R^{p^2\times T_d}\alpha$, which is used by the transformer for all layers. This is referred to as patch embedding. Then, patch embeddings are integrated with positional embeddings to control the positional information of all patches, i.e, $\alpha_{pos} \in R^{(N+1)\times T_d}$, and to generate features β_0 as the input for an encoder:

$$\beta_0=[x_{class}; x_p^1\alpha; x_p^2\alpha;...; x_p^N\alpha]+\alpha_{pos} \quad (8)$$

The encoder comprises multiple stacked layers, each consisting of multiheaded self-attention (MSA) and multilayer perceptron (MLP) submodules. Within each encoder block, the input features are first normalized and passed through the MSA module, followed by a residual connection. The output is then normalized again and processed through the MLP, with another residual connection applied. These operations are mathematically expressed as:

$$\beta_i'=MSA(L_{norm}(\beta_{i-1}'))+\beta_{i-1}', i=1,...,n \quad (9)$$

$$\beta_i'=MSA(L_{norm}(\beta_i'))+\beta_i', i=1,...,n \quad (10)$$

After a defined number n of encoders, the output of the last encoder, βn^0, is passed through the final MLP, known as the MLP head:

$$F_v=MLP_{head}(L_{norm}(\beta_n^0)) \quad (11)$$

$MLP_{head}(\bullet)$ consists of a single fully connected (FC) layer that generates the final feature map F_v with dimension D, which is subsequently used in eq. (1).

4. Experiments

4.1. Datasets and evaluation protocols

We evaluated our proposed ensemble framework on three publicly available datasets and a combined version of these datasets.

The Mendeley Rice Leaf Disease Image Samples dataset (Sethy *et al.* 2020; https://data.mendeley.com/datasets/fwcj7stb8r/1), henceforth referred to as the Mendeley dataset, contains a total of 5,932 images representing four diseases: bacterial blight, blast, brown spot, and tungro. These images were collected from rice fields in Sambalpur and Bargarh Districts, Odisha, India, using a Nikon DSLR-D5600 camera (Japan) equipped with an 18–55 mm high-resolution lens. Dataset statistics are summarized in Table 1, and representative images for each disease class are displayed in Fig. 3. Various evaluation protocols exist in the literature; to ensure fair comparison, we strictly adopted a widely used protocol from prior studies (Hassan and Maji 2022; Sharma R *et al.* 2022; Singh *et al.* 2023; Sudhesh *et al.* 2023). Specifically, we employed fivefold cross-validation to assess the performance of the proposed E2ETCA framework.

The Kaggle Rice Diseases Image dataset (https://www.kaggle.com/datasets/minhhuy2810/rice-diseases-image-dataset), hereafter referred to as the Kaggle dataset, is a publicly accessible collection comprising 1,600 images of three rice diseases, brown spot, hispa, and leaf blast, along with images of healthy plants. The dataset was gathered

Table 1 Distribution of sample images for each rice disease class for the Mendeley, Kaggle, and BRRI datasets[1)]

Class	BH	BB	BS	FS	HI	NB	ST	HE	LB	TU	SB	Total
Mendeley	–	1,584	1,600	–	–	–	–	–	1,440	1,308	–	5,932
Kaggle	–	–	400	–	400	–	–	400	400	–	–	1,600
BRRI	71	138	111	93	73	286	201	234	–	–	219	1,426
Combined	71	1,722	2,111	93	473	286	201	634	1,840	1,308	219	8,958

[1)] BH, brown plant hopper; BB, bacterial blight/bacterial leaf blight; BS, brown spot; FS, false smut; HI, hispa; NB, neck blast; ST, stemborer; HE, healthy; LB, leaf blast/blast; TU, tungro; SB, sheath (blight and/or rot). Bacterial blight and bacterial leaf blight are treated as a single class; leaf blast and instances labeled as blast are grouped under a unified blast category.

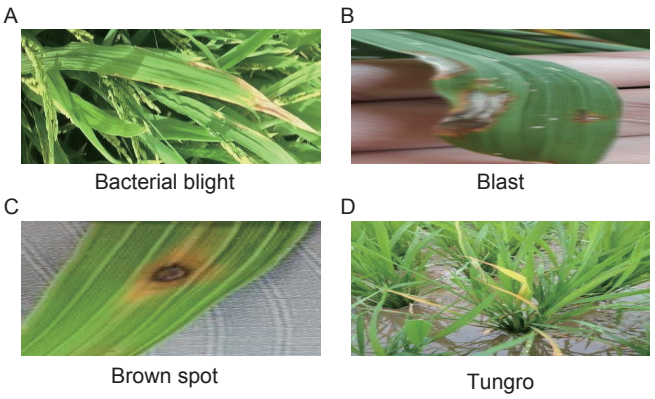


Fig. 3 Example images for each rice disease from the Mendeley Rice Leaf Disease Image Samples dataset.

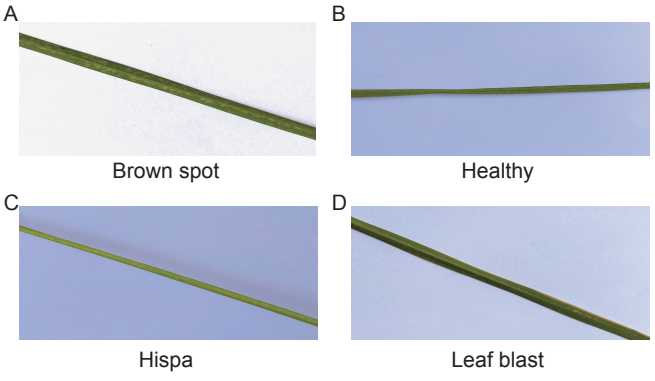


Fig. 4 Example image for each rice disease class of the Kaggle Rice Diseases Image dataset.

from a rice farmland. Detailed composition is provided in Table 1, and sample images for each class are illustrated in Fig. 4. To enable direct comparison with SOTA methods, we applied fivefold cross-validation to evaluate the proposed framework, consistent with previous works (Wang *et al.* 2021; David *et al.* 2022; Al-Gaashani *et al.* 2023).

The Bangladesh Rice Research Institute dataset (https://drive.google.com/drive/folders/1ewBesJcguriVTX8sRJseCDbXAF_T4akK?usp=drive_open)(Rahman *et al.* 2020) (henceforth referred to as the BRRI dataset) contains 1,426 images across nine classes, collected from paddy fields at the Bangladesh Rice Research Institute. The images were captured over a seven-month period under diverse environmental conditions, including winter, summer, and overcast weather, and from heterogeneous backgrounds to ensure a representative sample. To enhance model robustness, images were acquired using four different cameras. The dataset comprises five disease classes, three pest classes, and one class for healthy plants; sheath blight and sheath rot are grouped into a single class due to their co-occurrence, similar symptomatic locations, and shared treatment protocols. Dataset statistics are presented in Table 1, and representative images for each class are displayed in Fig. 5. For fair comparison with existing approaches (Rahman *et al.* 2020; Gautam *et al.* 2022), all nine classes were utilized, and experiments were conducted using 10-fold cross-validation, with average performance metrics reported across folds.

To evaluate the proposed E2ETCA framework on a large-scale, multi-class dataset, we constructed a *combined* dataset by integrating the Mendeley, Kaggle, and BRRI datasets. This merged dataset includes 11 classes: false smut (FS),



Fig. 5 Example images from each class of the Bangladesh Rice Research Institute dataset.

Table 2 Metrics used to evaluate the proposed E2ETCA framework

Metric	Expressions
Accuracy	$\frac{\lvert TP+TN\rvert}{\lvert TP+TN+FP+FN\rvert}$
Recall/Sensitivity	$\frac{\lvert TP\rvert}{\lvert TP+FN\rvert}$
Precision	$\frac{\lvert TN\rvert}{\lvert TN+FP\rvert}$
Specificity	$\frac{\lvert TN\rvert}{\lvert TN+FP\rvert}$
F1-score	$\frac{2\times Precision\times Recall}{Precision+Recall}$

brown plant hopper (BH), bacterial leaf blight (BB), neck blast (NB), stemborer (ST), hispa (HI), sheath blight and/or sheath rot (SB), brown spot (BS), blast (LB), tungro (TU), and healthy (HE) plants, totaling 8,958 images. Notably, bacterial blight and bacterial leaf blight are consolidated into a single class labeled "bacterial blight", and leaf blast and blast are unified under the label "leaf blast". The distribution of samples per class in the combined dataset is detailed in Table 1. In each experiment, the dataset was randomly partitioned into training and test sets at an 80:20 ratio, and all experiments were repeated five times to minimize the influence of random variation.

The proposed E2ETCA framework was evaluated using multiple performance metrics, including accuracy, precision, recall, and F1-score. The definitions of these evaluation metrics and their corresponding calculation formulas are presented in Table 2. Quantitative assessment was based on true positives (TP), TN, false positives (FP), and false negatives (FN). In this context, TP and TN represent correctly predicted labels, whereas FP and FN indicate misclassifications. Specifically, a correct prediction for a given disease class is counted as a TP. A prediction that assigns a non-target disease to that class is classified as an FP. Conversely, if a disease belongs to the target class but is misclassified into another category, it is recorded as an FN.

4.2. Training and testing

Training: Labeled rice disease images were input into the proposed E2ETCA framework, yielding a predicted label *fik*, representing the predicted probability for the *i*th training sample belonging to the *k*th rice disease class. The cross-entropy loss between the predicted and true labels was computed as defined in eq. (2), and the Adam optimizer was used to update the model's weights and biases. A batch size of 16 was employed during training. Deep learning models typically require large datasets to achieve optimal performance and mitigate overfitting (Paymode and Malode 2022). To address limited data availability, we applied data augmentation techniques to synthetically expand the dataset, enabling the model to better generalize and capture underlying patterns. For each original image, we generated augmented variants using resizing, random erasing, normalization (Zhang *et al*. 2021), vertical and horizontal flipping, rotation, and color jitter (Krizhevsky *et* al. 2012). Representative examples of augmented images are illustrated in Fig. 6.

Testing: During evaluation, rice disease images were fed into the trained E2ETCA framework, and the class with the highest predicted probability was selected as the output. The predicted labels were compared against the ground truth to assess model performance. We utilized the evaluation metrics listed in Table 2, accuracy, precision, recall, and F1-score, to quantitatively evaluate the effectiveness of the proposed framework.

4.3. Implementation details

All experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU with 24 GB of memory, operating under a Linux environment. The implementation was carried out using Python 3.10 and PyTorch 1.10.0.

We also used ImageNet (https://www.image-net.org/download.php) for pre-training our model. The dimension *D* of the last layer was 1,024. The remaining hyperparameters and their values are summarized in Table 3.

4.4. Comparison with SOTA methods

Evaluation of the Mendeley dataset The accuracy, precision, recall, and F1-score on the Mendeley dataset are presented in Table 4, demonstrating the superiority of the proposed E2ETCA framework over all SOTA approaches. Compared to an Inception layer-based model (Hassan and Maji 2022), a CNN with an attention-based model (Peng *et al.* 2023), and several other CNN-based methods (Sharma R *et al.* 2022; Singh *et al.* 2023; Sudhesh *et al.* 2023), the E2ETCA framework achieved superior performance. The accuracy, precision, recall, and F1-score of our method were all 100.0%, indicating that E2ETCA achieves SOTA performance on this dataset.

Evaluation of the Kaggle dataset The accuracy, precision, recall, and F1-score on the Kaggle dataset are reported in Table 5. When compared to existing CNN-based approaches (Kathiresan *et al.* 2021; David *et al.* 2022; Van Ho *et al.* 2022; Agustin *et al.* 2023) and CNN with attention-based models (Wang *et al.* 2021; Al-Gaashani *et al.* 2023), the proposed E2ETCA framework achieved the highest performance. Specifically, its accuracy, precision, recall, and F1-score reached 99.2%, representing improvements of 0.5, 0.2, 0.4, and 0.3 percentage points, respectively, over the best results from prior methods.

Evaluation of the BRRI dataset The accuracy, precision, recall, and F1-score on the BRRI dataset are shown in Table 6. Consistent with the results on the Mendeley and Kaggle datasets, the E2ETCA framework outperformed existing approaches, including SqueezeNet V1.1, MobileNet V2, NASNet, VGG-16, ResNet, CNN, and VGG-19 (Rahman *et al.* 2020; Gautam *et al.* 2022). The E2ETCA framework achieved accuracy, precision, recall, and F1-scores of 99.5, 99.1, 98.9, and 98.9%, respectively, surpassing the best-performing baseline methods by 2.4, 3.2, 2.9, and 2.1 percentage points in each metric.

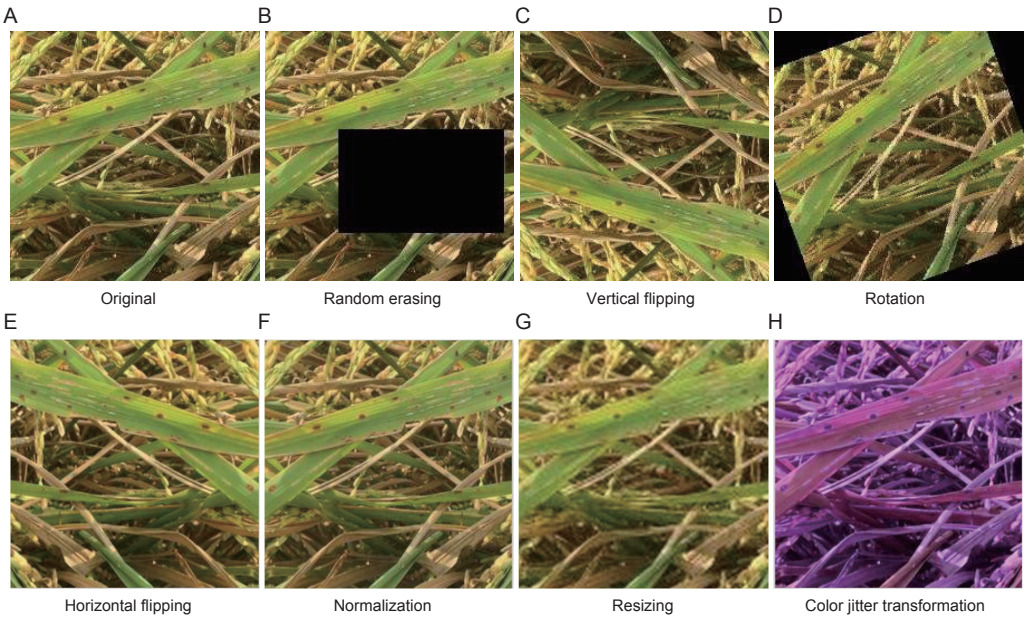


Fig. 6 Examples of augmented images, together with the original.

Table 3 Implementation hyperparameters with values used during training and testing

Parameter	Value
Image size	224×224
Number of epochs	300
Optimizer	Adam
Learning rate	10^{-3}
Batch size	16
Weight decay	10^{-4}

Table 4 Comparison of the proposed E2ETCA framework with existing methods applied to the Mendeley dataset (%)[1)]

Author (year)	Method	Accuracy	Precision	Recall	F1-score
Hassan and Maji (2022)	Inception	99.7	99.7	99.7	99.7
Sharma R *et al.* (2022)	CNN	99.6	99.6	99.6	99.7
Peng *et al.* (2023)	Attention+ DenseNet	97.9	98.0	97.9	97.9
Singh *et al.* (2023)	CNN	99.8	99.8	99.8	99.8
Sudhesh *et al.* (2023)	Xception	100.0	100.0	100.0	100.0
This study	Proposed E2ETCA	**100.0**	**100.0**	**100.0**	**100.0**

[1)] Bold values indicate the best benchmark.

5. Discussion

5.1. Overview

We propose an end-to-end ensemble framework (E2ETCA) that integrates multiple deep learning models: Inception V3, DenseNet-201, and ViT. To evaluate its effectiveness, we compare the E2ETCA framework with a traditional non-end-to-end machine learning approach in which features extracted by the same models are fed into an SVM, denoted as FCA-SVM. We conduct an in-depth analysis of the individual model performances under both end-to-end and non-end-to-end (FCA-SVM) settings. Additionally, we assess the performance of both frameworks on the combined dataset.

Table 5 Comparison of the proposed E2ETCA framework with existing methods applied to the Kaggle dataset (%)[1)]

Author (year)	Method	Accuracy	Precision	Recall	F1-score
Kathiresan *et al.* (2021)	DenseNet	66.5	68.2	65.4	66.8
Wang *et al.* (2021)	Attention+ MobileNet	94.7	92.6	87.4	89.6
David *et al.* (2022)	CNN	64.8	–	–	–
Van Ho *et al.* (2022)	ResNet+ DenseNet	–	–	–	96.0
Agustin *et al.* (2023)	Yolo V5	80.0	–	–	–
Al-Gaashani *et al.* (2023)	Attention+ ResNet	98.7	99.0	98.8	98.9
This study	Proposed E2ETCA	**99.2**	**99.2**	**99.2**	**99.2**

[1)] Bold values indicate the best benchmark.

Table 6 Comparison of the proposed E2ETCA framework with existing methods applied to the BRRI dataset (%)[1)]

Author (year)	Method	Accuracy	Precision	Recall	F1-score
Rahman *et al.* (2020)	SequeezeNet V1.1	92.5	–	–	–
	MobileNet V2	96.1	–	–	–
	NasNet	97.0	–	–	–
	VGG-16	97.1	–	–	–
Gautam *et al.* (2022)	VGG-19	90.0	90.0	90.0	91.0
	ResNet	91.0	90.0	90.0	91.0
	CNN	96.7	96.9	96.0	96.8
This study	Proposed E2ETCA	**99.5**	**99.1**	**98.9**	**98.9**

[1)] Bold values indicate the best benchmark.

5.2. Comparison of E2ETCA with FCA-SVM

To compare E2ETCA and FCA-SVM, we selected the BRRI dataset, which contains the largest number of classes (nine). The experimental protocol followed the procedure outlined in Section 4.1, with performance evaluated using 10-fold cross-validation. The accuracy, precision, recall, and F1-score on the BRRI dataset are provided in Table 7, the confusion matrix is illustrated in Fig. 7, and per-class performance comparisons are depicted in Fig. 8. The results show that the proposed end-to-end E2ETCA framework outperforms the non-end-to-end FCA-SVM framework across all disease classes and in terms of average performance metrics. Specifically, E2ETCA achieved average precision, recall, F1-score, and specificity of 99.1, 98.9, 98.9, and 99.9%, respectively, compared to 96.9, 95.6, 96.2, and 99.5% for FCA-SVM, representing improvements of 2.2, 3.3, 2.7, and 0.4 percentage points, respectively.

Regarding per-class performance, the E2ETCA framework achieved 100.0% precision, recall, F1-score, and specificity for all rice disease classes, with the exception of BB and HI. This discrepancy is likely due to class imbalance in the training and test sets. The BB and HI classes contain only 71 and 73 samples, respectively (Table 1). As a result, HI is frequently misclassified as BB or BH, as evidenced by the confusion matrix in Fig. 7.

5.3. Impact of individual models

We evaluated each base model included in our ensemble framework independently: Inception V3, DenseNet-201, and ViT. Each base model was trained and tested in an end-to-end manner, while the same features were also extracted and fed into an SVM classifier in a non-end-to-end approach. The BRRI dataset, which contains the largest number of classes, was selected for this evaluation. The resulting accuracy, precision, recall, and F1-score are presented in Table 8 and Fig. 9. The proposed E2ETCA framework achieved an accuracy of 99.1%, with corresponding precision, recall, and F1-scores of 98.8, 98.7, and 98.7%, respectively. These results represent improvements of 3.0, 2.6, 2.6, and 2.7% over the Inception V3 base model, and of 1.3, 0.9, 0.9, and 0.9% over ViT. A similar trend was observed when using the SVM classifier:

Table 7 Precision, recall, F1-score, and specificity of E2ETCA and FCA-SVM on BRRI dataset (%)[1)]

Class[2)]	E2ETCA				FCA-SVM			
	Precision (SD)	Recall (SD)	F1-score (SD)	Specificity (SD)	Precision (SD)	Recall (SD)	F1-score (SD)	Specificity (SD)
BB	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	97.6 (3.4)	98.8 (1.7)	100.0 (0.0)
BH	91.7 (5.9)	100.0 (0.0)	95.6 (3.1)	99.5 (0.4)	91.7 (5.9)	95.8 (5.9)	93.6 (5.1)	99.5 (0.4)
BS	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
FS	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	88.9 (0.0)	94.1 (0.0)	100.0 (0.0)
HE	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	90.5 (2.1)	97.1 (4.1)	93.7 (3.1)	98.0 (0.4)
HI	100.0 (0.0)	90.5 (6.7)	94.9 (3.6)	100.0 (0.0)	100.0 (0.0)	90.5 (6.7)	94.9 (3.6)	100.0 (0.0)
NB	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
SB	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	93.7 (4.5)	90.5 (0.0)	92.0 (2.1)	98.9 (0.8)
ST	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	96.8 (2.2)	100.0 (0.0)	98.4 (1.1)	99.4 (0.4)
Average	**99.1 (0.7)**	**98.9 (0.7)**	**98.9 (0.7)**	**99.9 (0.0)**	**96.9 (1.6)**	**95.6 (2.2)**	**96.2 (1.9)**	**99.5 (0.2)**

[1)] The standard deviation (SD) is shown in parentheses. Bold values indicate the average benchmarks.
[2)] BB, bacterial blight/bacterial leaf blight; BH, brown plant hopper; BS, brown spot; FS, false smut; HE, healthy; HI, hispa; NB, neck blast; SB, sheath (blight and/or rot); ST, stemborer.

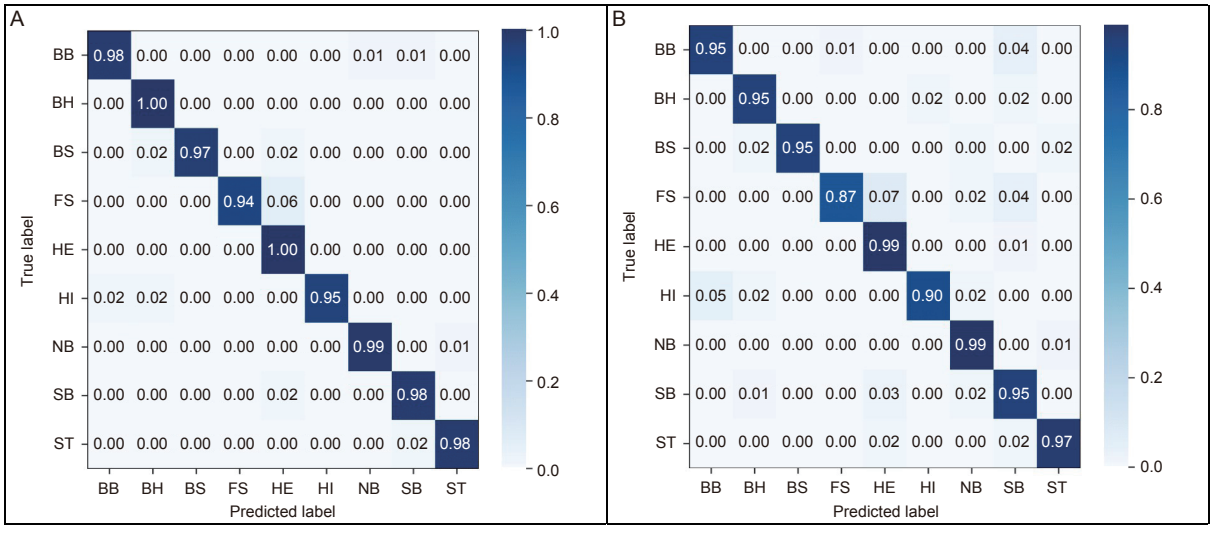


Fig. 7 Normalized confusion matrices of E2ETCA (A) and FCA-SVM (B) on the BRRI dataset. BB, bacterial leaf blight/bacterial blight; BH, brown plant hopper; BS, brown spot; FS, false smut; HE, healthy; HI, hispa; NB, neck blast; SB, sheath (blight and/or rot); ST, stemborer.

performance improved by 0.8, 0.7, 0.8, and 0.8% in accuracy, precision, recall, and F1-score, respectively, compared to ViT. Therefore, the ensemble framework outperformed each individual base model under both end-to-end and non-end-to-end configurations.

In terms of individual model performance, the attention-based ViT model surpassed the CNN-based models. Specifically, ViT achieved accuracy, precision, recall, and F1-score values that were 1.7, 1.7, 1.7, and 1.8% higher than those of Inception V3, and 0.7, 0.4, 0.7, and 0.7% higher than those of DenseNet-201. The ViT model leverages MSA to extract features both within patches (local intra-patch information) and across patches (global inter-patch relationships). This dual capability enables the model to capture fine-grained local details as well as long-range contextual dependencies, thereby surpassing the performance of conventional CNN-based architectures that rely primarily on local receptive fields.

5.4. Evaluation on the combined dataset

To evaluate the generalization and scalability of the proposed E2ETCA framework, we conducted experiments on a large-scale dataset created by combining Mendeley, Kaggle, and BRRI datasets. Table 9 presents the comparative evaluation results of the proposed E2ETCA framework and FCA-SVM, where feature representations extracted by the FCA are used as input to a SVM classifier. The corresponding confusion matrix is shown in Fig. 10. On the combined dataset, E2ETCA achieved an average precision of 90.7%, recall of 91.5%, F1-score of 90.8%, and specificity of 99.4%. In contrast, the FCA-SVM model attained 83.3, 76.6, 77.9, and 98.4% for these metrics. These results indicate that E2ETCA outperformed the non-end-to-end framework by margins of 7.4% in precision, 14.9% in recall, 12.9% in F1-score, and 1.0% in specificity.

With regard to the results for each class, E2ETCA achieved a better evaluation performance than FCA-SVM. E2ETCA yielded higher F1-scores than FCA-SVM for all rice disease classes, with relative improvements ranging from 3.1% for LB to 27.4% for BH (Table 9). Furthermore, the performance differences between the end-to-end E2ETCA and the non-end-to-end FCA-SVM are clearly visible for the large dataset.

For example, for the BS class, the F1-score difference between E2ETCA and FCA-SVM on the BRRI dataset alone was negligible (0.0%, Table 7). However, on the combined dataset, this difference increased to 7.5% (Table 9). This performance gain is attributed to the substantially larger number of samples available in the combined dataset, 2,111 images compared to just 111 in BRRI. The end-to-end E2ETCA framework perfectly trains the model with a large sample size and greater diversity

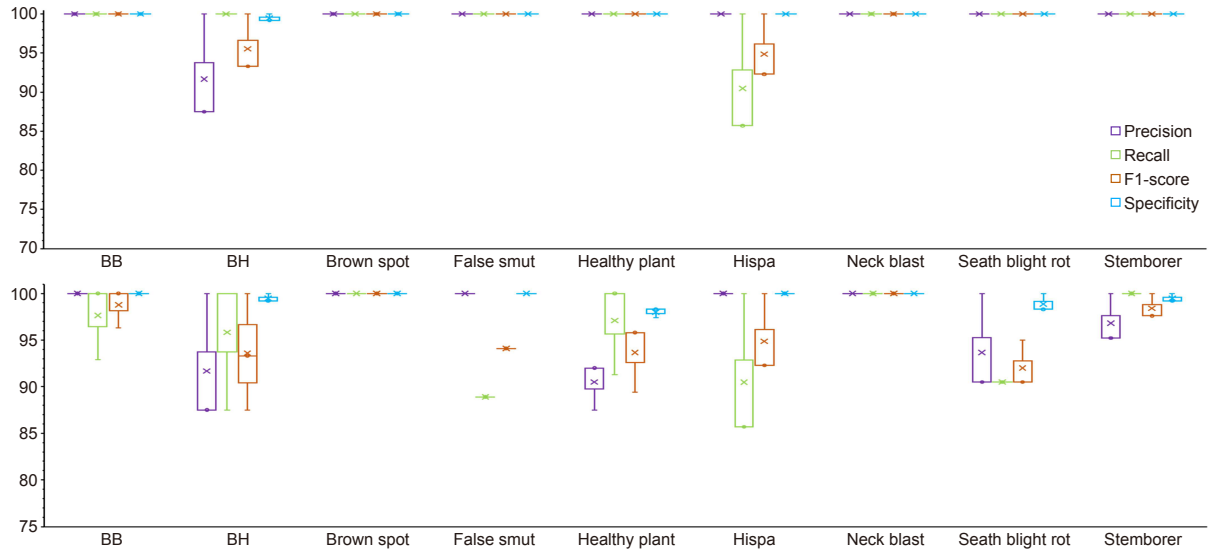


Fig. 8 Precision, recall, F1-score, and specificity of E2ETCA and FCA-SVM on BRRI dataset. BB, bacterial blight/bacterial leaf blight; BH, brown plant hopper.

Table 8 Performance of individual models: Inception V3, DenseNet-201, and ViT in an end-to-end manner and with the extracted features fed to an SVM[1)]

Model	End-to-end				Extracted features fed to SVM			
	Accuracy (SD)	Precision (SD)	Recall (SD)	F1-score (SD)	Accuracy (SD)	Precision (SD)	Recall (SD)	F1-score (SD)
Inception V3	96.1 (1.1)	96.2 (1.1)	96.1 (1.1)	96.0 (1.1)	90.8 (9.0)	92.4 (6.4)	90.7(9.0)	90.9 (8.7)
DenseNet-201	97.1 (1.0)	97.5 (0.5)	97.1 (1.0)	97.1 (1.0)	83.2 (5.0)	88.5 (2.3)	83.2 (5.0)	84.4 (3.8)
ViT	97.8 (0.8)	97.9 (0.8)	97.8 (0.8)	97.8 (0.8)	96.3 (0.3)	96.4 (0.4)	96.3(0.3)	96.3 (0.2)
Proposed (E2ETCA)	**99.1 (0.2)**	**98.8 (0.7)**	**98.7 (0.7)**	**98.7 (0.7)**	**97.1 (0.7)**	**97.1 (0.7)**	**97.1 (0.7)**	**97.1 (0.7)**

[1)] The standard deviation (SD) is shown in parentheses. Bold values indicate the average benchmarks.

since this class is represented across all three datasets.

6. Conclusion

This study presented an approach based on end-to-end training of a CNN and an attention-based ensemble framework (E2ETCA) for diagnosing rice diseases, which combines the strengths of the CNN-based approach of Inception V3 and DenseNet-201 with a vision transformer for the attention-based model. Features from each base model are fused through pointwise addition in a dedicated fusion layer, followed by a softmax classification layer for final prediction. The framework was trained and evaluated on 11 disease classes , including FS, BH, BB, NB, ST, HI, SB, BS, LB, TU, and HE. Various image preprocessing and argumentation techniques were applied to reduce the overfitting of the framework.

The model was comprehensively evaluated using various metrics, such as accuracy, precision, recall, F1-score, and specificity. Experimental results across three benchmark datasets, Mendeley, Kaggle, and BRRI, demonstrate that E2ETCA outperforms existing SOTA approaches for rice disease diagnosis on three publicly available datasets. Specifically, the proposed E2ETCA framework achieved absolute improvements in classification accuracy ranging from 1.2 to 3.0%. In addition, a comparative analysis with FCA-SVM, where extracted features are fed to an SVM, revealed that E2ETCA achieved superior performance both on a per-class level and in terms of overall averages for all disease classes. These improvements were especially significant on the large-scale dataset.

Future work should focus on constructing a much larger dataset that incorporates samples from various countries

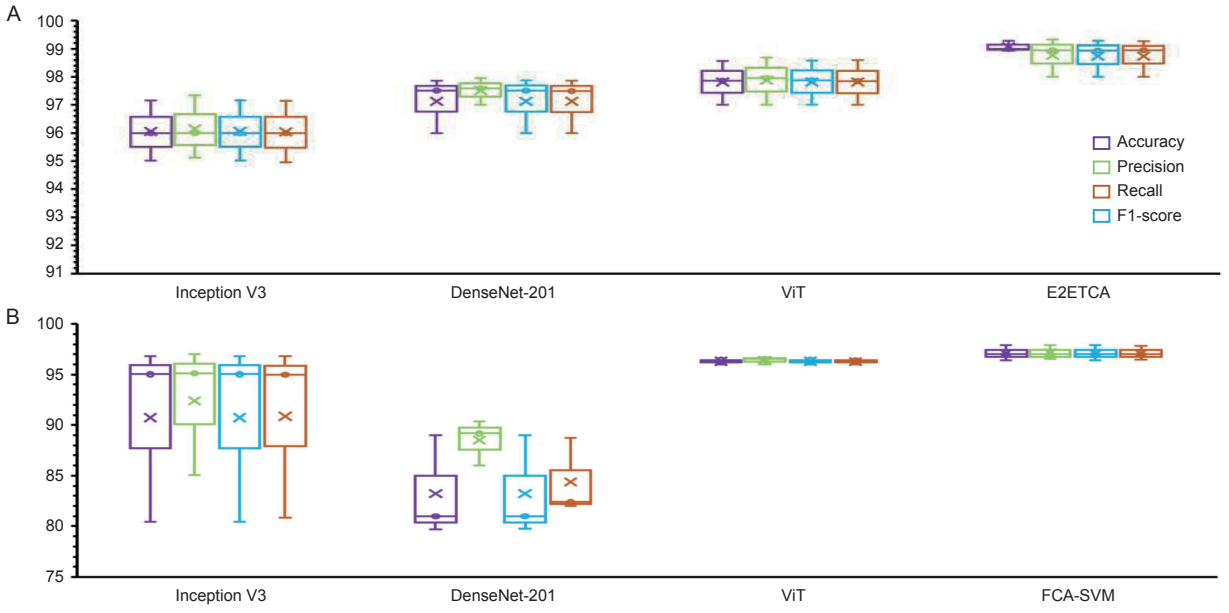


Fig. 9 Performance of individual models: Inception V3, DenseNet-201, and ViT. A, training and tests were performed in an end-to-end manner. B, extracted features from the corresponding network were fed to an SVM for evaluation.

Table 9 Precision, recall, F1-score, and specificity of E2ETCA and FCA-SVM on combined dataset (%)[1)]

Class[2)]	E2ETCA				FCA-SVM			
	Precision (SD)	Recall (SD)	F1-score (SD)	Specificity (SD)	Precision (SD)	Recall (SD)	F1-score (SD)	Specificity (SD)
BS	96.9 (1.0)	95.4 (1.7)	96.1 (0.8)	99.0 (0.3)	86.7 (11.9)	91.8 (4.1)	88.6 (6.5)	94.9 (5.3)
LB	95.2 (2.4)	92.9 (2.1)	94.0 (0.5)	98.7 (0.7)	93.9 (2.2)	88.2 (4.0)	90.9 (1.9)	98.5 (0.6)
BH	91.8 (10.8)	100.0 (0.0)	95.4 (6.2)	99.9 (0.1)	85.4 (11.3)	67.9 (33.7)	68.0 (26.2)	99.9 (0.1)
BB	97.2 (1.1)	99.1 (0.8)	98.2 (0.6)	99.3 (0.3)	88.1 (16.2)	91.7 (3.6)	88.9 (9.3)	95.9 (6.1)
TU	100.0 (0.0)	99.8 (0.3)	99.9 (0.2)	100.0 (0.0)	93.8 (10.8)	85.2 (14.8)	87.8 (7.2)	98.6 (2.5)
HE	80.1 (3.8)	83.1 (5.1)	81.3 (0.8)	98.4 (0.4)	74.6 (3.9)	73.8 (17.0)	73.3 (10.7)	98.1 (0.5)
HI	73.7 (9.2)	70.5 (13.0)	70.4 (4.1)	98.4 (0.8)	68.8 (6.1)	67.3 (11.8)	66.9 (4.4)	98.2 (0.9)
NB	94.8 (4.3)	99.6 (0.8)	97.1 (2.3)	99.8 (0.2)	77.5 (26.4)	73.7 (33.2)	74.9 (30.4)	99.5 (0.4)
SB	89.6 (8.0)	88.4 (4.9)	88.7 (4.8)	99.7 (0.2)	85.3 (9.7)	72.1 (20.6)	76.9 (16.7)	99.7 (0.2)
ST	96.3 (1.1)	95.6 (4.8)	95.9 (2.2)	99.9 (0.0)	84.5 (22.0)	78.1 (20.3)	80.9 (20.5)	99.6 (0.5)
FS	82.4 (7.5)	81.9 (6.1)	82.0 (5.7)	99.8 (0.1)	78.1 (18.5)	52.8 (29.5)	60.0 (30.6)	99.9 (0.1)
Average	**90.7 (4.5)**	**91.5 (3.6)**	**90.8 (2.6)**	**99.4 (0.3)**	**83.3 (12.6)**	**76.6 (17.5)**	**77.9 (14.9)**	**98.4 (1.6)**

[1)] The standard deviation (SD) is shown in parentheses. Bold values indicate the average benchmarks.
[2)] BS, brown spot; LB, leaf blast/blast; BH, brown plant hopper; BB, bacterial blight/bacterial leaf blight; TU, tungro; HE, healthy; HI, hispa; NB, neck blast; SB, sheath (blight and/or rot); ST, stemborer; FS, false smut.

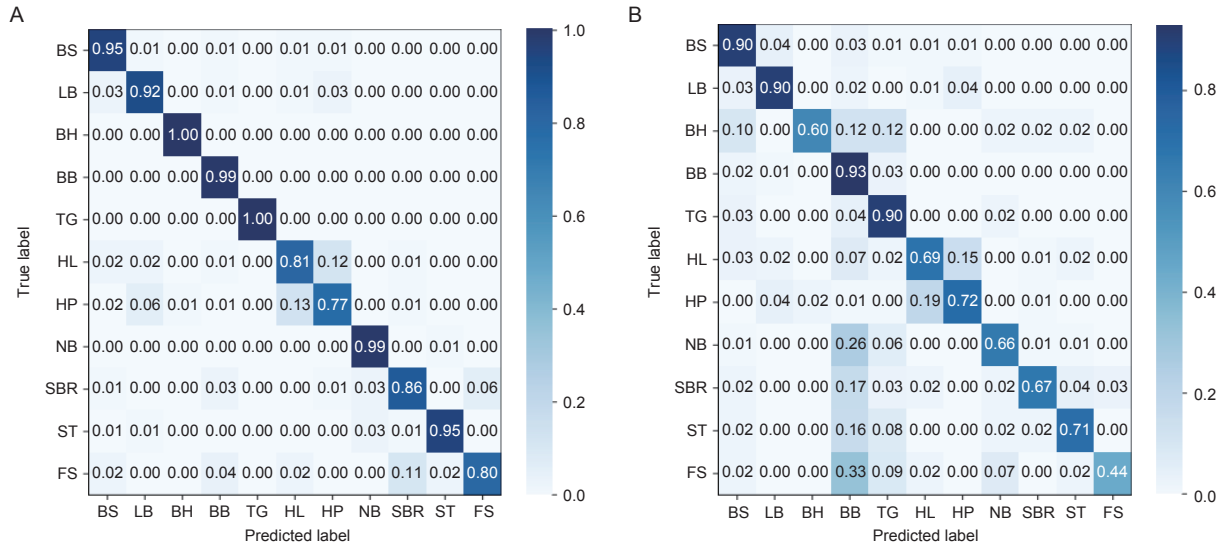


Fig. 10 Normalized confusion matrices for the combined dataset. A, E2ETCA. B, FCA-SVM. BS, brown spot; LB, leaf blast/blast; BH, brown plant hopper; BB, bacterial blight/bacterial leaf blight; TU, tungro; HE, healthy; HI, hispa; NB, neck blast; SB, sheath (blight and/or rot); ST, stemborer; FS, false smut.

and then apply existing approaches to demonstrate their performances. Additionally, optimizing the E2ETCA framework for deployment on edge devices or smartphones will be critical to enabling practical applications by farmers, especially in resource-limited regions. The proposed framework offers significant potential not only for rice disease diagnosis but also as a foundational model for disease detection in other crops.

Acknowledgements

We are grateful to the Begum Rokeya University, Rangpur, and the United Arab Emirates University, UAE for partially supporting this work.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors declare that they did not use AI in the preparation and writing of this manuscrlpt.

References

Agrios G. 2005. *Plant Pathology*. 5th ed. Elsevier Academic Press, Burlington, Ma. USA. pp. 79–103.

Agustin M, Hermawan I, Arnaldy D, Muharram A T, Warsuta B. 2023. Design of livestream video system and classification of rice disease. *JOIV: International Journal on Informatics Visualization*, **7**, 139–145.

Ahad M T, Li Y, Song B, Bhuiyan T. 2023. Comparison of cnn-based deep learning architectures for rice diseases classification. *Artificial Intelligence in Agriculture*, **9**, 22–35.

Ahmed K, Shahid T R, Alam S M I, Momen S. 2019. Rice leaf disease detection using machine learning techniques. In: *2019 International Conference on Sustainable Technologies for Industry 4.0* (STI). IEEE. pp. 1–5.

Al-Gaashani M S, Samee N A, Alnashwan R, Khayyat M, Muthanna M S A. 2023. Using a resnet50 with a kernel attention mechanism for rice disease diagnosis. *Life*, **13**, 1277.

Andrianto H, Faizal A, Armandika F. 2020. Smartphone application for deep learning-based rice plant disease detection. In: *2020 International Conference on Information Technology Systems and Innovation* (ICITSI). IEEE. pp. 387–392.

Anwar S M, Majid M, Qayyum A, Awais M, Alnowami M, Khan M K. 2018. Medical image analysis using convolutional neural networks: A review. *Journal of Medical Systems*, **42**, 1–13.

Bejani M M, Ghatee M. 2021. A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, **54**, 6391–6438.

Bhartiya V P, Janghel R R, Rathore Y K. 2022. Rice leaf disease prediction using machine learning. In: 2022 *Second International Conference on Power, Control and Computing Technologies* (ICPC2T). IEEE. pp. 1–5.

Borhani Y, Khoramdel J, Najafi E. 2022. A deep learning based approach for automated plant disease classification using vision transformer. *Scientific Reports* **12**, 11554.

Caruana R, Niculescu-Mizil A, Crew G, Ksikes A. 2004. Ensemble selection from libraries of models. In: *Proceedings of the Twenty-First International Conference on Machine Learning*. Association for Computing Machinery. p. 18.

David T, Alfred R, Obit J H, Fui F S, Gobilik J, Iswandono Z, Haviluddin H. 2022. Optimization of convolutional neural network in paddy disease detection. In: *International Conference on Computational Science and Technology*. pp. 399–412.

Deng R, Tao M, Xing H, Yang X, Liu C, Liao K, Qi L. 2021. Automatic diagnosis of rice diseases using deep learning. *Frontiers in Plant Science*, **12**, 701038.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929.

Durmus H, Gu¨nes E O, Kırcı M. 2017. Disease detection on the leaves of the tomato plants by using deep learning. In: 2017 *6th International Conference on Agro-Geoinformatics*. IEEE. pp. 1–5.

Gautam V, Trivedi N K, Singh A, Mohamed H G, Noya I D, Kaur P, Goyal N. 2022. A transfer learning-based artificial intelligence model for leaf disease assessment. *Sustainability*, **14**, 13610.

Ghosal S, Sarkar K. 2020. Rice leaf diseases classification using cnn with transfer learning. In: *2020 IEEE Calcutta Conference* (CALCON). IEEE. pp. 230–236.

Hassan S M, Maji A K. 2022. Plant disease identification using a novel convolutional neural network. *IEEE Access*, **10**, 5390–5401.

Van Ho S, Vuong H G, Nguyen B Q, Trinh Q H, Tran M T. 2022. Ensemble of deep neural networks for rice leaf disease classification. In: 2022 *RIVF International Conference on Computing and Communication Technologies* (RIVF). IEEE. pp. 238–243.

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

Islam T, Sah M, Baral S, Choudhury R R. 2018. A faster technique on rice disease detectionusing image processing of affected area in agro-field. In: *2018 Second International Conference on Inventive Communication and Computational Technologies* (ICI-CCT). IEEE. pp. 62–66.

Jiang F, Lu Y, Chen Y, Cai D, L G. 2020. Image recognition of four rice leaf diseases based on deep learning and support vector machine. *Computers and Electronics in Agriculture*, **179**, 105824.

Joshi A A, Jadhav B D. 2016. Monitoring and controlling rice diseases using Image processing techniques. In: 2016 *International Conference on Computing*, *Analytics and Security Trends* (CAST). IEEE. pp. 471–476.

Kathiresan G, Anirudh M, Nagharjun M, Karthik R. 2021. Disease detection in rice leaves using transfer learning techniques. *Journal of Physics: Conference Series*, **1911**, 012004.

Kishore K K, Kannan E. 2022. Detection of rice plant disease using adaboost svm classifier. *Agronomy Journal*, **114**, 2213–2229.

Krizhevsky A, Sutskever I, Hinton G E. 2012. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, **25**, 1097–1105.

Liao F B, Feng X Q, Li Z Q, Wang D Y, Xu C M, Chu G, Ma H Y, Yao Q, Chen S. 2023. A hybrid CNN-lstm model for diagnosing rice nutrient levels at the rice panicle initiation stage. *Journal of Integrative Agriculture*, **23**, 711–723.

Lu Y, Yi S, Zeng N, Liu Y, Zhang Y. 2017. Identification of rice diseases using deep convolutional neural networks. *Neurocomput ing*, **267**, 378–384.

Mohammed A, Kora R. 2023. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University — Computer and Information Sciences*, **35**, 757–774.

Nalini S, Krishnaraj N, Jayasankar T, Vinothkumar K, Britto A S F, Subramaniam K, Bharatiraja C. 2021. Paddy leaf disease detection using an optimized deep neural network. *Computers*, *Materials & Continua*, **68**, 1117–1128.

Opitz D, Maclin R, 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, **11**, 169–198.

Paymode A S, Malode V B. 2022. Transfer learning for multi-crop leaf disease image classification using convolutional neural network vgg. *Artificial Intelligence in Agriculture*, **6**, 23–33.

Peng J, Wang Y, Jiang P, Zhang R, Chen H. 2023. Ricedranet: Precise identification of rice leaf diseases with complex back grounds using a res-attention mechanism. *Applied Sciences*, **13**, 4928.

Peng S, Tang Q, Zou Y. 2009. Current status and challenges of rice production in china. *Plant Production Science*, **12**, 3–8.

Phadikar S, Sil J, Das A K. 2012. Classification of rice leaf diseases based on morphological changes. *International Journal of Information and Electronics Engineering*, **2**, 460–463.

Putra O V, Ningrum N T, Puspitasari N S, Wibowo A T, Rachmawaty E. 2022. Hit-lidia: A framework for rice leaf disease classification using ensemble and hierarchical transfer learning. *Lontar Komputer*: *Jurnal Ilmiah Teknologi Informasi*, **13**, 196.

Rahman C R, Arko P S, Ali M E, Khan M A I, Apon S H, Nowrin F, Wasif A. 2020. Identification and recognition of rice diseases and pests using convolutional neural networks. *Biosystems Engineering*, **194**, 112–120.

Ramesh S, Vydeki D. 2020. Recognition and classification of paddy leaf diseases using optimized deep neural network with jaya algorithm. *Information Processing in Agriculture*, **7**, 249–260.

Sanghvi H A, Patel R H, Agarwal A, Gupta S, Sawhney V, Pandya A S. 2023. A deep learning approach for classification of covid and pneumonia using densenet–201. *International Journal of Imaging Systems and Technology*, **33**, 18–38.

Serbetci A, Akgul Y S. 2020. End-to-end training of cnn ensembles for person re-identification. *Pattern Recognition*, **104**, 107319.

Sethy P K, Barpanda N K, Rath A K, Behera S K. 2020. Image processing techniques for diagnosing rice plant disease: A survey. *Procedia Computer Science*, **167**, 516–530.

Sharma M, Kumar C J, Deka A. 2022a. Early diagnosis of rice plant disease using machine learning techniques. *Archives of Phytopathology and Plant Protection*, **55**, 259–283.

Sharma M, Nath K, Sharma R K, Kumar C J, Chaudhary A. 2022b. Ensemble averaging of transfer learning models for identification of nutritional deficiency in rice plant. *Electronics*, **11**, 148.

Sharma R, Singh A, Kavita, Jhanjhi N Z, Masud M, Jaha E S, Verma S. 2022. Plant disease diagnosis and image classification using deep learning. *Computers*, *Materials & Continua*, **71**, 2125–2140.

Simhadri C G, Kondaveeti H K. 2023. Automatic recognition of rice leaf diseases using transfer learning. *Agronomy*, **13**, 961.

Singh S P, Pritamdas, K, Devi K J, Devi S D. 2023. Custom convolutional neural network for detection and classification of rice plant diseases. *Procedia Computer Science*, **218**, 2026–2040.

Stephen A, Punitha A, Chandrasekar A. 2023. Optimal deep generative adversarial network and convolutional neural network for rice leaf disease prediction. *The Visual Computer*, **40**, 919–936.

Strange R N, Scott P R. 2005. Plant disease: A threat to global food security. *Annual Review of Phytopathology*, **43**, 83–116.

Sudhesh K, Sowmya V, Kurian S, Sikha O. 2023. Ai based rice leaf disease identification enhanced by dynamic mode decomposition. *Engineering Applications of Artificial Intelligence*, **120**, 105836.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. 2015. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. pp. 1–9.

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. 2016. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. pp. 2818–2826.

Talukder M S H, Sarkar A K. 2023. Nutrients deficiency diagnosis of rice crop by weighted average ensemble learning. *Smart Agricultural Technology*, **4**, 100155.

Temniranrat P, Kiratiratanapruk K, Kitvimonrat A, Sinthupinyo W, Patarapuwadol S. 2021. A system for automatic rice disease detection from rice paddy images serviced via a chatbot. *Computers and Electronics in Agriculture*, **185**, 106156.

Turkoglu M, Yanikog˘lu B, Hanbay D. 2022. Plantdiseasenet: Convolutional neural network ensemble for plant disease and pest detection. *Signal*, *Image and Video Processing*, **16**, 301–309.

Udayananda G, Shyalika C, Kumara P. 2022. Rice plant disease diagnosing using machine learning techniques: A comprehensive review. *SN Applied Sciences*, **4**, 311.

Uddin M Z, Muramatsu D, Takemura N, Ahad M A R, Yagi Y. 2019. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. *IPSJ Transactions on Computer Vision and Applications*, **11**, 1–18.

Uddin M Z, Ngo T T, Makihara Y, Takemura N, Li X, Muramatsu D, Yagi Y. 2018. The ou-isir large population gait database with real-life carried object and its performance evaluation. *IPSJ Transactions on Computer Vision and Applications*, **10**, 1–11.

Wang Y, Wang H, Peng Z. 2021. Rice diseases detection and classification using attention based neural network and bayesian optimization. *Expert Systems with Applications*, **178**, 114770.

Zeng N, Gong G, Zhou G, Hu C. 2023. An accurate classification of rice diseases based on icai-v4. *Plants*, **12**, 2225.

Zhang J, Fukuda T, Yabuki N. 2021. Automatic object removal with obstructed fac¸ades completion using semantic segmentation and generative adversarial inpainting. *IEEE Access*, **9**, 117486–117495.

Zhang Y, Zhong L, Ding Y, Yu H, Zhai Z. 2023. Resvit-rice: A deep learning model combining residual module and transformer encoder for accurate detection of rice diseases. *Agriculture*, **13**, 1264.

Executive Editor-in-Chief Fanghao Wan
Managing Editor Juan Zhang