Ingénierie des Systèmes d'Information
Vol. 30, No. 3, March, 2025, pp. 665-673
Journal homepage: http://iieta.org/journals/isi

An Automatic Rice Plant Disease Classification Using Hierarchical Vision Transformers with Spatial Reduction Attention Mechanism

Manisha Gnanavel*, Ezhumalai Periyathambi

Department of CSE, R.M.D Engineering College, Thiruvallur 601206, India

Corresponding Author Email: manisha.cse@rmd.ac.in

https://doi.org/10.18280/isi.300311

Received: 25 October 2024
Revised: 5 January 2025
Accepted: 14 March 2025
Available online: 31 March 2025

Keywords:
computer vision, multi-head attention, rice leaf disease, vision transformers

ABSTRACT

Rice is a staple crop for world food security, but its yield is susceptible to several diseases. Timely recognition of rice leaf diseases (RLD) is critical to decrease yield losses. Recent studies offered improved solutions by harnessing deep networks to precisely diagnose and categorize RLDs. Conversely, due to extreme scale deviations, data redundancy, and the high-resolution multi-spectral nature of leaf images, the conventional deep-learning classifiers underperform in detecting rice diseases. Also, training deep learning models involves major challenges such as class imbalance, overfitting, and vanishing gradient problems. Nowadays, vision transformers are infiltrated into the field of image processing, in which the self-attention unit is employed to learn local and distant correlations among the pixels in an image. However, the processing and storage overheads of analyzing image patches are very high. In this context, we propose a new vision transformer-based RLD classifier, called Faster Hierarchical Vision Transformer (FHViT) which employs a Spatial Reduction Attention Mechanism (SRAM) to speed up the classification process. The SRAM module enables the transformer to estimate the significance of each pixel in image patches and optimize their effect on the result. We evaluate our model on an open-access Ade F. Rice Leaf Diseases (AFRLD) database and relate its performance with other advanced models in terms of performance indicators. Our model delivers 99.2% detection accuracy, 99.3% precision, 99.2% sensitivity, 99.1% specificity, 99.0% recall, and 99.0% F1 measure. The extensive experimentations demonstrate that the FHViT realizes a viable solution for RLD diagnosis.

1. INTRODUCTION

The global population is anticipated to rise from 8.01 bn now (1st January 2024) to 8.5 bn in 2030, 9.7 bn in 2050, and 11.2 bn in 2100 [1]. By 2050, it is unavoidable to increase global food security by about 70% to fulfil the caloric needs of the world population. Rice is a staple plant for food security and a primary nutrient source, being the second-most produced crop globally and an essential food for over half of the inhabitants around the world [2, 3]. India is the second-largest rice producer (135.76 million metric tons (Mt)) in the world after China (145.95 Mt), contributing 24% of production worldwide [4]. Also, world rice production is projected to rise by 11.4%, reaching 567 Mt by 2030 [5].

According to the International Rice Research Institute (IRRI), RLDs can decline rice production by around 80%, causing a drop in agricultural revenues, reduced food supply, and more costs for buyers [6]. Detection and classification of RLDs are very important since they directly affect the growth, development, and production of plants [7]. This further leads to the utilization of detrimental chemicals and insecticides, which can have adverse effects on human well-being and the environment [8]. Hence, efficient disease management in rice crops is critical for guaranteeing sustainable yield and sufficient food supply. Identifying the rigorousness of the rice disease is regularly characterized by the level and spread of the disease over the leaves area [9]. Traditionally, the pathology of rice plants mainly depends on manual techniques by capturing and analyzing rice leaf images through dedicated tools. However, these techniques are laborious, critical, and unproductive since the rice field images contain multifaceted background information, such as weeds, soil, and redundant parts of rice plants. Furthermore, detecting this RLD manually does not deliver timely recognition of RLDs, which can result in considerable production losses [10]. Farmers, agricultural experts, or plant pathologists need to visually examine large areas of rice fields, which can be extremely laborious. Manual classification is intrinsically prone to errors. Skilled agricultural workers might misidentify or overlook diseases, particularly when signs are subtle or not easily distinguishable from other conditions. Subjectivity is another issue; different experts may understand the signs differently, leading to inconsistencies in diagnosis. Manual classification methods are not scalable and some diseases may go unnoticed by human inspectors.

Automatic and exact classification of RLDs is made feasible through Deep Learning (DL) algorithms. This causes an enhancement in crop yield and grade [11]. Convolutional Neural Networks (CNN) have demonstrated strong performance in recognizing and categorizing RLDs using

images captured from damaged leaves, stalks, and fruits of plants, providing early inferences to reduce production loss [12]. Related to manual methods, DL approaches fetch numerous advantages, including higher accuracy, speed, and the potential to process big data. Numerous research works prove this statement that DL approaches outstrip conventional techniques in detecting and diagnosing RLDs [13]. These approaches can efficiently identify RLDs at an initial stage, rendering them a more rapid and dependable method of identifying and classifying RLDs [10].

As neural networks evolve, their depth also increases. This leads to vanishing gradient problems and complex training errors. Besides, the performance of these networks mostly depends on large training databases. Vision transformers, with their outstanding capacity to extract significant attributes in images, have developed one of the most modern and prevailing networks that are being applied in the domain of vision-based applications [14]. Visual transformer is widely used for image processing tasks and it applies the self-attention mechanism to analyze images. In this research, we develop a novel RLD classification approach using a transformer, called FHViT which employs SRAM to speed up the processing speed. Besides, the SRAM module enables the transformer to estimate the significance of several pixels in input image patches and dynamically optimize their effect on the output. This article is arranged as mentioned below: In section II, we discuss some existing studies on RLD classification. In Section III, we explore the architecture and operation of basic transformers comprehensively. Section IV explores the implementation of our FHViT approach to identify and categorize RLDs from high-resolution pictures. Section V provides the empirical study and evaluation metrics employed in this work. To end, Section VI summarizes this work.

2. LITERATURE SURVEY

Timely identification of RLDs is of undue significance to ensuring food security and monitoring the spread of disease. A neural network-based approach classified RLDs by minimizing the model parameters [15]. Using a new database of 4199 rice plant photographs, the proposed model is trained to detect five different RLDs. The performance of this approach is assessed on a dataset with rice leaf images. The study [16] developed a MobileNetV2-based transfer learning framework. It is trained on the ImageNet database and uses an attention mechanism to increase the training efficiency of the feature engineering. A hybrid CNN and EfficientNet B7 deep architecture has been proposed for classifying four diseases in grape plants, such as leaf blight, black measles, and black rot [17]. This model employs a logistic regression method to down-sample the collected attributes. An ensemble approach with ResNet-50, DenseNet-121, and ResNeSt-50 has been developed in the study [18]. This model eliminates misperception among the diverse types of disease and reduces misclassification errors. By applying the concept of the ensemble approach, this approach identifies 6 different RLDs. The performance of four deep networks, including DenseNet-121, VGG-16, Inception-V4, and ResNet-50, has been assessed for detecting and classifying RLDs [10]. From a comprehensive experimental study, the authors prove that the DenseNet-121 outdoes other models regarding detection accuracy.

Leaf infection classification frameworks was introduced by integrating a support vector machine with CNNs employed to detect and classify particular rice crop infections like false smut, sheath rot, bacterial leaf blight, rice blast, and brown leaf spot [18]. Three CNN structures such as ResNet-18, ResNet-34, and ResNet-50 recognize and categorize normal and unhealthy leaves including hispa, brown spot, and leaf blast [19]. The AlexNet model categorize RLDs like bacterial leaf blight, brown leaf spot, and false smut [20]. Two deep networks, DenseNet-169 and Xception models, classify rice crop diseases [21]. A compact vision transformer, called MobileViT, was developed for RLD detection on mobile devices [22]. This model substitutes the convolutional layer in MobileViT with a flipped residual configuration that uses a 7×7 convolutional filter to capture global relationships among various pixels in rice field images efficiently. An encoding module excerpts diverse grades of attributes in a picture [23]. Likewise, to increase the capacity of the transformer encoding module to capture short-range statistics, inception modules have been introduced [24].

From this survey, we conclude that the transformer models with attention mechanisms have the potential to upturn the detection and classification accuracy of RLDs. The attention mechanism is used to define local and global relationships among the pixels within an image (local dependencies) and between patches (global dependencies). However, due to low resolution, high processing and storage overheads, and other reasons, it is not appropriate for dense pixel-level classification applications like RLD classification. To handle this issue, we develop a new transformer-based RLD classification model, called FHViT with a spatial reduction attention mechanism to speed up the classification process. Besides, the SRAM module enables the transformer to estimate the significance of several pixels in input image patches and dynamically optimize their effect on the output. Conventional deep networks severely rely on physically selected attributes. These attributes often fail to interpret the subtle and complex patterns intrinsic in plant diseases, limiting the model's accuracy. CNNs have high computational costs, mainly when processing high-resolution images. These models are often overfitted to the training database. Therefore, it needs a huge labeled database for training. When models are overfitted, their capacity to calculate precisely on new data reduces (i.e., limited generality).

3. TRANSFORMERS IN IMAGE CLASSIFICATION

The visual transformer is a pioneering network that reimagines how humans analyze and interpret images in numerous visual applications. The transformer consists of three important components: a linear embedding module, an encoder, and a final classification module. Figure 1 illustrates the general architecture of the visual transformer. Let $P = \{C_i, l_i\}_{i=1}^{n}$ indicate n set of rice leaf images, in which C_i is an input image and l_i are its equivalent class label. Initially, an input image C is divided into fixed-size patches. Assume a rice leaf image C with the size of $w \times h \times \delta$, in which h is height, w is width, and δ denotes feature space size. To handle a 2D picture, the transformer splits every picture into blocks of width and length of size α (i.e., the picture is converted into square blocks). Thus, we get small blocks q_i with a size of $\alpha \times \alpha \times \delta$ from the leaf images. This division converts the image $C \in \mathbb{R}^{h \times w \times \delta}$ into a linearized heap of 2D blocks $c \in \mathbb{R}^{m \times \alpha^2 \delta}$, in which $(h \times w)$ is the size of the original picture,

$(\alpha \times \alpha)$ is the dimension of every sub-image, and $m = \frac{hw}{\alpha^2}$ is the number of image blocks extracted from the leaf picture. This generates a set of blocks $(q_1, q_2, q_3, \ldots q_m)$ of length m. In general, the patch dimension α is designated as 32×32 or 16×16, in which a smaller patch dimension leads to a lengthy stack. This network handles every block as a distinct patch. Therefore, every patch is linearized into a particular array by combining the feature maps of a picture element in a block and then directly mapping it to the designated dimension of input. It converts the linearized patches into sub-images with dimension s through a trainable mapping element as explained in the subsequent sections of this article.

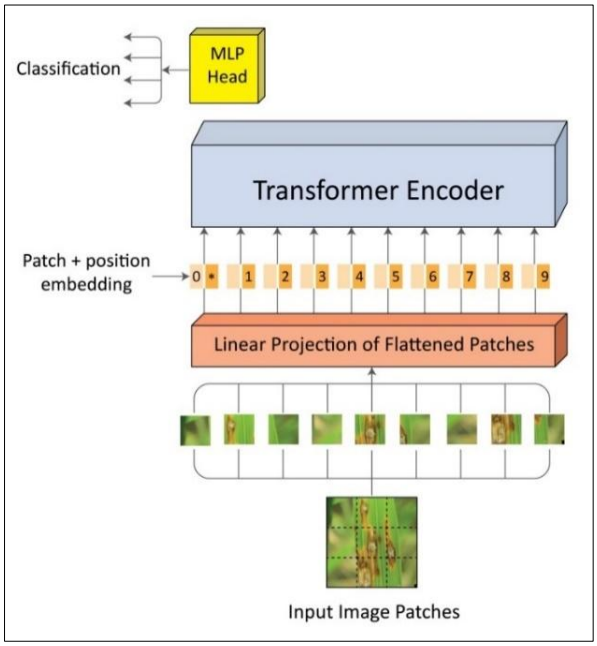

Figure 1. General architecture of transformer [25]

3.1 Patch embedding

The input image given to the transformer is divided into blocks with a certain size and is embedded into a set of attribute values as a vector. These attribute vectors are then explicitly visualized in a latent feature space. Interpreting the attributes in the latent feature space is useful for detecting the sub-image blocks with analogous attributes. The offset among attributes can be calculated from the attributes vector to define the level of the relationship. Arbitrary values are originally given and appraised during the learning process within the embedding layer. In the learning process, related attributes become nearer to each other in the latent feature space. This is crucial to detect or excerpt related attributes. Conversely, finding the location of attributes makes it easy to calculate the correlation among them. Before encoding the stack of patches, it is linearly related into a matrix of the dimension s using a trained embedding matrix β. These embedding patterns are then integrated with a learnable token. These sub-images are significant in this work to achieve RLDs detection and classification.

3.2 Positional embeddings

The position encoding technique is used to restructure the image series in their original positions and encode attribute vectors to their correct location. The attribute map and position embedding values are included to create a new vector in the latent feature space. It enables the network to distinguish between various points in the image and compute spatial relationships. To keep the spatial arrangement of the blocks as in the original picture, the position statistics β_{PI} is computed and added to the block representations. The patches with the sub-image ϵ_0 are described by Eq. (1).

$$\epsilon_0 = [T_c\,; x_1\beta; x_2\beta; \ldots . x_n\beta] + \beta_{PI}, \beta \in \mathbb{R}^{\rho^2 c \times s}, \quad \beta_{PI} \in \mathbb{R}^{(n+1)\times s} \tag{1}$$

3.2 Encoder in ViT

The resulting array of embedded blocks ϵ_0 is sent to the encoder of the transformer network. This module contains two elements: (i) a Multi-head Self-Attention (MSA) module for creating attention vectors from particular embedded graphic patches. It allows the transformer to focus on the most significant areas in the input picture (e.g., brown spot); and (ii) a Multilayer Perceptron (MLP)—a classification module and comprises 2 dense modules with a Gaussian Error Linear Unit (GeLU). Figure 2 shows the structure of encoder module in the transformer network. In this study, we use 12 MSA modules in our transformer structure. The encoder employs residual skip connections and is preceded by a Layer Normalization Module (LNM). The LNM keeps the learning procedure on track and permits the network to adapt the eccentricities in the learned data. Eq. (2) gives the statistical representation of MSA.

$$\epsilon_l' = \mathrm{MSA}(\mathrm{LNorm}(\epsilon_l - 1) + \epsilon_l - 1, \quad l = 1,2 \ldots \mathrm{L} \tag{2}$$

The statistical representation of MLP operations is given in Eq. (3).

$$\epsilon_l = \mathrm{MLP}(\mathrm{LNorm}(\epsilon_l') + \epsilon_l', \quad l = 1,2 \ldots \mathrm{L} \tag{3}$$

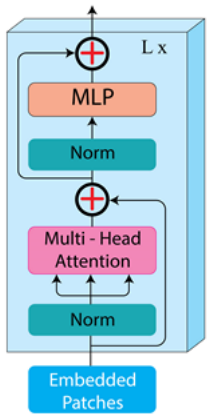

Figure 2. Encoder module in transformer

Then, the result from each self-attention module is transferred to an FFN network. An FFN network usually contains a fully connected module followed by an activation unit (e.g., Residual Linear Unit (ReLU)). It is used to add non-linearity and enable the network to capture relationships among sub-image blocks. In the final layer of the encoder, we take the first element in the stack E_L^0 and send it to an external MLP module for defining the label ψ as defined in Eq. (4).

$$\psi = \mathrm{LNorm}(E_L^0) \tag{4}$$

The multiple attention heads allow the model to capture local and global relationships and calculate the significance of a patch encoding. This module includes 4 layers as shown in Figure 3: (i) the primary projection module to relate the projected dimension of the image, (ii) the scaled scalar product attention component, (iii) the concatenate component to integrate learnable encoded patches with the other projections, and (iv) a component to get a patch mapping.

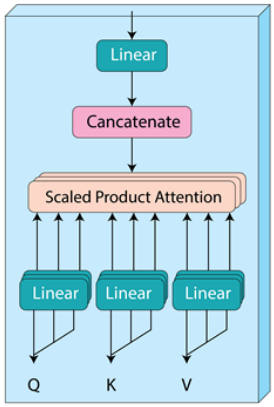

Figure 3. Structure of MSA unit

In vision-based applications, attention can be computed from the sum of weights of the picture series ϵ. The head assigns the scores by computing and compressing the scalar product of the query (Q), key (K), and value (V) as given in Figure 4. Eq. (5) provides the mathematical representation of this scalar product of each pixel and derived map χ_{QKV}.

$$[Q, K, V] = \epsilon \chi_{QKV}, \quad \chi_{QKV} \in \mathbb{R}^{s \times 3S_{key}} \tag{5}$$

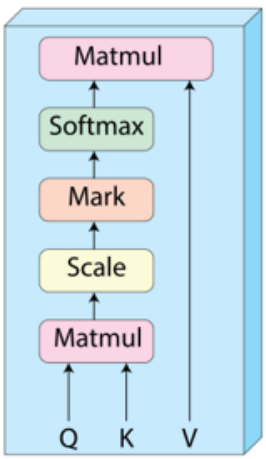

Figure 4. Scaled product attention

The outcomes define the relative importance of patches in the heap. Next, these results are compressed and transferred to a classification module. The operation of this module is described by Eq. (6).

$$\mathrm{A} = softmax\left(\frac{QK^T}{\sqrt{S_k}}\right), \quad \mathrm{A} \in \mathbb{R}^{n \times n} \tag{6}$$

where, $S_k$is dimension of the key. Finally, the value of each vector of block projection is multiplied by the output of the classification module to determine the patch with the higher attention. The whole self-attention (η) mechanism is calculated using the following Eq. (7).

$$\eta(\epsilon) = \mathrm{A}.V \tag{7}$$

The MSA unit computes the scaled dot-product score for h classifiers individually. Then, this network assimilates the outputs of every attention head and computes the final score using an FFN network with weighted parameters w to the desired dimension. Eq. (8) defines this process.

$$MSA(\epsilon) = Concat\big(\eta_1(\epsilon);\ \eta_2(\epsilon); \ldots . \eta_h(\epsilon)\big)w, \quad w \in \mathbb{R}^{hS_{key} \times S} \tag{8}$$

4. FHVIT IN DETECTING RICE DISEASE

In this work, we propose a hierarchical visual transformer to create multi-scale attribute vectors. Rice leaf images can have irregular lesion areas that may vary in shape, size, and location. These lesions may not follow any consistent pattern, making it challenging for RLD models to detect them. The attention mechanism allows FHViT to extract long-range correlations and complex spatial relations in a picture.

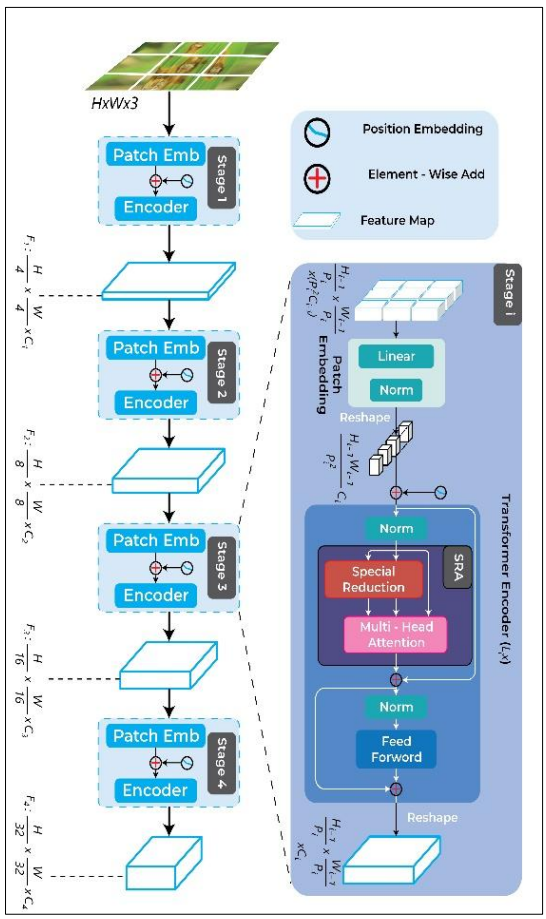

Figure 5. Structure of FHVIT

Additionally, FHViT uses a hierarchical structure that processes images at various degrees of granularity. The lower modules extract fine-grained details, whereas the upper modules emphasize high-level features. When the attention mechanism is combined with the hierarchical structure of

FHViT, the model benefits from both high-resolution processing and the ability to handle complex and irregular disease patterns, all while maintaining computational efficiency. Figure 5 illustrates the basic structure of FHViT. This model has 4 identical modules that create attribute vectors of various scales. Every stage contains a patch embedding unit and L_i encoding module. In the initial stage, the input picture of dimension $w \times h \times \delta$ is divided into $\frac{hw}{\alpha^2}$ image blocks, each of size $\alpha \times \alpha \times \delta$. In this study, initially, we set $\alpha = 4$ and $\delta = 3$. Now, we give the flattened sub-image blocks to a linear mapping and get embedded image blocks of dimension $\frac{hw}{\alpha^2} \times \delta$. Then, the embedded image blocks with a position embedding are transferred using an encoding module with L_1 layers, and the result is reformed to an attribute vector F_1 of dimension $\frac{w}{\alpha} \times \frac{h}{\alpha} \times \delta$. Similarly, using the attribute vector from the earlier stage as input, we find the following attribute vectors: F2, F3, and F4, whose strides are 8, 16, and 32 pixels in terms of the input image. With the attribute hierarchy {F1, F2, F3, F4}, FHViT can be easily implemented in most downstream applications. This model adapts an SRAM to accelerate the calculation of FHVIT.

The SRAM performs a critical role in optimizing the efficiency of FHViT, where high-resolution images with irregular lesion patterns need to be processed efficiently. The key advantages of SRAM include faster processing, reduced memory usage, improved generalization, and robustness. By combining these strengths, SRAM allows FHViT to be applied more effectively in real-world applications such as rice disease detection, where timely, accurate, and efficient diagnosis is essential for maintaining healthy crops and ensuring agricultural sustainability. The proposed SRAM decreases the size of K and V matrices by a factor of R_i^2 as presented in Figure 6. Here, i specify the stage index in the FHViT network. The spatial reduction is achieved in two steps including (i) adding adjacent tokens with a size δ in a non-overlapping window of size R_i^2 into a token of dimension $R_i^2\delta$, and (ii) mapping each of the added tokens to a token of dimension δ linearly and implementing normalization procedure. The temporal and storage costs are reduced since the number of tokens is decreased through the process of spatial reduction.

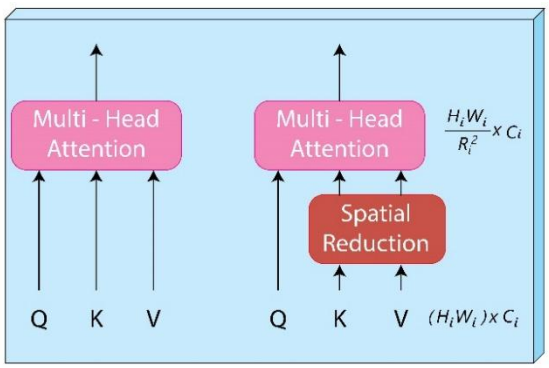

Figure 6. Attention mechanism (a) MSA (left), (b) SRAM (right)

5. EXPERIMENTS AND RESULTS

To assess the enactment of our FHViT network, it is realized using a processor called Google Tensor with 8 CPUs, 8 threads, 12GB memory, and two memory channels at a maximum frequency of 2.80 GHz. The tests are conducted using MATLAB 2023/Computer Vision Toolbox. The proposed model is validated against cutting-edge rice leaf image classification models including simple CNN [15], MobileNetV2 [16], VGG-16 [10], ResNet-50 [10] Inception V4 [10], ResNet34 [19], AlexNet [20], and Mobile ViT [22].

5.1 Image acquisition

This research utilizes a high-quality AFRLD database for learning and validation of the proposed framework [26]. The database comprises a total of 2710 rice leaf images captured under diverse dimensions and settings. Each image is categorized as Narrow Brown Spot (NBS), Leaf Scald (LS), Leaf Blast (LB), Brown Spot (BS), Bacterial Leaf Blight (BLB), and healthy (HL). In a preprocessing step, each picture in the database was cropped square from the central point to retain the most imperative (diseased) part of the picture. The database comprises high-quality images with a balanced class distribution. Table 1 shows the statistics of the AFRLD dataset.

Table 1. Statistics of the AFRLD dataset

Type of Sample	Number of Samples Used for Training/ Distribution	Number of Samples Used for Testing/ Distribution	Total Images/ Distribution
HL	371 (13.69%)	93 (3.43%)	464 (17.12%)
BLB	350 (12.91%)	88 (3.24%)	438 (16.16%)
BS	373 (13.76%)	93 (3.43%)	466 (17.20%)
LB	363 (13.39%)	91 (3.36%)	454 (16.75%)
LS	358 (13.21%)	90 (3.32%)	448 (16.53%)
NBS	352 (12.98%)	88 (3.24%)	440 (16.23%)
Total	2167 (79.96%)	543 (20.03%)	2710 (100%)

5.2 Preprocessing

Once gathering the rice leaf images, we preprocess the images to improve the quality of each pixel using contrast enhancement and filtering methods [27]. The filtering technique is used to increase the picture quality by reducing artifacts and noise in the input picture. There are numerous methods found in the literature for image smoothing (e.g., Gaussian blur filter, median blur filter, etc.), to remove noise from the input picture [28]. For the removal of reflection, artifacts, and noises from the input images, we exploit a simple thresholding method. Contrast improvement is a significant procedure to increase the quality of the input picture. This is realized by increasing the contrast of features or dropping the indecision of every pixel. We apply a contrast improvement and filtering technique to improve the picture quality [29]. This method enhances the contrast of the region of interest and makes the input picture more appropriate for analyzing the input images further.

5.3 RLD classification using FHViT

Before processing the leaf pictures, it is split into a series of certain size blocks. These patches are then linearly embedded. A token is added to perform as a representative of the entire image, which can be used for predictions. FHViT also includes position encoding and transfers the output heap of vectors to an encoding module. As the configuration of the developed transformer is straightforward, we optimize the network using

the VitModel.from_pretrained() function. As the network considers each original image to be of identical size, we employ ViTImageProcessor to rescale the normalized images. In this study, we state the recovery points of the network as parameters. Each input picture is rescaled to a particular dimension (224×224) and normalized across the feature space with mean values of (0.5, 0.5, 0.5) and SD of (0.5, 0.5, 0.5). Our framework is pre-trained on Google Tensor using the AFRLD database. To fine-tune higher-resolution pictures, we perform a 2D interpolation function of the pre-trained position encoding using their order in the original picture. We add a linear layer finally to carry out the classification task.

5.4 Performance measures

To assess the efficiency of the developed FHViT, we relate the enactment of our network with other prevailing models in terms of 5 significant measures including detection accuracy (ACC), specificity (SPE), sensitivity (SEN), precision (PRE), Recall (REC) and F1-measure (F1-M). These measures are required to be kept at greater value to increase the disease detection efficiency of the FHViT framework. The evaluation measures are calculated using Eqs. (9)-(14).

$$\mathrm{ACC} = \frac{T^- + T^+}{T^- + T^+ + F^- + F^+} \tag{9}$$

$$\mathrm{PRE} = \frac{T^+}{T^+ + F^+} \tag{10}$$

$$\mathrm{SEN} = \frac{\mathrm{T}^+}{\mathrm{T}^+ + \mathrm{F}^-} \tag{11}$$

$$\mathrm{SPE} = \frac{F^+}{T^- + F^+} \tag{12}$$

$$\mathrm{REC} = \frac{T^+}{T^+ + F^-} \tag{13}$$

$$\mathrm{F1-M} = \frac{2 \times PRE \times REC}{PRE + REC} \tag{14}$$

In the above equations, T^+ (true positive) indicates the number of samples correctly designated as diseased images; F^- (false negative) is the number of diseased samples wrongly categorized as healthy ones. T^- (True negative) is the number of samples correctly identified as normal, and F^+ (false positive) signifies the number of normal samples wrongly categorized as diseased ones.

5.5 Evaluation of FHViT

The effectiveness of the developed FHViT network is assessed by relating the empirical outcomes with that of 8 existing RLD recognition deep learning networks, including CNN [15], MobileNetV2 [16], VGG-16 [10], ResNet-50 [10] InceptionV4 [10], ResNet34 [19], AlexNet [20], and MobileViT [22].

Table 2. Disease-wise classification performance of FHViT on the AFRLD dataset

Type of Sample	ACC	PRE	SEN	SPE	REC	F1-M
HL	0.999	0.998	0.986	0.997	0.995	0.996
BLB	0.998	0.987	0.974	0.994	0.991	0.992
BS	0.997	0.988	0.987	0.977	0.974	0.984
LB	0.956	0.973	0.957	0.963	0.957	0.966
LS	0.946	0.959	0.961	0.967	0.943	0.971
NBS	0.985	0.982	0.984	0.987	0.986	0.976

Table 3. Mean value of results obtained by various RLD detection models

Algorithm	ACC	PRE	SEN	SPE	REC	F1-M
VGG-16	0.783	0.815	0.881	0.880	0.851	0.878
Inception-V4	0.886	0.919	0.908	0.905	0.888	0.912
MobileViT	0.934	0.906	0.927	0.925	0.903	0.935
ResNet-50	0.937	0.910	0.945	0.943	0.947	0.930
AlexNet	0.965	0.919	0.948	0.949	0.954	0.949
CNN	0.975	0.940	0.965	0.963	0.961	0.968
MobileNet-V2	0.984	0.969	0.974	0.971	0.973	0.972
ResNet-34	0.985	0.974	0.976	0.973	0.977	0.983
FHViT	0.992	0.993	0.992	0.991	0.990	0.990

Table 4. SD value of results obtained by various RLD detection models

Algorithm	ACC	PRE	SEN	SPE	REC	F1-M
VGG-16	0.034	0.043	0.027	0.027	0.060	0.021
Inception-V4	0.038	0.043	0.028	0.031	0.053	0.043
MobileViT	0.022	0.043	0.029	0.029	0.040	0.019
ResNet-50	0.034	0.043	0.014	0.015	0.024	0.013
AlexNet	0.007	0.043	0.031	0.031	0.014	0.048
CNN	0.018	0.044	0.030	0.030	0.036	0.031
MobileNet-V2	0.009	0.016	0.011	0.011	0.015	0.019
ResNet-34	0.008	0.018	0.015	0.014	0.011	0.008
FHViT	0.002	0.006	0.006	0.009	0.009	0.007

To achieve a more accurate result, this study employs 10-fold cross-validation (CV). In this approach, the entire dataset is fragmented into 10 portions. In every autonomous trial, one portion is used for testing and the other portions are pooled for

learning. Then, we calculate the mean value of outcomes across all 10 folds. Hence, all the outputs are specified on an average of 10 runs. To execute our intended approach, a total of 2710 images of the AFRLD database have been preprocessed by applying preprocessing techniques. The designated class tags are compared with the ground truth class tags. Table 2 lists the disease-wise evaluation measures obtained from the FHViT network. From this table, it is witnessed that the FHViT network realizes the optimum results on HL image samples with 99.9% accuracy, 99.8% precision, 98.6% sensitivity, 99.7% specificity, 99.5% recall, and 99.6% F1-M. Our FHViT model achieves classification accuracy of 99.8%, 99.7%, 95.6%, 94.6%, and 98.5% on BLB, BS, LB, LS, and NBS image samples, respectively.

The results obtained by different models in detecting rice crop disease are listed in Tables 3 and 4. Figure 7 and Figure 8 display the performance of the FHViT against other existing RLD models regarding evaluation measures. The models using the VGG-16 network use a two-stage learning process that realizes nominal performance in detecting RLDs. It shows 78.3% ACC, 81.5% PRE, 88.1% SEN, 88.0% SPE, 85.1% REC, and 87.8% F1-M. However, this model provides unreliable results and reduced recognition accuracy.

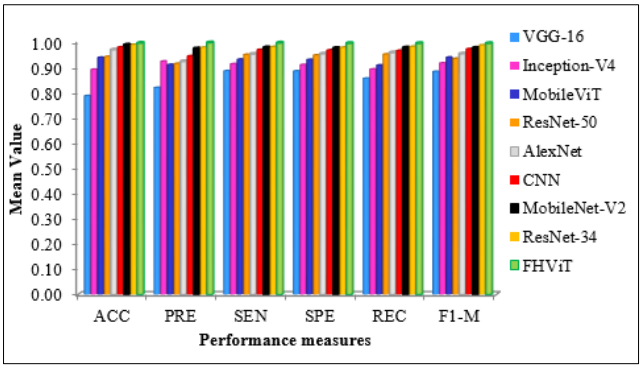

Figure 7. Performance measures of FHViT related to other models in terms of mean value

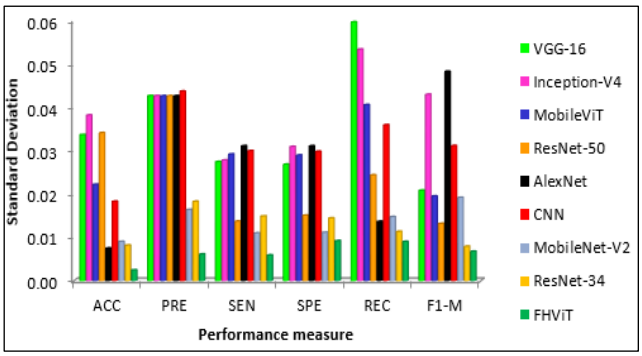

Figure 8. Performance measures of FHViT related to other models regarding SD value

By applying grid size reduction units and skip connections, the Inception-V4 model achieves better results than the CNN-based model. The skip connections enable the model to capture residual mappings, which helps increase the classification performance and convergence speed. This model provides 88.6% ACC, 91.9% PRE, 90.8% SEN, 90.5% SPE, 88.8% Recall, and 91.2% F1-M. To realize effective classification, MobileViT assimilates the idea of MobileNets and ViT using their novel MobileViT-block that learns both short- and long-range relationships successfully. It generates improved outcomes than Inception-V4 and VGG-16 models regarding the performance indicators (93.4% ACC, 90.6% PRE, 92.7% SEN, 92.5% SPE, 90.3% REC, and 93.5% F1-M). ResNet-50 includes the convolution block attention unit which improves the attribute exacerbation ability by capturing both spatial position and channel data of the image. It can realize better results regarding the performance indicators including 93.7% accuracy, 91.0% precision, 94.5% sensitivity, 94.3% specificity, 94.7% recall, and 93.0% F1 measure. The Alexnet exploits ReLU activation, dropout regularization, and data augmentation techniques to improve classification performance. This network provides better ACC (96.5%), PRE (91.9%), SEN (94.8%), SPE (94.9%), REC (95.4%), and F1-M (94.9%).

The CNN-based RLD detection model considered in this study implements the concept of low-rank approximation, network pruning, feature extraction, and hyperparameter optimization. It shows improved performance, such as accuracy of 97.5%, precision of 94%, sensitivity of 96.5%, specificity of 96.3%, Recall of 96.1%, and F1 measure of 96.8%. MobileNet incorporates several features such as inverted residuals, depthwise separable convolution, linear bottlenecks, and squeeze-and-excitation units. It shows 98.4% ACC, 96.9% PRE, 97.4% SEN, 97.1% SPE, 97.3% REC, and 97.2% F1-M. The concept of skip connections in the ResNet-34 model enables improved optimization and parameter flow, making the learning procedure easier and realizing enhanced enactment on standard databases. This model gains better performance measures such as classification ACC of 98.5%, PRE of 97.4%, SEN of 97.6%, SPE of 97.3%, recall of 97.71%, and F1-measure of 98.3%.

Our FHViT network outdoes other classification models regarding all the evaluation metrics. This model realizes better results compared to other classifiers with 99.2% accuracy, 99.3% precision, 99.2% sensitivity, 99.1% specificity, 99.0% recall, and 99.0% F1 measure. Also, it achieves improved SD values with 0.2% ACC, 0.6% PRE, 0.6% SEN, 0.9% SPE, 0.9% REC, and 0.7% F1-M. The SD values of the proposed model are less as compared to all other classifiers about the performance indicators. Hence, the FHViT provides more dependable outputs for classifying RLDs. Thus, this empirical analysis demonstrates that the FHViT is the most feasible network for detecting rice plant diseases.

6. CONCLUSION

In this research, we propose a fast hierarchical visual transformer network for RLD classification using an SRAM to alleviate the inadequacies of the conventional CNNs and basic transformers. The SRAM allows the transformer to estimate the importance of each pixel in image patches and dynamically optimize their effect on the output. We evaluate our model on an open-access AFRLD database and relate its enactment with advanced methods regarding designated evaluation measures. The developed approach revealed better performance in the detection and diagnosis of 5 infections in terms of 99.2% accuracy, 99.3% precision, 99.2% sensitivity, 99.1% specificity, 99.0% recall, and 99.0% F1 measure. Besides, it realizes better SD value with 0.2% accuracy, 0.6% precision, 0.6% sensitivity, 0.9% specificity, 0.9% recall, and 0.7% F1 measure. The proposed model relies heavily on large, labeled datasets for training to avoid overfitting problems. Imbalanced

datasets are also an issue, where certain diseases are underrepresented, which may lead to model bias, affecting the capacity of the model to generalize to rare diseases or uncommon scenarios. To avoid this problem, we plan to apply a Generative adversarial network (GAN) to produce synthetic datasets for the learning process. The model helps farmers apply pesticides and other inputs only when necessary, based on disease detection, rather than applying them across the entire farm indiscriminately. Automating disease detection and monitoring with the ViT model reduces the necessity for physical examinations by farmers. This not only saves time but also cuts labor costs, especially in large-scale farming operations.

REFERENCES

[1] United Nations, Department of Economic and Social Affairs, Population Division. (2019). World Population Prospects 2019: Highlights (ST/ESA/SER.A/423). UN: New York, NY, USA.

[2] Mohidem, N.A., Hashim, N., Shamsudin, R., Che Man, H. (2022). Rice for food security: Revisiting its production, diversity, rice milling process and nutrient content. Agriculture, 12(6): 741. https://doi.org/10.3390/agriculture12060741

[3] Lin, H.I., Yu, Y.Y., Wen, F.I., Liu, P.T. (2022). Status of food security in east and southeast Asia and challenges of climate change. Climate, 10(3): 40. https://doi.org/10.3390/cli10030040

[4] Vinci, G., Ruggieri, R., Ruggeri, M., Prencipe, S.A. (2023). Rice production chain: Environmental and social impact assessment—A review. Agriculture, 13(2): 340. https://doi.org/10.3390/agriculture13020340

[5] Shew, A.M., Durand-Morat, A., Putman, B., Nalley, L.L., Ghosh, A. (2019). Rice intensification in Bangladesh improves economic and environmental welfare. Environmental Science and Policy, 95: 46-57. https://doi.org/10.1016/j.envsci.2019.02.004

[6] Shoaib, M., Shah, B., El-Sappagh, S., Ali, A., Ullah, A., Alenezi, F., Gechev, T., Hussain, T., Ali, F. (2023). An advanced deep learning models-based plant disease detection: A review of recent research. Frontiers in Plant Science, 14: 1158933. https://doi.org/10.3389/fpls.2023.1158933

[7] Saleem, M.H., Potgieter. J., Arif, K.M. (2019). Plant disease detection and classification by deep learning. Plants, 8(11): 468. https://doi.org/10.3390/plants8110468

[8] Jung, M., Song, J.S., Shin, A.Y. (2023). Construction of deep learning-based disease detection model in plants. Scientific Reports, 13: 7331. https://doi.org/10.1038/s41598-023-34549-2

[9] Jones, R.A.C. (2021). Global plant virus disease pandemics and epidemics. Plants, 10(2): 1-41. https://doi.org/10.3390/plants10020233

[10] Bock, C.H., Chiang, K.S., Del Ponte, E.M. (2022). Plant disease severity estimated visually: A century of research, best practices, and opportunities for improving methods and practices to maximize accuracy. Tropical Plant Pathology, 47(1): 25-42. https://doi.org/10.1007/s40858-021-00439-z

[11] Andrew, J., Eunice, J., Popescu, D.E., Chowdary, M.K., Hemanth, J. (2022). Deep learning-based leaf disease detection in crops using images for agricultural applications. Agronomy, 12(10): 2395. https://doi.org/10.3390/agronomy12102395

[12] Aggarwal, M., Khullar, V., Goyal, N., Singh, A., Tolba, A., Thompson, E.B., Kumar, S. (2023). Pre-trained deep neural network-based features selection supported machine learning for rice leaf disease classification. Agriculture, 13(5): 936. https://doi.org/10.3390/agriculture13050936

[13] Zhang, Y., Wa, S., Zhang, L., Lv, C. (2022). Automatic plant disease detection based on tranvolution detection network with GAN modules using leaf images. Frontiers in Plant Science, 13: 875693. https://doi.org/10.3389/fpls.2022.875693

[14] Kumar, V.S., Jaganathan, M., Viswanathan, A., Umamaheswari, M., Vignesh, J. (2023). Rice leaf disease detection based on bidirectional feature attention pyramid network with YOLO v5 model. Environmental Research Communications, 5(6): 065014. http://doi.org/10.1088/2515-7620/acdece

[15] Qing, Y., Liu, W., Feng, L., Gao, W. (2021). Improved transformer net for hyperspectral image classification. Remote Sensing, 13(11): 2216. https://doi.org/10.3390/rs13112216

[16] Hossain, S.M.M., Tanjil, M.M.M., Ali, M.A., Bin Islam, M.Z., Islam, M.S. (2020). Rice leaf diseases recognition using convolutional neural networks. Lecture Notes in Computer Science, 12447: 299-314. https://doi.org/10.1007/978-3-030-65390-3_23

[17] Chen, J., Zhang, D., Zeb, A., Nanehkaran, Y.A. (2021). Identification of rice plant diseases using lightweight attention networks. Expert Systems with Applications, 169: 114514. https://doi.org/10.1016/j.eswa.2020.114514

[18] Kaur, P., Harnal, S., Tiwari, R., Upadhyay, S., Bhatia, S., Mashat, A., Alabdali, A.M. (2022). Recognition of leaf disease using hybrid convolutional neural network by applying feature reduction. Sensors, 22(2): 575. https://doi.org/10.3390/s22020575

[19] Haridasan, A., Thomas, J., Raj, E.D. (2023). Deep learning system for paddy plant disease detection and classification. Environmental Monitoring and Assessment, 195(1): 120. https://doi.org/10.1007/s10661-022-10656-x

[20] Stephen, A., Punitha, A., Chandrasekar, A. (2023). Designing self-attention-based ResNet architecture for rice leaf disease classification. Neural Computing and Applications, 35(9): 6737-6751. https://doi.org/10.1007/s00521-022-07793-2

[21] Lwin, L.Y.W., Htwe, A.N. (2023). Image classification for rice leaf disease using AlexNet model. In 2023 IEEE Conference on Computer Applications (ICCA), Yangon, Myanmar, pp. 124-129. https://doi.org/10.1109/ICCA51723.2023.10181847

[22] Gopi, S.C., Kondaveeti, H.K. (2023). Transfer learning for rice leaf disease detection. In Proceedings of the 3rd International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, pp. 509-515. https://doi.org/10.1109/ICAIS56108.2023.10073711

[23] Li, G., Wang, Y., Zhao, Q., Yuan, P., Chang, B. (2023). PMVT: A lightweight vision transformer for plant disease identification on mobile devices. Frontiers in Plant Science, 14: 1256773. https://doi.org/10.3389/fpls.2023.1256773

[24] Lu, X., Yang, R., Zhou, J., Jiao, J., Liu, F., Liu, Y., et al. (2022). A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest. Journal of King Saud University - Computer and Information Sciences, 34: 1755-1767. https://doi.org/10.1016/j.jksuci.2022.03.006
[25] Yu, S., Xie, L., Huang, Q. (2023). Inception convolutional vision transformers for plant disease identification. Internet of Things, 21: 100650. https://doi.org/10.1016/j.iot.2022.100650
[26] Vaswani, A.S. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, California, USA, p. 6010. https://doi.org/10.5555/3295222.3295349
[27] Ade, F. (2022). Rice Leaf Diseases Dataset. https://www.kaggle.com/datasets/adefiqri12/riceleafsv3.
[28] Liu, C.C. (2020). Adaptive contrast enhancement of optical imagery based on Level of Detail (LOD). Remote Sensing, 12(10): 1555. https://doi.org/10.3390/rs12101555
[29] Park, K., Chae, M., Cho, J.H. (2021). Image preprocessing method of machine learning for edge detection with image signal processor enhancement. Micromachines, 12(1): 73. https://doi.org/10.3390/mi12010073