E2ETCA: End-to-end training of CNN and attention ensembles for rice diseasediagnosis[1]

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

[1]*Department of Computer Science and Engineering, Begum Rokeya University, Rangpur, Bangladesh*

[2]*Department of Computer Science and Software Engineering, United Arab Emirates University, UAE*

[3]*Department of Computer Science and Digital Technologies, University of East London, London, UK*

Abstract

Rice is one of the most important crops worldwide. Diseases of the rice plant can drastically reduce crop yield and even lead to complete loss of production. Early diagnosis can reduce the severity and help efforts to establish effective treatment and reduce the usage of pesticides. Traditional machine learning approaches have already been employed for automatic diagnosis. However, they heavily rely on manual preprocessing of images and handcrafted features, which is challenging, time-consuming, and may require domain expertise. Recently, a single end-to-end deep learning (DL)-based approach was employed to diagnose rice diseases. However, it is not highly robust, nor is it generalizable to every dataset. Hence, we propose a novel end-to-end training of convolutional neural network (CNN) and attention (E2ETCA) ensemble framework that fuses the features of two CNN-based state-of-the-art (SOTA) models along with those of an attention-based vision transformer model. These fused features are utilized for diagnosis by the addition of an extra fully connected layer with softmax. The whole procedure is performed end-to-end, which is very important for real-world applications. Additionally, we feed the extracted features into a traditional machine learning approach support vector machine for classification and further analysis. To verify the effectiveness of our proposed E2ETCA framework, we demonstrate it on three publicly available datasets: the Mendeley Rice Leaf Disease Image Samples dataset,

[1] #Correspondence Md. Zasim Uddin, E-mail: zasim@brur.ac.bd; Fady Alnajjar, E-mail:fady.alnajjar@uaeu.ac.ae

the Kaggle Rice Diseases Image dataset, the Bangladesh Rice Research Institute dataset, and a combination of these three datasets. On the basis of various evaluation metrics (accuracy, precision, recall, and F1-score), our proposed E2ETCA framework exhibits superior performance to existing SOTA approaches for rice disease diagnosis, which can also be generalizable in similar other domains.

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

Increasing population leads to an increasing demand for food. More than 800 million people have been found to lack access to enough food, and there are 1.3 billion people who receive less than one USD per day and cannot afford to pay for the food that they need (Strange and Scott, 2005). Plant diseases have severe adverse effects on crop production, leading to a negative impact on the economy and hindering efforts to meet the demand for food. According to the Food and Agriculture Organization (FAO), agricultural diseases and pests cost USD 220 billion annually and reduce crop yield by 20%–40% (Agrios, 2005).

Rice is one of the most important crops, with consumption that increases day by day (Shahbandeh, 2023). However, rice plant decrease rice production by 10%–15% (Peng et al., 2009). In extreme circumstances, they may reduce yield by 40%–50% or even lead to a complete loss of production (Jiang et al., 2020). These diseases can drastically lower productivity and quality, and so their prompt identification and control are crucial for good production (Deng et al., 2021).

Early diagnosis of rice diseases is essential to mitigate their severity and recover their previous production rate. Manual diagnosis is frequently employed to identify a disease based on how it manifests itself, but this requires human observation. For a large rice field, the cost, time, and effort involved in this procedure are high, and expertise is needed to perform the activities involved (Sethy et al., 2020; Zeng et al., 2023). Farmers frequently lack precise timing and knowledge of the diseases affecting their plants when they spray them to eradicate insects, other pests, and diseases (Andrianto et al., 2020). Moreover, excessive spraying pollutes the environment (Durmus et al., 2017). The only way to adequately diagnose rice diseases is by continuously monitoring the crop, which is only possible if an automated monitoring and diagnosis process can be developed (Sharma et al., 2022a).

Recent automated rice disease diagnosis methods can be categorized into two types: traditional machine learning (ML)-based approaches and modern deep learning (DL)-based approaches. Traditional ML-based techniques are widely used in rice disease diagnosis (Udayananda et al., 2022),

since they require less time and effort than manual diagnosis. For example, Islam et al. (2018) employed a naive Bayes method to classify rice disease on the basis of the red, green, and blue (RGB) values of the affected parts of a plant. Ahmed et al. (2019) removed the background and the affected area and then used *K*-nearest neighbor (K-NN), decision tree, naive Bayes, and logistic regression methods for rice disease classification. Although the currently available traditional ML-based methods give encouraging results, they suffer from a lack of generalizability, as well as depending on handcrafted features.

By contrast, DL-based approaches show promise as a basis for the development of more advanced methods of rice disease diagnosis and classification. For example, Lu et al. (2017) implemented a convolutional neural network (CNN)-based model for classifying 10 common rice diseases. Ghosal and Sarkar (2020) employed a pre-trained CNN-based VGG-16 model to classify plants suffering from three diseases along with healthy ones. Sharma et al. (2022a) investigated the use of six pre-trained DL-based models, namely, MobileNet, VGG-16, Inception V3, ResNet-50, Xception, and InceptionResNet V2, which were pre-trained on the largest ImageNet dataset. They considered three rice diseases, namely bacterial blight, rice blast, and brown spot, and compared the diagnostic results obtained using different DL-based approaches. We can see that the majority of DL-based methods rely on a single end-to-end pre-trained model, and because such models are prone to overfitting, this hinders the adaptability and generalizability of these methods to other datasets (Bejani and Ghatee, 2021).

Recently, ensemble learning has been adopted owing to its superior performance (Talukder and Sarkar, 2023). This is a learning process that combines an array of baseline models into a deeper composite model that is more efficient than its components. Additionally, the diversity of baseline models can help to reduce the risk of overfitting (Mohammed and Kora, 2023). Ensemble models have already been employed for rice disease diagnosis. For example, Ahad et al. (2023) used an ensemble model with DenseNet-121, EfficientNetB7, and Xception to compare individual CNN models. They evaluated performance on a publicly available dataset of nine classes of rice diseases. Sharma et al. (2022b) used an ensemble model with InceptionResNetV2, Xception, DenseNet-201, and VGG-19 CNN-based methods to generate binary, ternary, or quaternary ensemble classifiers and averaged the prediction probability. Deng et al. (2021) implemented an ensemble model including DenseNet-121, SE-ResNet-50, and ResNeSt-50, and they took the averaged output scores of submodels to obtain the final scores for prediction. To classify plant diseases, Turkoglu et al. (2022) proposed an ensemble model based on six state-of-the-art (SOTA) CNN-based models that provided improved accuracy through the use of a majority voting technique for model selection. However, these methods all adopted weighted average techniques or majority voting for the final prediction score, and the procedure as a whole was not end-to-end.

By contrast, end-to-end ensemble models incorporate data processing, feature extraction, and final prediction in a unified framework and are able to handle the sort of complex problems (Serbetci and

Akgul, 2020; Caruana et al., 2004) that arise in real-world applications. In this study, we propose an end-to-end ensemble framework (E2ETCA) for rice disease diagnosis that includes two CNN-based models and an attention-based vision transformer model. The features extracted from the considered models are consolidated to produce a final feature in the last layer with softmax for diagnosis. We evaluate the proposed E2ETCA framework on three publicly available rice disease datasets and a combination of these datasets. The results of this evaluation demonstrate that the E2ETCA framework achieves state-of-the-art performance. After extracting the features using the framework, we employ a traditional machine learning-based approach based on a support vector machine (SVM) for further comprehensive analysis in a non-end-to-end manner. Furthermore, we evaluate each component model of the E2ETCA framework separately and compare their performance to gain insights into how well each model performs independently in an end-to-end way as well as with the SVM in a non-end-to-end way.

2. Related work

2.1. *Traditional machine learning-based approaches*

Traditional ML-based approaches to rice disease diagnosis have been employed to automate the detection and classification of these diseases. Such approaches often require the extraction of handcrafted features from rice images before diagnosis can be performed. For example, Phadikar et al. (2012) considered morphological changes and the radial distribution of hue from the center of the boundary of the images as a feature for classification. They achieved accuracies of 79.5% and 68.1% using a naive Bayes and an SVM classifier, respectively, for a dataset with 1000 sample images of two rice diseases. Islam et al. (2018) introduced an approach to expedite the detection and classification of rice diseases by low-level image processing of affected areas, such as the localization and RGB values of the affected portion. Finally, they used traditional naive Bayes to classify three rice diseases.

By contrast, Joshi and Jadhav (2016) employed the minimum distance classifier (MDC) and the K-NN classifier to classify four rice diseases, taking the shape and color of a diseased rice leaf portion as features. They evaluated their approach on a dataset with 115 leaf sample images and obtained accuracies of 87.0% and 89.2% with MDC and K-NN, respectively. Kumar K and E (2022) took color, shape, and texture as the main ingredients for feature extraction and applied the discrete wavelet transform to identify affected areas for feature extraction and to remove the typical green portion of leaves. They employed an adaptive boosting SVM and achieved a maximum accuracy of 98.8% for the classification of three rice diseases. Bhartiya et al. (2022) used shape characteristics, including area, roundness, and area-to-lesion ratio, to distinguish between various rice disease types. Using a quadratic SVM classifier, they achieved an accuracy of 81.8%.

Although the traditional ML-based methods remain useful for rice disease diagnosis, they depend heavily on handcrafted feature extraction, which can be labor-intensive and requires domain expertise. Moreover, the use of handcrafted features to represent the complex visual characteristics of rice diseases can be challenging and may not capture all relevant information in the images, and it is therefore difficult to adopt this approach for real-world applications.

2.2. *Deep learning-based approaches*

Convolutional neural networks (CNNs) have emerged as a powerful and highly effective approach for preprocessing, feature extraction, and classification in an end-to-end manner in many fields of computer vision, such as biometrics (Uddin et al., 2018, 2019), medical imaging (Anwar et al., 2018), and person re-identification (Serbetci and Akgul, 2020). Such DL-based approaches have also been explored for rice disease detection and diagnosis (Liao et al., 2023; Jiang et al., 2020; Ramesh and Vydeki, 2020; Nalini et al., 2021). For example, Jiang et al. (2020) used a CNN along with an SVM to classify rice diseases, while Ramesh and Vydeki (2020) and Nalini et al. (2021) utilized *K*-means clustering algorithm for disease diagnosis. A variety of end-to-end pre-trained models (transfer learning) have also been employed in rice disease detection and diagnosis. For example, Temniranrat et al. (2021) compared four models, namely, Faster R-CNN, Retina-Net, YOLO V3, and mask R-CNN, for rice disease classification, and they demonstrated that YOLO V3 achieved the highest accuracy of 79.2% on a dataset of 8767 images for six classes. Studied in (Simhadri and Kondaveeti, 2023; Gautam et al., 2022) explored six different pre-trained CNN-based models, for example, VGG-16, VGG-19, ResNet, Inception, Mobilenet, and SqueezeNet. They find out that Inception works better than other models. By contrast, Sudhesh et al. (2023) implemented 10 transfer-learned CNN-based models along with pre-processed raw images using a dynamic mode decomposition (DMD)-based attention-driven mechanism. Besides transfer learning, generative adversarial networks (GANs) have also been explored. For example, Stephen et al. (2023) utilized three-dimensional and two-dimensional CNNs to extract features, and for classification, they employed a GAN based on an improved backtracking search method.

Vision transformer (ViT) models (Dosovitskiy et al., 2020) are attention-based models that demonstrate the power of attention mechanisms in capturing long-range dependencies within images, dividing an image into fixed-size, non-overlapping patches, which are then linearly embedded into a sequence of vectors. ViT models are widely used for image classification and rice disease diagnosis and classification. For example, Borhani et al. (2022) employed a ViT-based approach using image patches. They evaluated their model on a dataset of 120 images for three disease classes and achieved an accuracy of 91.7%. ViT models can also be combined with CNNs. For example, to combine the advantages of both a ViT and ResNet, Zhang et al. (2023) proposed a model called

ResViT-Rice, in which a ResViT-Rice block was embedded in ResNet, using a self-attention mechanism. They evaluated their model on a publicly available dataset consisting of 1548 images of two rice disease classes and a healthy class, for which they achieved an accuracy of 99.1%.

Ensemble models combine several independent models into a single unified framework, thereby improving generalization performance over the individual components (Opitz and Maclin, 1999) and also reducing overfitting (Serbetci and Akgul, 2020). Recently, ensemble learning has gained popularity in rice leaf disease detection and diagnosis. For example, Ahad et al. (2023) implemented an ensemble model consisting of DenseNet-121, EfficientNet B7, and Xception. They achieved an accuracy of 97.6% for a dataset of 900 images of nine classes, which represents an increase in accuracy of 17.0% compared with a single model (Seresnext101). Putra et al. (2022) implemented two ensemble models, one a combination of MobileNet and DenseNet and the other a combination of DenseNet and XceptionNet, through feature fusion, and they achieved over 40.0% greater accuracy compared with the individual models for the first ensemble model, and 60.0% greater accuracy for the second, for a dataset of three diseases with 300 images. Deng et al. (2021) employed an ensemble model consisting of DenseNet-121, SE-ResNet-50, and ResNet-50, using an averaging technique for final prediction. They achieved an accuracy of 91.0% using a dataset of 33026 images of six diseases. Turkoglu et al. (2022) proposed an ensemble model combining six state-of-the-art CNN-based models to classify plant diseases, and they were able to achieve an improved accuracy when they evaluated the performances of different combinations of base models with an SVM.

However, all these methods used weighted average techniques or majority voting for the final prediction score, and the procedure as a whole was not end-to-end. By contrast, an end-to-end ensemble model would achieve better general performance with reduced overfitting training and test time. Furthermore, such a model could be applied in real time, which is necessary for rice disease detection and diagnosis. In this paper, we propose an end-to-end ensemble framework (E2ETCA) that fuses the features from two CNN-based models along with an attention-based model.

3. Proposed framework

3.1. Overview of framework

Traditional ensemble methods usually involve training the component models separately, either in parallel or sequentially, which is not feasible for deep learning because of computational cost. Furthermore, it is necessary to select the best models using voting. Therefore, instead of training multiple methods from scratch, we propose a end-to-end training of convolutional neural network (CNN) and attention (E2ETCA) ensemble framework of multiple base learners in a unified framework, including two of which are CNN-based and the other is an attention-based ViT. An overview of the proposed framework is presented in Fig. 1. Our framework is based on Inception V3 (Szegedy et al.,

2016) and DenseNet-201 (Huang et al., 2017), which are two well-known CNN-based models, and ViT-L/32, a patch-based ViT attention network (Dosovitskiy et al., 2020). Each model flattens the input feature maps and performs pointwise addition, with the output then being fed to the fully connected embedding layer. Finally, a one-layer classification network with softmax is employed to calculate the 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 a function $h(x)$: $\mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{D}$ 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:

$$F_{\text{final}} = F_{\text{i}} \oplus F_{\text{d}} \oplus F_{\text{v}}, \qquad (1)$$

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 $f(h(x))$: $\mathbb{R}^{D} \rightarrow \mathbb{R}^{K}$.

Finally, the cross-entropy loss L_{CEL} is employed to measure the classification error:

$$L_{CEL} = -\sum_{k=1}^{K} y_i^k \log\left(f_i^k\right), \qquad (2)$$

where f_i^k and y_i^k denote the estimated and actual probabilities of the *i*th image with *k*th rice disease class.

3.2 CNN-based model

Two state-of-the-art CNN-based models, namely, Inception V3 (Szegedy et al., 2016) and DenseNet-201 (Hang et al., 2017), are employed.

Inception V3 (Szegedy et al., 2016) consists of three different Inception modules, denoted as Block_A, Block_B, and Block_C in Fig. 2. Block_A is employed for spatial aggregation by replacing a large filter with two or more small filters. It also uses convolutions of unit size to reduce the number of channels and computational complexity while retaining vital information and having the ability to learn various characteristics. Block_B adds additional convolutions to those of Block_A. Finally, Block_C includes factorization as well as convolutions. This structure increases the output of the filter bank, enabling the use of high-dimensional representations. Further details can be found in Szegedy et al. (2015, 2016).

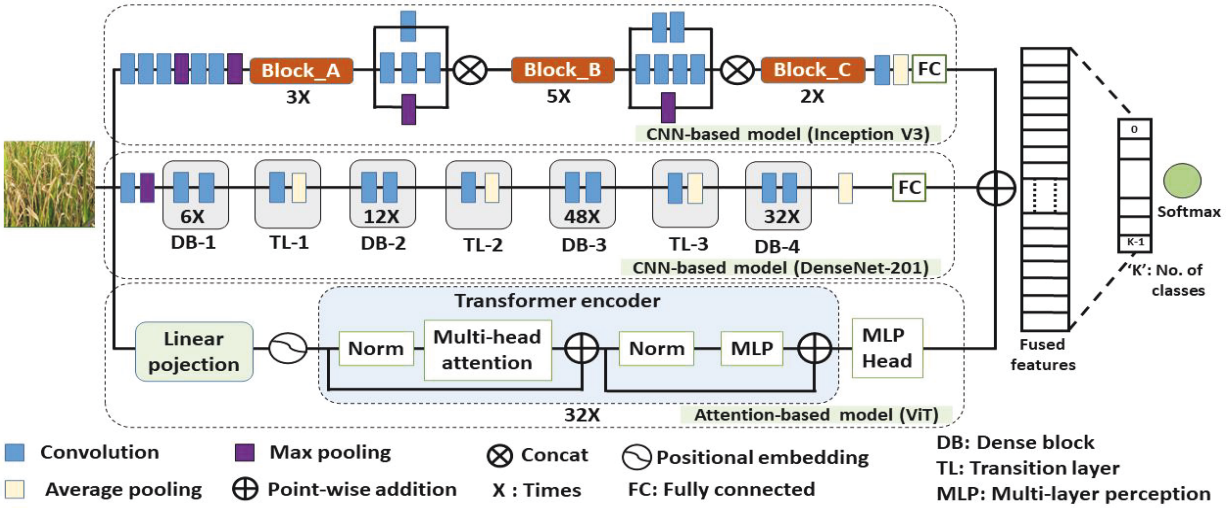


Fig. 1: Overview of the proposed E2ETCA framework consisting of three modules: two CNN-based models and an attention-based model. The extracted features are consolidated and employed to classify rice diseases with classification layers, including an FC layer along with softmax. The overall training and evaluation are performed end-to-end.

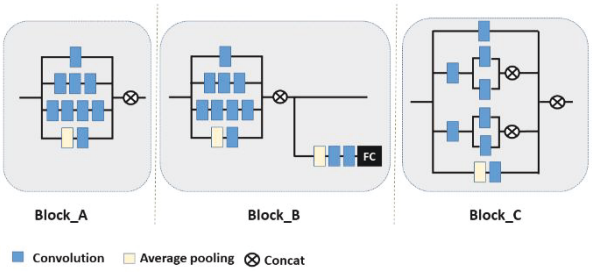


Fig.2: The three blocks of Inception V3, namely, Block_A, Block_B, and Block_C, where each block consists of different sizes of convolution layers with routes. All routes are finally concatenated to get the final features.

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

$$\mathrm{In}_{i-1} = h_{hybrid}(x). \qquad (3)$$

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

$$\mathrm{In}_i = h_{\mathrm{block}}(\mathrm{In}_{i-1}). \quad (4)$$

After several hybrid functions, Inception blocks, and concatenation of convolution and pooling via different routes, the feature vector passes to the fully connected layer to produce the final features as follows:

$$F_i = FC(\mathrm{In}_{n-1}), \quad (5)$$

where In_{n-1} is the feature map generated from the $(n-1)$th layer and FC(·) converts the feature map of dimensions $P_{n-1} \times Q_{n-1}$ into a D-dimensional feature vector F_i that is employed to represent the final features for fusion in Eq. (1).

DenseNet-201 (Huang et al., 2017) is a version of DenseNet that is efficient in terms of both computation and memory consumption (Sanghvi et al., 2023), and in which 201 layers of the architecture are used. 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 $x \in \mathbb{R}^{H \times W \times C} \rightarrow E_1 \in \mathbb{R}^{P_1 \times Q_1}$.

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

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

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 each dense block, there is a transition layer involving convolution of files of size 1×1 and an average pooling of size 2×2 to decrease the spatial dimensions of the feature maps. After several dense blocks and transition layers, the feature vector undergoes an average pooling and then passes through a fully connected layer to extract 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) capture both global and local contexts by treating images as sequences of fixed-size patches. After that, it adds positional embeddings to each patch to form a sequence that is then passed to the transformer encoder. An input image $x \in \mathbb{R}^{H \times W \times C}$ is

reshaped to equally shaped sequences $x_p \in \mathbb{R}^{N \times p^2 C}$, where (*p, p*) is the (height, width) of each image patch, and *N* is the number of patches, which serves as the effective input sequence length for the transformer:

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

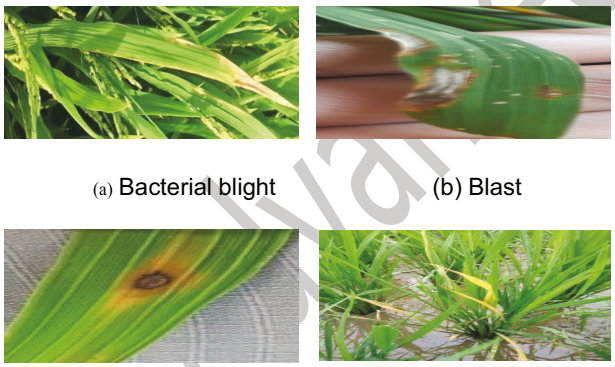
Then, each patch is flattened and mapped to a T_d-dimensional latent vector with a trainable linear projection $\alpha \in \mathbb{R}^{p^2 \times T_\text{d}}$, which used by the transformer for all layers. This is referred to as patch embedding. Then, patch embeddings are integrated with positional embeddings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

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

An encoder is made up of several alternating layers of multiheaded self-attention (MSA) and multilayer perceptron (MLP) blocks. In the encoder, the input feature is passed through MSA block, followed by the MLP block. A layer norm and residual a connection are added before and after each block, respectively. The operations of an encoder block can be expressed as

$$\beta_i^{'} = MSA\left(L_{\text{norm}}\left(\beta_{i-1}^{'}\right)\right) + \beta_{i-1}^{'}, \quad i = 1, \ldots, n, \quad (9)$$

$$\beta_i = MSA\left(L_{\text{norm}}\left(\beta_i^{'}\right)\right) + \beta_i^{'}, \quad i = 1, \ldots, n, \quad (10)$$



(a) Bacterial blight (b) Blast

(c) Brown spot (d) Tungro

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

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

$$F_{\mathrm{v}} = MLP_{\mathrm{head}}\left(L_{\mathrm{norm}}(\beta_n^0)\right). \quad (11)$$

$\mathrm{MLP}_{\mathrm{head}}(\cdot)$ consists of a single FC to produce the final feature map F_{v} with dimension D for further use in Eq. (1).

4. Experiments

4.1. Datasets and evaluation protocols

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

The *Mendeley Rice Leaf Disease Image Samples dataset* (Sethy et al., 2020)[1] (henceforth referred to as the Mendeley dataset) consists of a total of 5932 images of four diseases, namely, bacterial blight, blast, brown spot, and tungro. They were collected from the rice fields of Sambalpur and Bargarh Districts, Odisha, India, using a Nikon DSLR-D5600 camera with an 18–55 mm high-resolution lens. The dataset details are given in Table 1, and examples of images from each class are shown in Fig. 3. A variety of protocols are available in the literature. For a fair comparison, we strictly followed a popular protocol used in previous studies (Hassan and Maji, 2022; Sharma et al., 2022c; Singh et al., 2023; Sudhesh et al., 2023), that is, we used fivefold cross-validation to validate the proposed E2ETCA framework.

The *Kaggle Rice Diseases Image dataset*[2] (henceforth referred to as the Kaggle dataset) is a publicly

[1] https://data.mendeley.com/datasets/fwcj7stb8r/1

[2] https://www.kaggle.com/datasets/minhhuy2810/rice-diseases-i mage-dataset

Table 1: Distribution of sample images for each rice disease class for the Mendeley, Kaggle, and BRRI datasets. Bacterial blight and bacterial leaf blight are considered together, as are leaf blast and blast.[a]

Class	BH	BB	BS	FS	HI	NB	ST	HE	LB	TU	SB	Total
Mendeley	—	1584	1600	—	—	—	—	—	1440	1308	—	**5932**
Kaggle	—	—	400	—	400	—	—	400	400	—	—	**1600**

BRRI	71	138	111	93	73	286	201	234	—	—	219	**1426**
Combined	71	1722	2111	93	473	286	201	634	1840	1308	219	**8958**

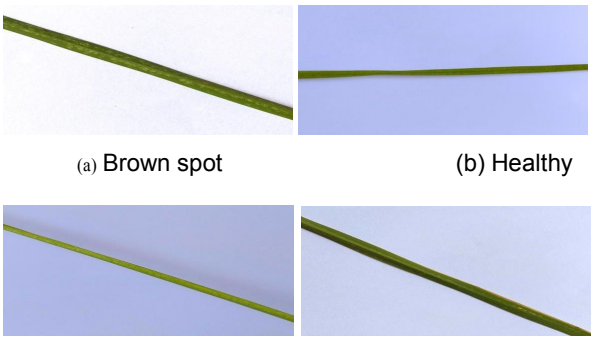
[a] BH, brown plant hopper; BB, bacterial blight/bacterial leaf blight; BS, brown spot; FS, false smut; HI, hispa; NB, neckblast; ST, stemborer; HE, healthy; LB, leaf blast/blast; TU, tungro; SB, sheath (blight and/or rot).

available dataset collected from rice farmland and consisting of 1600 images of three diseases, namely, brownspot, hispa, and leaf blast, and healthy plants. The dataset details are given in Table 1, and examples of images from each class are shown in Fig. 4. We used fivefold cross-validation to validate the proposed framework for a fair comparison with the SOTA approach (Wang et al., 2021; David et al., 2022; Al-Gaashani et al., 2023).

The *Bangladesh Rice Research Institute dataset* (Rahman et al., 2020)[3] (henceforth referred to as the BRRI dataset) consists of 1426 images with nine classes and was collected from paddy fields of the Bangladesh Rice Research Institute. The images were gathered from heterogeneous backgrounds over a period of seven months under different conditions, such as winter, summer, and overcast, to get, as far as possible, a fully representative set of images. To increase the robustness of the model, four different cameras were used. This dataset has a total of five classes of disease, three classes of pests, and a remaining class for healthy plants and others, with sheath blight and sheath rot being considered as the same class owing to their simultaneous occurrence, similar locations, and similar treatment methods. The dataset details are given in Table 1, and examples of images from each class are shown in Fig. 5. For a fair comparison with existing methods (Rahman et al., 2020; Gautam et al., 2022), we employed all nine classes, and the experiment was performed with 10-fold cross-validation, with the average evaluation result being used for each experiment.

To validate the proposed E2ETCA framework on a large-scale dataset with many classes, we generated a *combined dataset* by combining the Mendeley, Kaggle, and BRRI datasets. It consisted of a total of 11 classes, namely, false smut, brown plant hopper, bacterial leaf blight, neck blast, stemborer, hispa, sheath blight and/or sheath rot, brown spot, blast, tungro, and healthy plant, comprising 8958 sample images. It consisted of a total of 11 classes, namely, false smut, brown plant hopper, bacterial leafblight, neck blast, stemborer, hispa, sheath blight and/or sheath rot,

[3] https://drive.google.com/drive/folders/1ewBesJcguriVTX8sRJseCDbXAF_T4akK?usp=drive_open



(a) Brown spot (b) Healthy

(c) Hispa (d) Leaf blast

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

brown spot, blast, tungro, and healthy plant, comprising 8958 sample images. It should be noted that bacterial blight and bacterial leaf blight are considered to be in the same class under the name "bacterial blight", and similarly, leaf blast and blast under the name "leaf blast". The numbers of samples of each class of the combined dataset are shown in Table 1. In each experiment, the combined dataset was randomly divided into two subsets, namely, a training set and a test set, in a ratio of 80:20, and all experiments were repeated five times to reduce the effects of randomness.

Evaluation protocol: Different evaluation criteria, including accuracy, precision, recall, and F1-score, were employed to validate the proposed E2ETCA framework. The evaluation metrics and their calculation formula are given in Table 2. Quantification was performed in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Here, TP and TN denote successfully predicted labels, and FP and FN denote incorrectly predicted labels. For example, the correct classification of a given class of disease is considered a TP. On the other hand, whenever a disease does not belong to a given class, a prediction that it is in that class is considered an FP. Whenever a disease belongs to a desired class but is predicted as another class, this is considered an FN.

(a) Bacterial blight (b) Brown plant hopper (c) Brown spot

(d) False smut (e) Stemborer (f) Hispa

(g) Neck blast (h) Sheath blight (i) Healthy

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	$\frac{\lvert TN \rvert}{TN + FP\rvert}$
Specificity	$\frac{\lvert TN \rvert}{\lvert TN + FP \rvert}$
F1-score	$\frac{2 \times Precision \times Recall}{Precision + Recall}$

4.2. *Training and testing*

Training: Rice disease images along with labels were fed to the proposed E2ETCA framework and an output label f_i^k (i.e., for the *i*th training example with the *k*th rice disease class) was obtained as

prediction. Then, the predicted label and true label were fed into a calculation of the cross-entropy loss according to Eq. (2), and the Adam optimizer was employed to fit the weight and bias value. Here, the training batch size was 16. DL-based models need a large number of samples to improve their performance and prevent overfitting (Paymode and Mal-ode, 2022). One solution for a small dataset is to augment samples artificially for training such that the model explores the insight of the dataset more precisely. For each sample, we augmented through resizing of the image, random erasing, normalization (Zhang et al., 2021), vertical flipping, horizontal flipping, rotation, and color jitter (Krizhevsky et al. 2012). Examples of augmentedsamples for a rice disease image are shown in Fig. 6.

Testing: Rice disease images were passed to the framework, generating a class with maximum probability as output. The output and true labels were then used to evaluate the performance. We employed the evaluation metrics given in Table 2, namely, accuracy, precision, recall, and F1-score, to measure the performance of the proposed E2ETCA framework.

4.3. Implementation details

We conducted all experiments on a single NVIDIA GeForce RTX 3090 GPU with 24 GB memory running on a Linux. We used Python 3.10 and PyTorch 1.10.0 for all our implementations.

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}

We also used ImageNet[4] for pre-training our model. The dimension D of the last layer was 1024. The other hyperparameters with their values are shown in Table 3

4.4. Comparison with SOTA methods

4.4.1. Evaluation of Mendeley dataset

The accuracy, precision, recall, and F1-score on the Mendeley dataset are presented in Table 4, which demonstrates the superiority of our proposed E2ETCA framework over all SOTA approaches. Compared with an Inception layer-based model (Hasan and Maji, 2022), an CNN with attention-based model (Peng et al., 2023) and several other CNN-based approaches (Sharma et al., 2022c; Singh et al., 2023; Sudhesh et al., 2023), our proposed E2ETCA framework exhibited better performance. The accuracy, precision, recall, and F1-score of our method were all 100.0%. This indicates that our proposed E2ETCA framework achieves SOTA performance.

4.4.2. Evaluation of Kaggle dataset

The accuracy, precision, recall, and F1-score on the Kaggle dataset are presented in Table 5. Compared with existing CNN-based approaches (David et al.,2022; Kathiresan et al., 2021; 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 achieves the best performance. For example, its accuracy, precision, recall, and F1-score were all 99.2%, representing improvements upon the best results among the existing approaches by 0.5%, 0.2%, 0.4%, and 0.3%, respectively.

4.4.3. Evaluation of BRRI dataset

The accuracy, precision, recall, and F1-score on the BRRI dataset are presented in Table 6. As with the Mendeley and Kaggle datasets, the performance of our

[4]https://www.image-net.org/download.php

proposed E2ETCA framework exceeded those of existing approaches, namely, SequeezeNet V1.1, MobileNet V2, NasNet, VGG-16, ResNet, CNN, and VGG-19 (Gautam et al., 2022; Rahman et al., 2020). The accuracy, precision, recall, and F1-scores of E2ETCA were 99.5%, 99.1%, 98.9%, and 98.9%, respectively, surpassing by 2.4%, 3.2%, 2.9%, and 2.1%, respectively, the best results among the other methods.

5. Discussion

5.1. *Overview*

We have proposed an end-to-end ensemble framework (E2ETCA) incorporating several models: Inception V3, DenseNet-201, and ViT. Here, we compare the performance of our end-to-end ensemble with a traditional non-end-to-end ML-based approach with SVM by feeding the features extracted

using the proposed framework. This method is denoted as FCA-SVM. Similarly, we analyze in depth the performance of each of the considered models in both the end-to-end approach and the non-end-to-end approach with SVM by feeding the extracted features from these models. Furthermore, we evaluate the performance on the combined dataset using the proposed E2ETCA framework along with FCA-SVM.

5.2. *Comparison of E2ETCA with FCA-SVM*

To compare the performance of E2ETCA with that of FCA-SVM, we selected the dataset with the highest number of classes, namely, the BRRI dataset, with nine classes. We followed the same protocol for BRRI as described in Sec. 4.1. Experiments were evaluated using 10-fold cross-validation. The accuracy, precision, recall, and F1-score on the BRRI dataset are presented in Table 7, the confusion matrix is shown in Fig. 7, and a comparison for each rice disease class is shown in Fig. 8. We can see that the proposed end-to-end E2ETCA framework performed better than the non-end-to-end FCA-SVM framework for each class and in terms of the average of all results. The average precision, recall F1-score, and specificity for all classes using the proposed E2ETCA framework were 99.1%, 98.9%, 98.9%, and 99.9%, respectively, compared with 96.9%, 95.6%, 96.2%, and 99.5%, respectively, for FCA-SVM. These represent improvements of 2.2%, 3.3%, 2.7%, and 0.4%, respectively, for E2ETCA over FCA-SVM.

Regarding the evaluation performance for each rice disease class, we can observe that our proposed framework gave precision, recall, F1-score, and specificity of 100.0% for all classes except for

(a) Original (b) Random erasing (c) Vertical flipping (d) Rotation

(e) Horizontal flipping transformation (f) Normalization (g) Resizing (h) Color jitter

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

Table 4: Comparison of the proposed E2ETCA framework with existing methods applied to the Mendeley dataset. Bold values indicate the best benchmark.

Author (year)	Method	Accuracy	Precision	Recall	F1-score
Hassan and Maji (2022)	Inception	99.7	99.7	99.7	99.7
Sharma et al. (2022c)	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**

Table 5: Comparison of the proposed E2ETCA framework with existing methods applied to the Kaggle dataset. Bold values indicate the best benchmark.

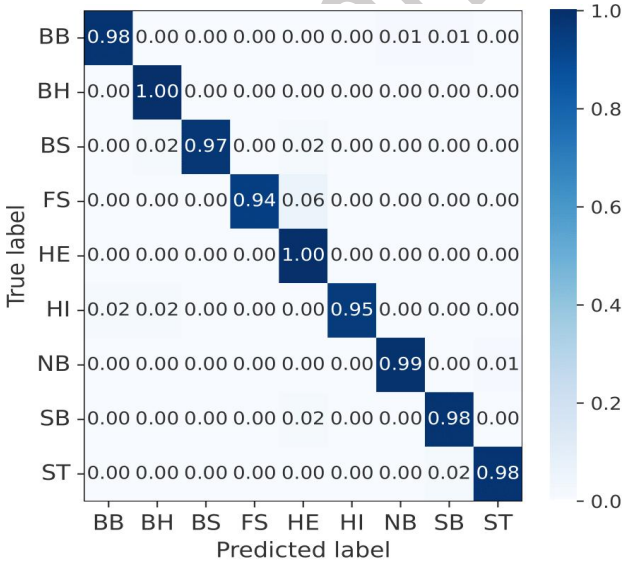
Author (year)	Method	Accuracy	Precision	Recall	F1-score
Kathiresan et al. (2021)	DenseNet	66.5	68.2	65.4	66.8
Wang et al. (2021)	Attention + MobileNet	94.7	92.6	87.4	89.6
David et al. (2022)	CNN	64.8	—	—	—
Van Ho et al. (2022)	ResNet + DenseNet	—	—	—	96.0
Agustin et al. (2023)	Yolo V5	80.0	—	—	—

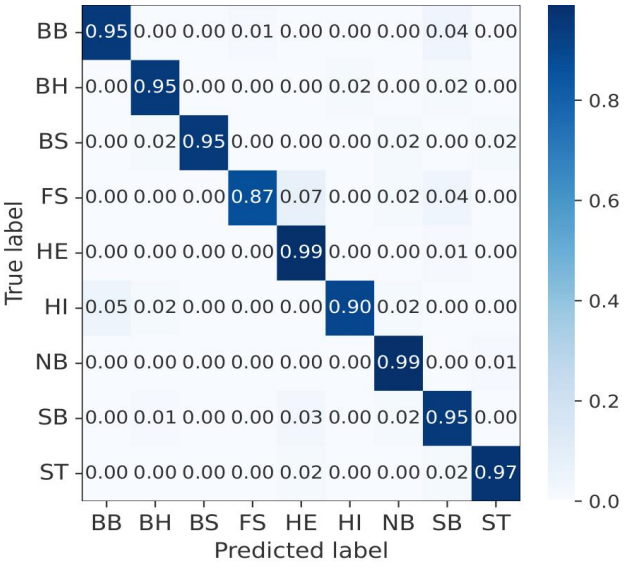
Al-Gaashani et al. (2023)	Attention + ResNet	98.7	99.0	98.8	98.9
This study	Proposed E2ETCA	**99.2**	**99.2**	**99.2**	**99.2**

bacterial blight/bacterial leaf blight (BB) and hispa. We believe that this discrepancy is due an imbalance in training and test samples for these two classes. The numbers of BB and HI disease class samples are 71 and 73 (see Table 1). Therefore, it is easy for hispa to be misclassified as bacterial blight/bacterial leaf blight (BB) and brown plant hopper (BH) (see the confusion matrix in Fig. 7).

Table 6: Comparison of the proposed E2ETCA framework with existing methods applied to the BRRI dataset. Bold values indicate the best benchmark.

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**





(a) E2ETCA (b) FCA-SVM

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

5.3. Impact of individual models

We evaluated each of the base models considered in our ensemble framework separately: Inception V3, DenseNet-201, and ViT. We trained and tested each base model end-to-end and employed the SVM with the extracted features from the corresponding base model in a non-end-to-end manner. We again selected the BRRI dataset, which has the highest number of classes, for the evaluation. The accuracy, precision, recall, and F1-score are shown in Table 8 and Fig. 9. We can see that the proposed E2ETCA framework gave an accuracy, precision, recall, and F1-score of 99.1%, 98.8%, 98.7%, and 98.7%, respectively, representing improvements of 3.0%, 2.6%, 2.6%, and 2.7%, respectively, compared with the Inception V3 base model and 1.3%, 0.9%, 0.9% and 0.9%, respectively compared with ViT. We can observe a similar tendency when the SVM was employed. For example, it improved accuracy, precision, recall, and F1-score by 0.8%, 0.7% 0.8%, and 0.8%, respectively, compared with ViT. Thus, our ensemble framework achieved better performance than each of the individual base models when applied in either an end-to-end or a non-end-to-end manner.

Regarding the evaluation performance for each base model, the attention-based ViT model was superior to the CNN-based models. For example, the accuracy, precision, recall, and F1-score achieved by ViT were better by 1.7%, 1.7%, 1.7%, and 1.8% than those from Inception V3 and better by 0.7%, 0.4%, 0.7%, and 0.7% than those from DenseNet-201. We know that the ViT model extracts feature

intra-patch-wise locally and inter-patch-wise globally, along with using a multi-head attention mechanism. Thus, it was able to extract fine-grained local details and global features and thereby exceed the evaluation performance of the CNN-based models.

Table 7: Precision, recall, F1-score, and specificity of E2ETCA and FCA-SVM on BRRI dataset. The standard deviation (SD) is shown in parentheses. Bold values indicate the average benchmarks.[a]

Class	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)**

[a] BB, bacterial blight/bacterial leaf blight; BH, brown plant hopper; BS, brown spot; FS, false smut; HE, healthy; HI, hispa; NB, neckblast; SB, sheath (blight and/or rot); ST, stemborer.

5.4. Evaluation of combined dataset

To validate the proposed E2ETCA framework on a large-scale dataset, we used the combined dataset generated from the Mendeley, Kaggle, and BRRI datasets. The evaluation results for

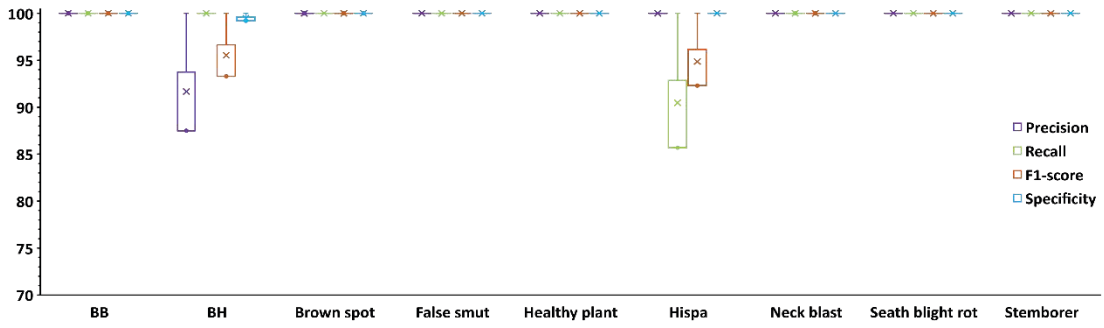
precision, recall, F1-score, and specificity of the proposed end-to-end E2ETCA framework and those obtained when the extracted features were fed to an SVM (FCA-SVM) are shown in Table 9, and the confusion matrix is shown in Fig. 10. We can observe that the average precision, recall, F1-score, and specificity for all rice disease classes from E2ETCA were 90.7%, 91.5%, 90.8%, and 99.4%, respectively, compared with 83.3%, 76.6%, 77.9%, and 98.4% from FCA-SVM. This demonstrates that our proposed end-to-end framework was superior in terms of precision, recall, F1-score, and specificity by 7.4%, 14.9%, 12.9%, and 1.0%, respectively, compared with the non-end-to-end framework.

With regard to the results for each class, E2ETCA achieved a better evaluation performance than FCA-SVM. We can see that the end-to-end ensemble framework E2ETCA gave results ranging from a minimum 3.1% F1-score for leaf blast (LB) to a maximum 27.4% for brown plant hopper (BH) (see 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.

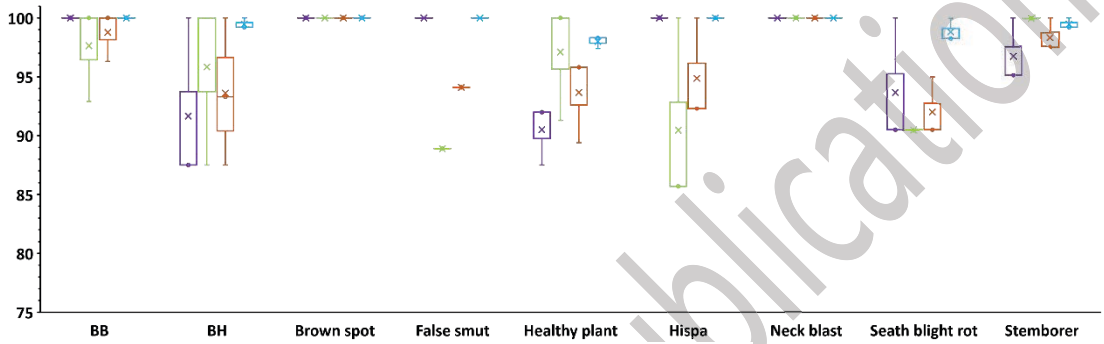
For example, the performance difference in F1-score for a brown spot on the BRRI dataset (see Table 7) was 0.0%, but 7.5% on the combined dataset (see Table 9). We believe that this is because there were only 111 sample images of the brown spot rice disease class on the BRRI dataset but 2111 on the combined dataset. The end-to-end E2ETCA framework perfectly trains the model with a large sample size and greater diversity because this brown spot disease class is available on all three datasets.

6. Conclusion

We have 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. The proposed framework considers the features extracted from each of the base models and merges them through pointwise addition in a separated layer along with a final classification layer with softmax. In total, 11 disease classes have been considered to evaluate the proposed framework: false smut, brown plant hopper, bacterial leaf blight, neck blast, stemborer, hispa, sheath blight or sheath rot, brown spot, blast, tungro, and healthy. We have also employed different image preprocessing and argumentation techniques to reduce the overfitting of the framework.



(a) E2ETCA



(b) FCA-SVM

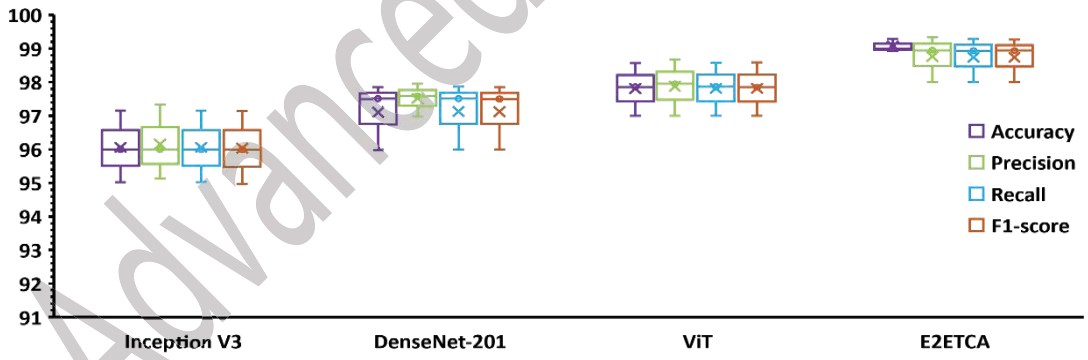
Fig. 8: Precision, recall, F1-score, and specificity of E2ETCA and FCA-SVM on BRRI dataset. BB, bacterial blight/bacterial leaf blight; BH, brown plant hopper.

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

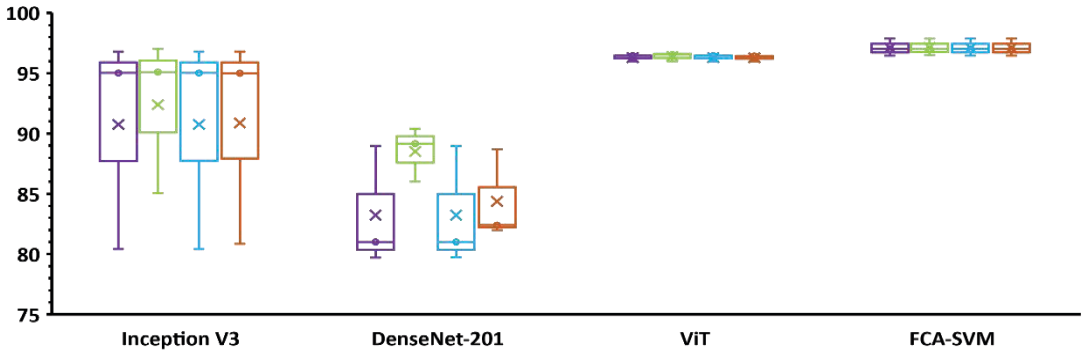
Model	End-to-end				Extracted features fed to SVM			
	Accuracy (SD)	Precision (SD)	Recall (SD)	F1-score (SD)	Accuracy (SD)	Precision (SD)	Recall (SD)	F1-score (SD)
Inception V3	96.1 (1.1)	96.2 (1.1)	96.1 (1.1)	96.0 (1.1)	90.8 (9.0)	92.4 (6.4)	90.7(9.0)	90.9 (8.7)
DenseNet-201	97.1 (1.0)	97.5 (0.5)	97.1 (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)**

We have evaluated the proposed framework using various metrics: accuracy, precision, recall, and F1-score. After analyzing the results, we can conclude that the proposed framework outperforms existing SOTA approaches for rice disease diagnosis on three publicly available datasets. Furthermore, we have added deep insight for each base model of our proposed framework, and the comparative analysis performed in this paper shows that the proposed E2ETCA framework improves the accuracy by 1.2%-3.0% for each base model for each dataset. In addition, we have performed a detailed analysis of how the proposed end-to-end framework differs when the extracted features are fed to an SVM for evaluation in a non-end-to-end manner. We have found that the proposed end-to-end framework achieves better accuracy both on each class as well as on average for all disease classes. Furthermore, the proposed framework works better for a large-scale dataset.

However, because of the great importance of this work, we need to develop a much larger dataset covering samples from various countries and then apply existing approaches to demonstrate their performances. We also need to ensure that these algorithms are applicable on edge devices or smartphones so that they can be used even by farmers in communities with a lack of access to other sophisticated technology. Our proposed method will have a high impact not only for rice disease diagnosis, but also for similar other crops' disease analysis in the long run.



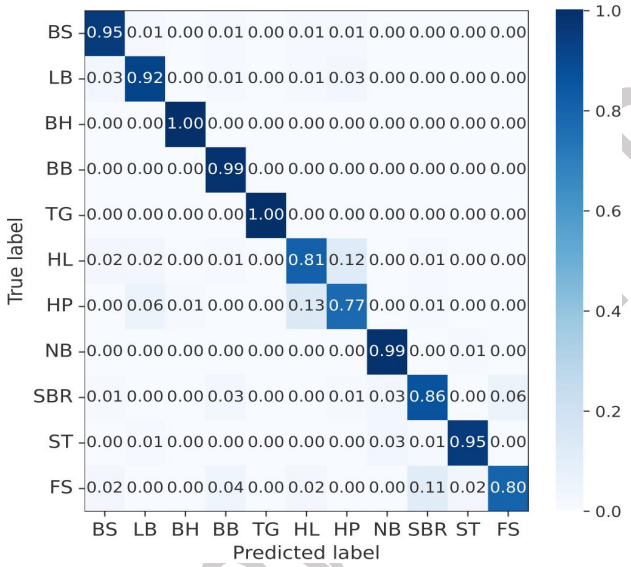
(a) End-to-end training and testing

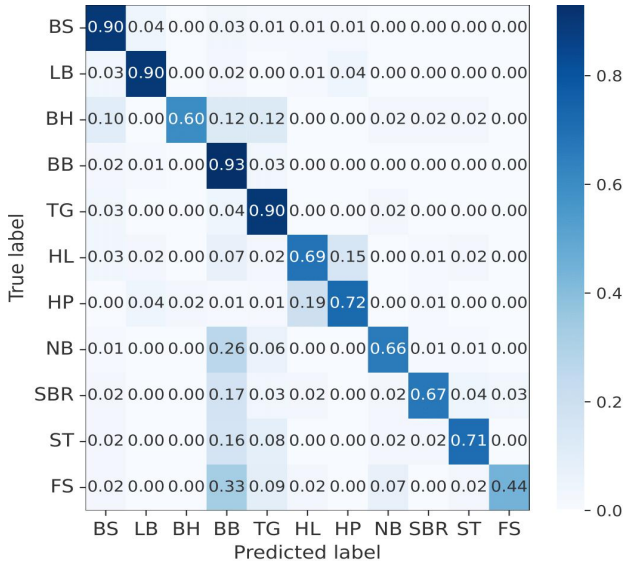


(b) Extracted features fed to SVM.

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.





(a) E2ETCA (b) FCA-SVM

Fig. 10: Normalized confusion matrices for the combined dataset. BB, bacterial blight/bacterial leaf blight; BH, brown plant hopper; 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).

Table 9: Precision, recall, F1-score, and specificity of E2ETCA and FCA-SVM on combined dataset. The standard deviation (SD) is shown inparentheses. Bold values indicate the average benchmarks.[a]

Class	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)**

[a] 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.

Acknowledgments

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

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this study.

References

Agrios, G., 2005. *Plant pathology 5th edition: Elsevier academic press*. *Burlington, Ma. USA* , 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., Shahidi, T.R., Alam, S.M.I., Momen, S., 2019. Rice leaf disease detection using machine learning techniques, in *2019International 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., et al., 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*, 1–48.

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*, 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., et al., 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.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: *Proceedings of theIEEE 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., & Li, 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)* (pp. 471-476). IEEE.

Kathiresan, G., Anirudh, M., Nagharjun, M., & Karthik, R. (2021, May). Disease detection in rice leaves using transfer learning techniques. In: *Journal of Physics: Conference Series* (Vol. **1911**, No. 1, p. 012004). IOP Publishing.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, **25**.

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

Liao, F.b., FENG, X.q., LI, Z.q., WANG, D.y., XU, C.m., Guang,
C., Qing, Y., Song, C., et al., 2023. A hybrid cnn-lstm model for diagnosing rice nutrient levels at the rice panicle initiation stage. *Journal of Integrative Agriculture*.

Lu, Y., Yi, S., Zeng, N., Liu, Y., Zhang, Y., 2017. Identification of rice diseases using deep convolutional neural networks. *Neurocomputing* **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.*

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(1)**,

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 backgrounds 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.

Shahbandeh, M., 2023. Global rice consumption 2022/23, by country.

Sharma, M., Kumar, C. J., & Deka, A. (2022). Early diagnosis of rice plant disease using machine learning techniques. *Archives of Phytopathology and Plant Protection*, **55(3)**, 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., Jhanjhi, N., Masud, M., Jaha, E.S., Verma, S., et al., 2022c. Plant disease diagnosis and image classification using deep learning. *Computers, Materials & Continua* **71**.

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,* 1–18.

Strange, R.N., Scott, P.R., 2005. Plant disease: a threat to global food security. *Annu. Rev. Phytopathol. 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*, 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*, 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 pestdetection. *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 ComputerVision 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.

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.

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 façades 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**, 126