An Efficient Hybrid Convolutional Vision Transformer Framework with Spatial Attention for Rice Leaf Disease Identification and Categorization

Pushpa Athisaya Sakila Rani
pushpasakila@gmail.com

Muslim Arts College, Affiliated to Manonmaniam Sundaranar University

Research Article

Keywords: Agriculture, Convolutional Neural Network, Hybrid CVT, Rice leaf disease, Vision Transformer

Posted Date: March 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8969006/v1

Additional Declarations: No competing interests reported.

An Efficient Hybrid Convolutional Vision Transformer Framework with Spatial Attention for Rice Leaf Disease Identification and Categorization

*A. Pushpa Athisaya Sakila Rani
pushpasakila@gmail.com
Department of Computer Science, Muslim Arts College, Affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India.
*Corresponding Author

Abstract

Disease detection and categorization in rice leaf play a crucial role in mitigating crop damage and supporting sustainable agriculture. Traditional approaches, which often rely on manual inspection, are limited by labor intensity, variability and error susceptibility. This paper introduces a Hybrid Convolutional Vision Transformer (CVT) model with Spatial Attention (SA) to enhance the detection accuracy and classification reliability in rice leaves. The proposed CVT framework integrates a Convolutional Neural Network (CNN), which is the backbone for initial feature extraction with a Vision Transformer (ViT) in advanced feature representation. Convolutional Neural Network captures the essential textures and shapes, while the Vision Transformer applies attention across image patches, effectively learning the complex spatial dependencies necessary for identifying disease-specific characteristics within diverse field environments. Further, SA module refines the model by assigning greater weight to diseased regions, reducing interference from non-leafbackground areas. Experimental results on rice leafdataset demonstrate that the hybrid CVT with SA model achieves over 98.12% feature extraction accuracy, 98.56% classification accuracy in dataset 1 and 98.26% feature extraction accuracy, 98.67% classification accuracy in dataset 2 across multiple rice leaf categories, outperforming baseline CNN and ViT models. Spatial Attention heat maps highlight the most important locations during decision-making process, making the model more interpretable. This hybrid CVT model offers a scalable solution for rice leaf disease detection and categorization, with potential applications in precise agriculture systems, including drone-based or mobile implementations for field monitoring. The presented model exhibits maximum performancethan the othertraditional methods.

Keywords: Agriculture, Convolutional Neural Network, Hybrid CVT, Rice leaf disease, Vision Transformer

I INTRODUCTION

A high and consistent rice yield is dependent on effective identification and management of leaf diseases. The demand for plant disease detection is outpacing the capability of traditional manual identification approaches, which are ineffectual and produce inaccurate diagnostic results. So, developing a clear and effective automated detection model for premium rice yield is essential [1]. Furthermore, indiscriminate use of insecticides to remove plant has an impact on ecosystem and rice quality. Thus, thorough rice leaf disease identification serves as the basis for avoiding and removing disease. At the same time, rice quality and production has to be improved while protecting the environment [2-4]. Recognizing plant is critical for agricultural production, since leaf catastrophes are recognized as the primary cause of crop output decrease. It takes time and effort to manually locate and identify disease. Agricultural technicians use hand lenses and microscopes to inspect and identify disease that are visible to the naked eye. This approach requires continuous crop monitoring. In large farms, this is an expensive, subjective, and labor-intensive task [5]. Conventional disease identification systems attempt in developing feature vectors to identify specific disease, resulting in the lack of generalizability [6]. Conversely, Deep Learning (DL)-based systems require pre-built object detection since the performance difference between generic object recognition and infectiondetection is rather significant. One could characterize this difference as variations in object properties and detecting criteria.

Because to DL's excellent feature learning and extraction capabilities, as well as more general advances in Artificial Intelligence (AI), advanced farming settings [7] has become much easier in terms of finding disease in various plants. On the study of rice illness [8], a Deep Convolutional Neural Network (DCNN) is utilized to suggest theactualcategorization technique for nine different types of rice infection. Using the RiceBioS app [9], biotic stress is recognized in rice leaf such as rice blast and bacterial leaf rot [10]. A DL classification system is developed in smart phone with rice sickness detection app.

Deep Learning approaches for image recognition have lately undergone major advancements. Strong performance on imagecategorization tasks by Convolutional Neural Networks (CNN) including Residual Network (ResNet) [15], Google Network (GoogLeNet) [14], Image Network (ImageNet) [12], Visual Geometry Group (VGG) [13] and Le Cun Network (LeNet) [11] has improved the knowledge and application of Neural Networks (NN). Several researchers are focusing on developing a technique for detecting diseases in rice.

Although the traditional CNN-based rice disease detection model [16-18] clearly extracts attributes of diseases, there are certain shortcomings in the present research. For starters, the visual backdrops are simple and the researchers use a limited number of ailments. Because of their complicated character, the presented models only go so far in generalizing rice leaf diseases. The existing models have not been sufficient in capturing the characteristics of diseases displayed in images. By extracting various meaningless features from the images, they regularly induce overfitting and poor model performance. Finally, models of disease detection in riceleaves are not particularly useful. The majority of research neither addresses any practical issues nor theoretical considerations. In another way, the current models are too big to be used on websites or mobile devices. [19-21].

Researchers have developed computer vision [22] algorithms usingvision transformers [23-26] to detect and quantify agricultural illnesses. Plants appear in digital images taken for disease identification. Backgrounds in the images of rice leaves diseases are been either simple or complex. A novel spatial Feature-Enhanced Attention (FEA) module [27] is created to increase backbone network performance for fine-grained image classification tasks.

This study uses Kaggle data to improve disease detection and classification in rice plants by combining hybrid Convolutional Vision Transformer (CVT) models with Spatial Attention (SA). While the transformer component provides global perspective by revealing connections throughout the whole image, convolutional layers of CVT model successfully capture local properties such as shapes and textures of leaves. By eliminating the unnecessary background noise, SA enhances the model's focus on important areaslike disease-affected areas. This integration improves accuracy, allowing for precise recognition and classification even in demanding agricultural contexts.

Modern ML algorithms utility has recently increased accuracy and efficiency of rice sickness diagnosis. Hybrid CVT [28] models combined with SAmechanisms are the particularly successful approaches. These models combine transformers' global contextual awareness with the ability of convolutional layers to extract localized features. The model enhances its ability to recognize numerous plant disease by including SA, which emphasizes the essential image features while decreasing noise. This technique delivers higher accuracy and recall than typical convolutional models and standalone transformers. Combining these technologies provides a new standard for disease control in agricultural systems, allowing for faster and more efficient management.

This paper is motivated by the critical need to resolve the challenges of rice leaf disease detection and classification, which are critical in preserving global food security. Traditional disease detection methods are time-consuming, error-prone and have limited scalability. As high-resolution agricultural imaging data becomes more widely accessible, new, automated systems capable of accurately detecting and categorizing disease in tough environments are urgently required. This study aims to reduce this gap by using hybrid CVTwith SAmodels, which provide accurate feature extraction and better model interpretability. Aside from pushing the boundaries of disease detection technology, thispaper seeks to offer the farmers with tools that promote sustainable disease management. Hence, agricultural production is increased.

*Organization of the paper:*The study is organized as follows: The related works are deliberated in section 2. Section 3 focuses on the few types of rice leaf disease. The presented methodologies are elucidated in section 4, whereas section 5 examines the results along with discussions. Section 6 represents this work's conclusion.

II RELATED WORKS

Diseases such as Tungro, Brown Spot, Blast and Bacterial Blight ripped at rice leaves. Jhatial et al.[29]an excellent and user-friendly rice leaf disease diagnostic model was built on You Only Look Once version 5 (YOLOv5) Deep Learning (DL) framework. The model had been updated to the most recent YOLOv5 version. In terms of object detecting performance and accuracy, YOLOv5 outperformedYou Only Look Once version 3 (YOLOv3) and You Only Look Once version 4 (YOLOv4). Although the proposed identification approach had several potential uses, the agricultural sector—particularly crop producing regions benefit from early disease detection—was now given a high priority. Their approach helped to detect and classify illnesses that affected the rice leaves.

Wijayanto et al.[30]this study looked into the effective data collection by drones for image analysis in West Javan rice fields that were paired with thermal and textural parameters to improve Bacterial Leaf Blight (BLB) detection. Researchers used drone data and powerful Machine Learning (ML) tools to investigate BLB impact levels. The normalized difference image proved to be quite effective because it included Haralick image features. Their findings revealed that when compared to the traditional approaches relied solely on spectral indices, incorporatedimage cues greatly enhanceddisease diagnostic accuracy. Using vegetation and image clues, Random Forest (RF) technique achieved very high classification accuracy.

Liu et al.[31]to monitor and identify disease in the rice canopy, a Generative Adversarial based Mask Region Convolutional Neural Network (GA-Mask R-CNN) target recognition model with images of rice stem bowers and rice leaf rollers from various sources. Bug generator of GA networks assisted the authors in making the classification network more responsive to data on insect bodies.It improved the reliability and accuracy of disease detection. Residual Network 101 (ResNet101)'s Mask R-CNN backbone network took an Efficient Channel Attention (ECA) module and multilevel residual connection to increase the model's recognition accuracy.Hence, gradient vanishing and explosion issues were decreasedwhen ensuring model stability and convergence. In comparison to prior disease monitoring models such as Faster-RCNN, Single Shot MultiBox Detector (SSD) and YOLOv5, the upgraded model performed better for rice disease detection. As a result, the program's quality and effectiveness were proven.

The rice plant disease detection system aimed to give accurate and timely disease predictions in aiding crop disease management. Early disease identification in rice plants allowed farmers and paddy researchers to respond swiftly for safeguarding the harvest. Daniya et al.[32] presented an overview of image processing as well asML strategies for disease detection in rice plants. These authors retrieved the images of disease rice plant leaves by combining multiple segmentation algorithms.

Li et al.[33]Brown spot, rice sheath blight and stem borer were all researched utilizing the video footage. The authors used image-training techniques to detect previously unseen movies. Their laboratory findings supported the proposed video detection approaches. Their technique to identify experimental design had been ideal for developing a Deep Convolutional Neural Network (DCNN) to recognize rice videos. Using Visual Geometry Group 16 (VGG16), ResNet-50 and ResNet-101 detection of strongly blurred images was ineffective. In similar line, same detection was used to run state-of- the-art YOLOv3 on rice lesion patches. It discovered the underperformance when the objects were shown with non-standard forms and fuzzy borders.

Majority of researches had concentrated only on the discovery of rice plants; nothing is known about the detection of rice disease. Furthermore, many new models incorporated multiple characteristics, making it unsuitable for use on mobile devices and necessitated complex technology. To overcome these limits, Jia [34]developed an improved You Only Look Once version 7 (YOLOv7) algorithm-based novel techniques for rice disease detection. Their approach was developed using Mobile Network Version 3 (MobileNetV3), a YOLOv7 object detection technology. The feature fusion component of YOLOv7 used the Coordinate Attention (CA) approach to extract spikelet features thoroughly. The suggested technique recognized numerous objects more efficiently and with less energy to YOLO algorithm improvement. This made it ideal for use in mobile device applications.

Patil et al.[35]developed the Rice-Fusion framework, a system for autonomous rice disease diagnosis in agriculture using AI and multimodal data fusion. Their research focused on three different diseases such as rice blast, bacterial blight and brown spot. Moreover, the healthy group received one type. The collection was unique with 3600 visuals, modalities and ambient features. Earlier and delayed fusion approach was one form of fusion paradigm that incorporated both modalities. Deep LearningNN served as the system's foundation, necessitated a large number of data samples for accurate model training. With an accuracy rate of 95.31%, their study showed that the Rice-Fusion data fusion approach outperformedthe current unimodal approaches such as CNN and MLP topologies. Table 1 compares the varied methods used in rice plant detection.

According to Mondal et al. [36],crop diseases were often deterrents to increase global demand for agricultural productivity that greatly influenced the yields. Such early detection was critical to reduce reliance on, losses and ensuring food security. Current methods of detecting target vehicles were traditionally intense and often relied on manual inspection or expert intervention. Automated and precise detection systems had become possible owing to recent advances in Artificial Intelligence (AI) and Deep Learning (DL). Of these, transformer-based architectures had proven a promising approach, in their capacity for handling complex visual data with efficiency. To utilize the strength of both transformers and residual networks for better detect crop diseases, X-ResFormer had been the scalable and reliable solution to the modern agricultural challenges. Development followed the rising trend of advancing edge AI in precision agriculture for sustainable farming practices.

Precision agriculture was the field of optimizing the farming field by utilizing advanced technology to assess and develop crop health. Earlier intervention against crop losses was essential in this domain with the task of plant disease detection. Conventional, current methods relied on manual labor or basic image tools that were slow, prone to human errors. In AI, the emergence of Vision Transformers became a transformative way of performing efficient and accurate analysis of plant diseases from complex visual datasets. Parez et al. [37] showed how visual intelligence reinvented industries in agriculture, solidifying sustainable farming through next generation, AI powered software.

Rice diseases affect the leaves have a big threat to food security all over the world hence need better methods of identifying the diseases. The usual visual techniques used in identifying diseases slow and less accurate; thus the development of Auto-ML. The idea of Artificial Intelligence (AI) has come to the frontline when it comes to the processing of large amounts of data and this has made it easy to classify rice leaf diseases. Overall, Mukherjee et al [38] provided a detailed overview of multiple approaches in ML application that highlight both strengths and weaknesses of the techniques with the focus on their applicability to agriculture. These authors is expected to helped to align research regarding rice disease control systems to be more sustainable and effective, so scale and reliability can be achieved in the field of agriculture.

Rice diseases controls were necessary to guarantee a stable crop production, especially in rice growing areas that were highly based on rice as food staple. Traditional ways of monitoring were laborious and commonly failed to find out the problems earlier, caused huge crop losses. Recently, Deep Learning (DL) advanced in particular Convolutional Neural Networks (CNN) had created new opportunities for automated, accurate and scale disease detection. The purpose of Rahman et al. [39] study was to understand the application of CNN in identifying rice diseases, demonstrated AI driven solutions in transforming precision agriculture and diminishing manual intervention.

Russakovsky et al. [40] ImageNet Large Scale Visual Recognition Challenge (ILSVRC) greatly influenced the status quo in computer vision and DL, due to the availability of large scale dataset and competition framework on visual recognition tasks. ImageNet consisted of a richly annotated dataset of images from which algorithms for large scale image classification, object detection and localization had been developed and evaluated. The pressure to overcome the DL limitations had pushed the boundary of DL enough to highlight Deep Convolution Neural Networks (DCNNs) as the greatest breakthrough approach helped by AlexNet architecture.

Table 1 compares the varied methods used in the detection of rice plant diseases.

Table 1: Comparison table of various methods used to detect rice plant diseases

Study	Focus Area	Key Technique/Method	Strengths	Limitations
Simonyan & Zisserman [41]	Large-scale image recognition	Very Deep Convolutional Networks	Demonstrates effectiveness of deep networks with small receptive fields	Computationally expensive due to depth
Szegedy et al. [42]	Advanced CNN design	Inception model Going deeper with convolutions	Improved efficiency and reduced parameters while maintaining accuracy	Complexity in architectural design
Temniranrat et al. [43]	Rice disease detection via chatbot	CNNs integrated with chatbot service	User-friendly, real-time rice disease detection system	Limited scalability to other crops
WAHYUN & Sitio [44]	disease detection in rice	Bayesian inference system	Simple and interpretable probabilistic model	Limited accuracy compared to advanced ML techniques
Wiatowski& Bölcskei [45]	Mathematical theory of CNNs	Analytical model of CNN feature extraction	Provides theoretical understanding of CNN capabilities	Focuses on theoretical aspects, lacks practical implementation

Pallathadka et al. [46]	Rice leaf disease detection	Machine learning algorithms	Demonstrates potential of ML in agricultural diagnostics	Limited exploration of DL methods
Vasantha et al. [47]	Rice leaf disease detection	Various ML algorithms	Comparative analysis of ML techniques	Lacks application of advanced DL models
Kathiresan et al. [48]	Rice disease detection	Transfer learning using pretrained models	Efficient and accurate detection with less data	Dependent on quality and relevance of pretrained models
Chung et al. [49]	Detection of Bakanae disease in rice	Machine vision techniques	Early detection through image processing	Limited focus on a single disease
Jiang et al. [50]	Recognition of rice leaf diseases	DL with SVM	Combines DL feature extraction with SVM classification for improved performance	Computationally intensive; specific to four diseases

2.1 Problem Identification

Infestations in rice fields are very sensitive, affecting both quality and yield. Current rice plant disease identification methods rely on standalone DL models, which exhibit significant intra-class similarity and inter-class heterogeneity in rice leaf images. These limits reduce accuracy and efficiency, especially in real-world agricultural contexts where the data is noisy and unstructured.

III Rice Leaf disease

1) Bacterial blight

Bacterial blight affecting rice leaves is a highly damaging disease. Attacks target the early stages of plant growth. Strong gusts or heavy rain cause this illness to spread throughout the field. If the nitrogen requirement is in proper levels, this disease is treated. Watering the plants assist inpreventing the disease and kept the field clean. Figure 1 displays the image of bacterial blight.

.

Figure 1: Bacterial blight

2) Blast

Rice blast is another deadly sickness which wipes out the whole agricultural field. While the fungus infects every stage of plant growth, this sickness targets the roots of plants. Environmental conditions have an influence on crop shape and size variations. This disease is managed with the use of licensed fungicides. The plant’s leaves with blast are shown in figure 2.

Figure 2: Blast leaf

3) **Brown leaf**

Maintaining crop health via fertilization, land leveling, soil preparation and drainage control helps to reduce damage. Seeds have to be protected against fungus transported by seeds and fungicides. The leaves shown in figure 3 depict the image of brown leaf.

Figure 3: Brown leaf

4) **Kernel smut**

Kernel smut develops blackish spores which replaces the endosperm, decrease grain quality and yield. Typically growing in warm and humid settings, the sickness is caused by either delayed harvesting or excessive nitrogen fertilization. Using resistant rice varieties, optimal fertilization, timely harvesting and fungicidal treatments are integrated in disease management approaches. Kernel smut is viewed in figure 4.

Figure 4: Kernel smut

5) **Tungro**

When Tungro originally develop at the earlier stage of evolution, it has been the most devastating and deadly disease known to man; it destroys an entire rice field. This disease affected leaves when nitrogen and zinc deficiencies occuralong with water stress and rodent damage.Tungro disease is depicted in the image of figure 5.

Figure 5: Tungro

IV MATERIALS AND METHODS

4.1 Dataset collection

To assemble the necessary information on various rice leaf diseases, the researchers has investigated and evaluated rice disease records as well as conducted interviews with Department of Agriculture—particularly those from the Regional Crop Protection Center. Images of several rice plant diseases are collected using the means available, which included digital cameras and smart phones. After gathering, all the imagesare pre-processed and included in the dataset.

Dataset 1: https://www.kaggle.com/datasets/nashehannafii/datasetleafblast

Dataset 2: https://www.kaggle.com/datasets/vbookshelf/rice-leaf-diseases

4.2 Data preprocessing

To provide high-quality input for model training, preparing a dataset for rice leaf detection involves many critical steps. Image data is initially cleaned by removing duplicates and low-quality samples such as blurry or poor lit images. The data is further labeled to represent various bug types or groups. Resizing and normalizing the images ensures that the input dimensions and pixel intensity ranges remain constant. While image improvement enhances the original dataset quality, augmentation increases the dataset's size. Increasing visual detail as well as smoothening by flattening and contracting the images. Rotation, flipping and cropping are some of the augmentation strategies used to diversify datasets and improve model robustness. These operations increase the dataset's quality, allowing for effective training of the detection model.

4.3 EXISTING METHODOLOGIES.

In existing methodologies,the methods such as CNN, Vision Transformer and SA are discussed and usedfor detecting and classifying infection in rice leaves.

4.3.1 CNN

Disease detection in rice leaves are performed using CNN model. Convolutional Neural Network is the core of most advanced computer vision solutions since they perform a wide range of tasks.

4.3.1.1 Utility of CNN in Feature Extraction and Segmentation

In CNN, each neural node has been linked with the next convolutional layer. It examines each element separately using the provided visual data. This element is called as an image feature. Convolutional layer aligns the features of input images before multiplying each pixel with matching feature pixel. After adding the values of each pixel in feature, divide the sum by pixel count of feature. While the feature map leads the filter, computed values cover the entire image. Every computed value is displayed on feature map. Each feature is treated to a method that results in original feature maps. To get convolutional layer, apply the equation (1).

$$C_{ijk} = \sum_{a=0}^{a-1}\sum_{q=0}^{H-1}\sum_{s=0}^{H-1} R_{i+q,j+s,a^h qsak}^{(l-1)} + n_{ijk} \text{----------} (1)$$

$C_{ijk} is$ the value of the neuron at position *(i,j,k)* in the output of the convolutional layer. $R_{i+q,j+s,a^h qsak}^{(l-1)}$indicates the filter (or kernel) applied to the image at a specific position.Bias often specified asn_{ijk}, is independent of the image pixel position. The indices *q, s* iterate over the kernel size H to apply the filter across the input image.h_{qsak}represent the precise weight value.Convolution operation determines the degree of fit of a feature (specified by the kernel/filter) with the local parts of the image. Kernel acts at every point in the original image and performs a weighted sum (Which includes the bias also) to generate feature map.Rectified Linear Unit (ReLu) has seized control and appended zeros to every negative value in feature map after the restoring of activation capacity. The activation ReLu function has form in equation (2).

$$F(x)= max(0,x) \text{ --------- } (2)$$

F(x) returns the maximum of 0 and x. If x is positive, the output remains the same. If x is negative, it is set to 0. Max pooling layer condenses the input image by aggregating the highest value from convolutional layer. In equation (3), the most important pooling layer has an equation that exposes the follows:

$$C_{ijk} = \max R_{qsa} \text{ ---------- } (3)$$

In the above equation (3), q, s ∈ $Q_{i,j}$. $Q_{i,j}$ returns a collection of pixels that cover the area. For each channel, C_{ijk} is calculated using H^2 pcs pixel value. Completely linked layer converts the low-resolution images generated by the model's last pooling layer into a single vector array. Categorization process occurs in the layer of fully linked links. The foundation of CNN has influenced the utilized methodology to identify disease in rice leaves. Classifier receives animage as input and the convolution layer creates neuron output to the input's logical region. This model projects images in this manner. Pooling layers reduces the spatial dimensions of the process. The method is repeated at every step until the last totally linked layer is obtained. This layer is responsible for determining the class scores to create the forecast. Figure 6 illustrates the usage of CNN for detecting the disease in rice leaves.

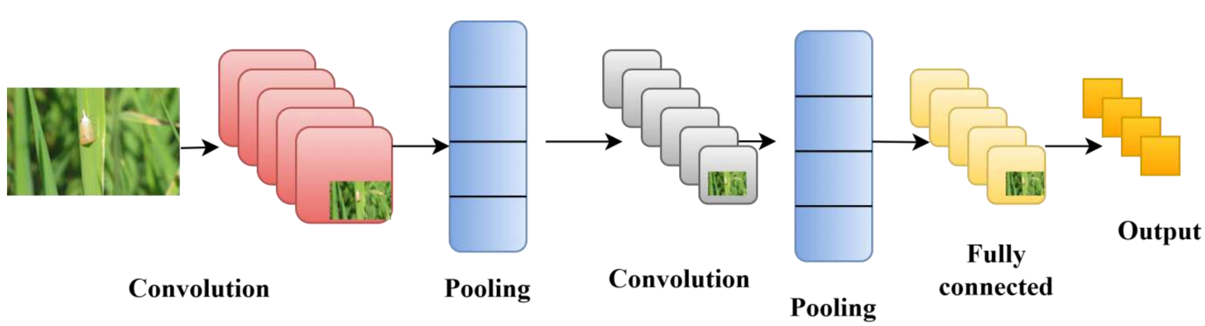

Figure 6: Rice leaf disease detection using CNN

Researchers retrained this algorithm with the images of rice leaves and their illnesses are presented in their findings. As training progressed, model's performance is also recorded and saved to a file.

4.3.2 Visual Transformer

Salamai et al. [51] produces outperforming classification performance using far less memory. Figure7 depicts the ViT model structure. Vision Transformer has been used to classify several leaf diseases across the wide range of crops. It generates multiple fixed-size patches from the images for classification. A transformer encoder learns inter-patch connections by flattening each patch in sequence, resulting in a feature vector that encodes the image. The final classification comes from a MLP head.

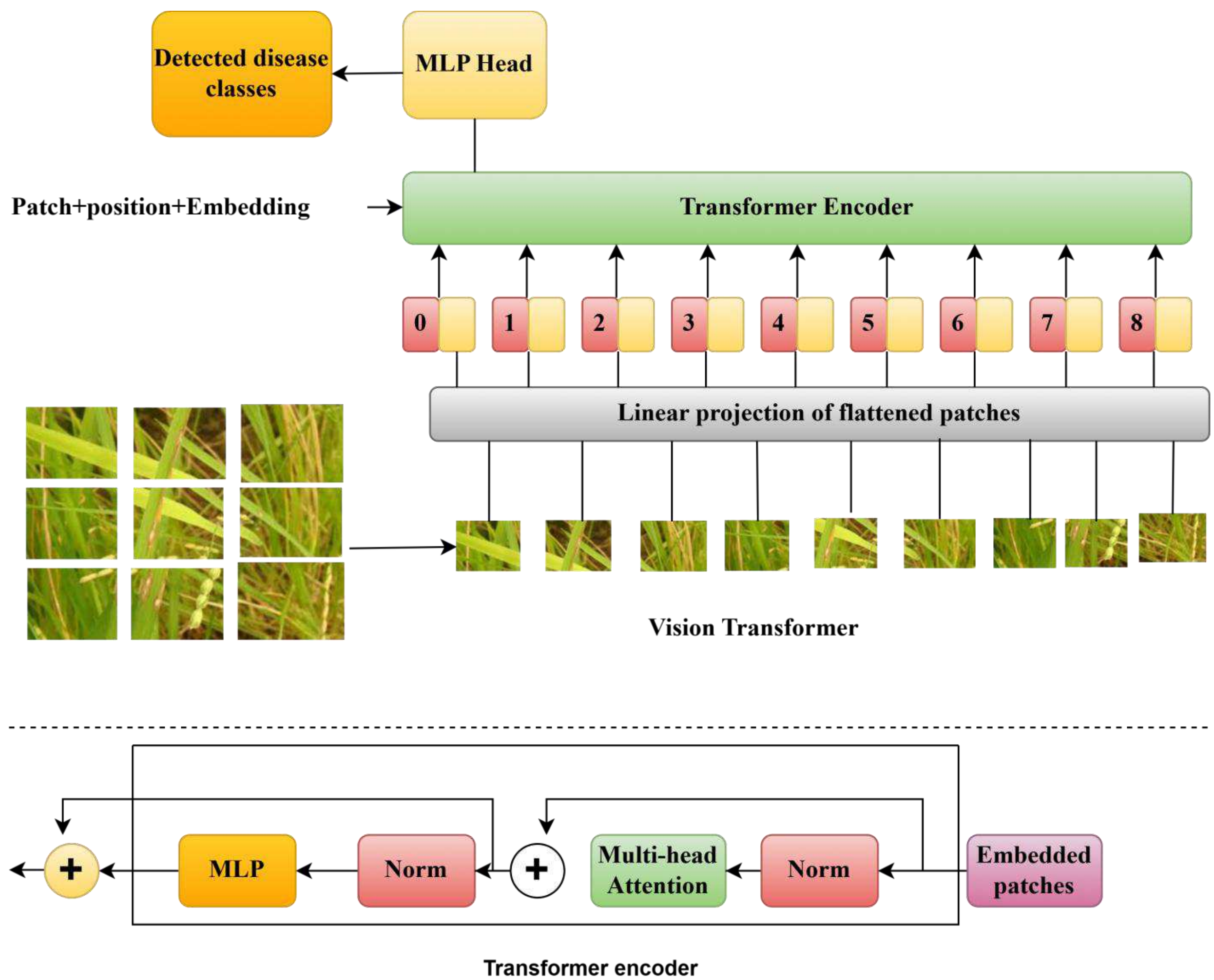

Figure 7: Vision Transformer for rice plant disease detection

Salamai et al. [51] developed a ViT-based lesion-aware paddy leaf disease diagnostic system for identifying and distinguishing lesion features. They used a multiscale contextual feature extraction network to gather local and global illness features across multiple scales and channel. In addition, a feature-tuning unit is used to enhance illness diagnosis by increasing spatial exchanges between visual meanings and weekly supervised Paddy Lesion Localization (PLL) strategy is devised for selecting lesion locations.

Least Important Attention Pruning (LeIAP) is utilized to locate and eliminate less significant attention heads in each transformer layer, which result in better F1 scores than the original FormerLeaf. Sparse Matrix-Matrix Multiplication (SPMM) is exploited for optimizing the methods training phase, which reduces computational and memory expenses. This strategy lowers computation complexity and resource utilization by calculating the associations between image patches, resulting in faster training. Vision Transformerprioritizes distant feature dependencies, makes it difficult to converge quickly on short datasets, even if it eliminates several TL concerns than CNN models. However, combining CNN's local feature extraction with ViT spatial-attention modules allow for more efficient collection of both local and global elements.

4.3.3 Spatial Attention

A unique spatial FEA module improves performance on fine-grained image categorization. In coarse-grained image classification problems, it is necessary to find the most discriminative aspect. The primary goal of fine-grained image categorization is to develop dependable modules or technologies capable of detecting significant inter-class variance as well as minor intra-class variance. Improving performance in fine-grained image categorization tasks is difficult because the present attention tactics focus too much on evident elements and too little on the nuances that typically verify identification judgments. While dealing with fine-grained image classification problems, efficiently extracting information-rich areas from images requires attention on the spatially important components of image.

Start by downsamping the features using a well-known pooling technique in image retrieval—Generalized Mean Pooling (GEM), as referred by Radenović et al [52]. GEM contains the P parameter. GEM uses mean pooling for $p = 1$, maximum pooling for $p\ \infty$ and the combination of maximum and mean pooling when $p < \infty$. Feature averages are useful for GEM pooling. It significantly boosts the p-power of a feature map by utilizing all its pixels. The following equation (4)is its predefined formula.

$$F=[f_1 \dots f_k \dots f_c]^T, f_k = (\frac{1}{|X_1|}\sum_{x\in X_1} x^p)^{\frac{1}{p}} \text{ ----------- } (4)$$

A pooling layer's input data is $Xi \in \mathbb{R}^{h\times w\times c}$, where c represents the channel count and $f \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times c}$ indicate its output. Features are extracted using convolution with a 7 x 7 kernel size, and gain final SA using Sigmoid after fitting the input dimension through up-sampling f. Here is the specific equation (5).

$$F = \sigma\left(conv_{7\times 7}\left(F_u(f)\right)\right) + X_1 \text{ --------- } (5)$$

In this instance,$conv_{7\times 7}$ symbolizes the 7 × 7 convolution and σ is the activation function utilized. Figure 8 shows a schematic view of SA module.

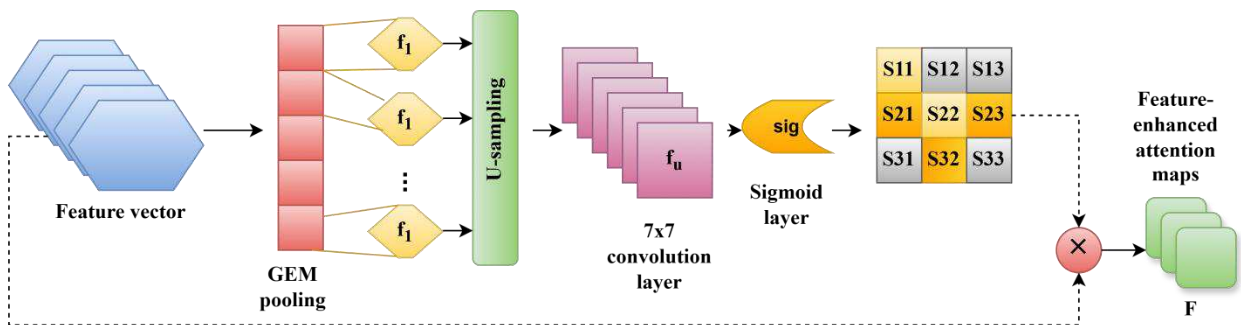

Figure 8: Spatial Attention model

4.4 PROPOSED METHODOLOGIES

Convolution Neural Network is combined with ViT and SA for improving the detection of infection in the rice leaves. A hybrid proposed model is discussed in this section.

4.4.1 Hybrid Convolutional Vision Transformer

Wu et al.[53]two significant enhancements in hybrid CVT include a new convolutional token embedding in the transformer hierarchy and a convolutional transformer block with a convolutional projection. Two convolutional approaches are used in ViT architecture namely; convolutional network for projecting and embedding tokens

First, convolutional token is the embedding layer tokens constructed to fit the 2D spatial grid and performs convolution using overlapping patches. The degree of overlap is determined by stride length. Token standards are more advanced. While the number of features (feature dimension) increases with each level, number of tokens or feature resolution decreases. This produces spatial downsampling and improved representation richness, as in CNN architecture.

Unlike previous systems that relied on transformers, ad hoc positions embedded in the tokens is not aggregated.Each phase includes a recommended stack of convolutional transformer blockings on other side. Classification token is further used in subsequent actions. A fully linked MLP head is used as last resort before the output classification predicts the class. The convolutional token embedding layer has to be considered first. Next, a convolutional projection is created, which showsthe successful controlling of computational costs in the Multi-Head Spatial-Attention module.

4.4.1.1. Convolutional Projection for Attention

To increase the efficiency, suggested Convolutional Projection layer allows undersampling of *K* and *V* matrices in order to represent the local spatial environment better. The suggested Transformer block is an adaptation of the original, which allows convolutional projection. Previous steps included more complex designs for convolution modules intended for the integration into transformer block, which increased computation costs. The model is replacing the original position-wise linear projections with depth-wise separable convolutions for constructing the Convolutional Projection layer for Multi-Head Spatial-Attention (MHSA).

Implementation Details

Figure 9shows position-wise linear projection of ViT.Convolutional projection uses a kernel-sized convolutional layer that is depth-wise separable. Finally, a one-dimensional framework is constructed to accommodate future token storage and usage. However, another option is flattening the layer is given in equation (6):

$$x_i^{\frac{q}{k}/v} = Flatten(conv2d(Reshape2D(x_i), s) \text{ -------------- } (6)$$

Apply the following guidelines: Both depth-wise and batch-wise separable convolution is possible. x_i at layer i is the unaltered token before the convolutional projection, whereas *Conv2d* defines the size of the convolution kernel. The *q/k/v* matrix token input is $x_i^{\frac{q}{k}/v}$ i. By combining the previous position-wise linear projection layer with a 1 x 1 kernel size convolution layer, one can easily construct the updated Transformer Block with the Convolutional Projection layer. Convolutional projections are depicted in figure 10.

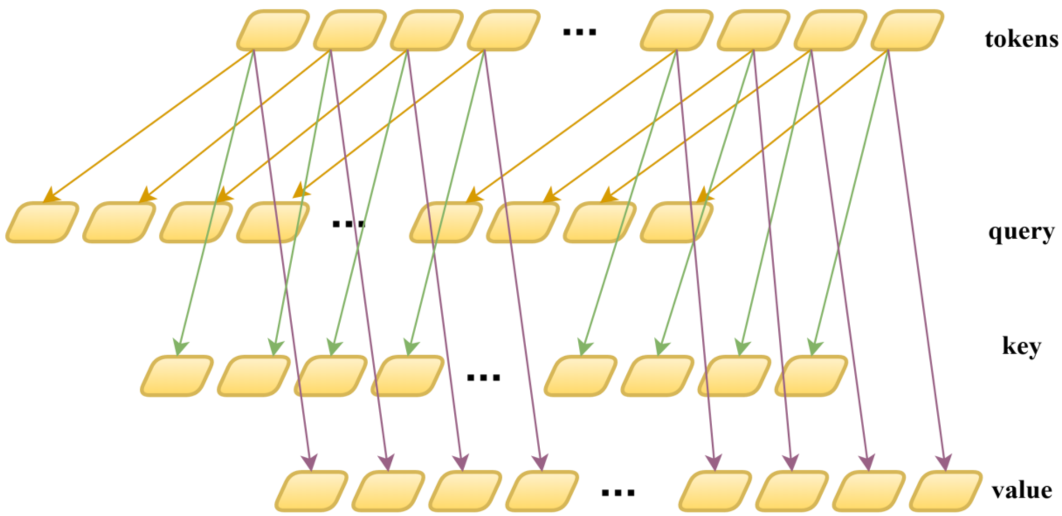

Figure 9: ViT linear projection

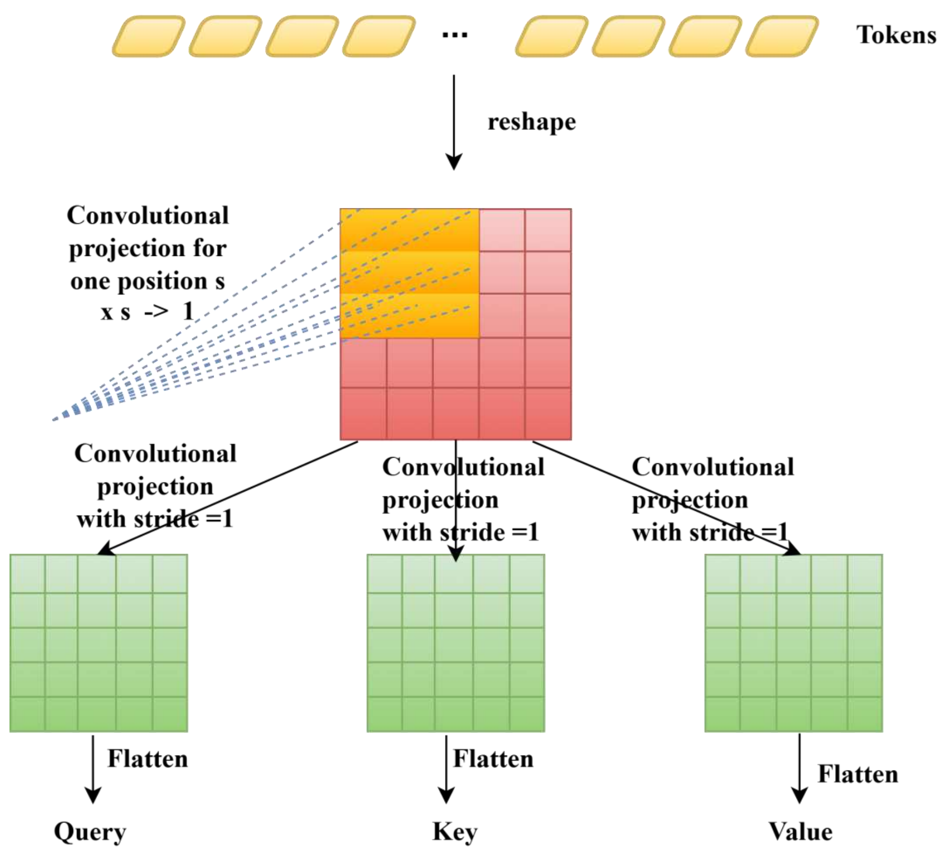

Figure 10: Convolutional projection

As seen in figure 11, convolutional projection, a convolution with a stride larger than 1 subsamples the key and value projections. Searching is done in one step, whereas key and value projection requires two. The token values required for value and key are four times smaller, as is the computational cost of the MHSA technique. Typically, there is a short lag while adjacent pixels or sections of an image react and appear identically. The suggested convolutional projection uses local context modeling to compensate for data loss caused by decreasing resolution.

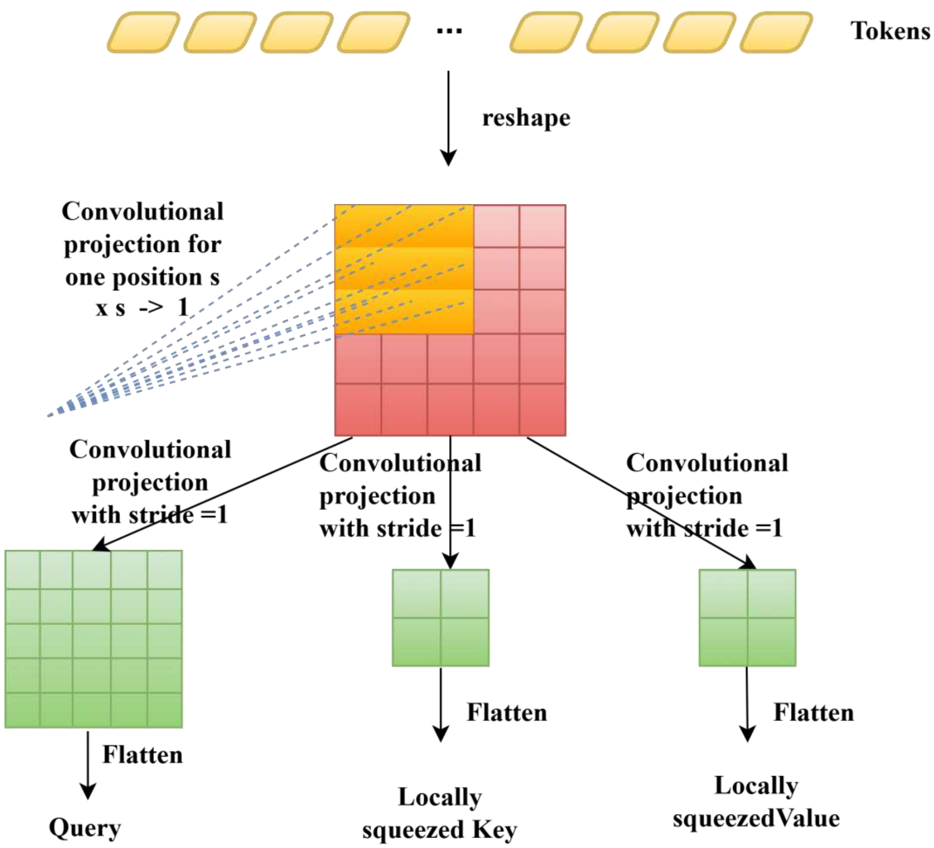

Figure 11: Squeezed convolutional projection

4.4.2 CVT with SA

Disease infections particularly in vital crop rice become the significant barriers to agricultural productivity. Reducing losses and ensuring sustainable farming systems rely on fast and accurate infection detection. Deep Learning has been recently improved forproducing hybrid methods likeCVT, which combines the capabilities of ViT with CNNs. These models are extremely useful for identifying rice leaf disease in complex agricultural contexts, since SA processes enhance these methods.

The method’s paradigm aims to address the shortcomings of current techniques by utilizing local feature extraction and transformers to capture long-range correlations in visual input via CNN. Convolutional layers successfully extract low-level information from leaf images such as edges and textures. Transformer layers, which usespatial-attention to encode global contextual links, assess these feature maps later. Spatial Attention systems increase the model's capability by prioritizing the image.

The architecture of CVT with SA is defined by three key components: SA modules, transformer encoders and convolutional blocks. To improve generality, diseasedimages are first resized, normalized and data augmented. Convolutional layers use preprocessed images to generate feature maps. A convolutional operation is mathematically stated as *F=conv (I; W);F* is the generated feature map,I is the input image and W is the convolution kernel weight.After accepting these flattened feature maps, transformer encoder uses the following equation (7) to compute spatial-attention:

$$A\ (Q,\ K,\ V)=Softmax\ (\frac{QK^T}{\sqrt{d_k}})V \text{ ---------- } (7)$$

Feature maps create three matrices: value from *V*, query from *Q* and key from *K*. The key's dimension is denoted as d_k. Combining several images aids the model in identifying diseases on rice leaves by capturing global trends.The SA module creates an attention map $M_s$that highlights disease regions in the feature map.Thus, the operation is enhanced. The map is computed as follows in equation (8):

$$M_s(x, y) = \frac{\sum_c |F_c(x,y)|}{\sum_{x,y,c} |F_c(x,y)|} \text{ -------- } (8)$$

Where $F_c(x, y)$ represent the channel *c* activity at the spatial position *x* and *y*. This normalized weighting ensures that the next layers emphasize regions with high feature intensity. Finally, CVT model generates bounding box localization coordinates as well as categorization labels that represent various diseases. Using CVT withSA for rice leaf disease detection requires many processes. First, data is collected and interpreted. Rotation, flipping and cropping are some of data augmentation strategies used to simulate the real-world scenarios and reinforce the model. For the tasks likedisease location and categorization, model is trained with task-specific loss functions and Adam optimizer. Loss function L, commonly aggregates cross-entropy loss. Mean squared error is calculated in equation (9):

$$L = L_c + \lambda L_l \text{ ------ } (9)$$

Where, λ balances the importance of two tasks. Loss function for classification and bounding box linear egression are represented as $L_c$andL_l.

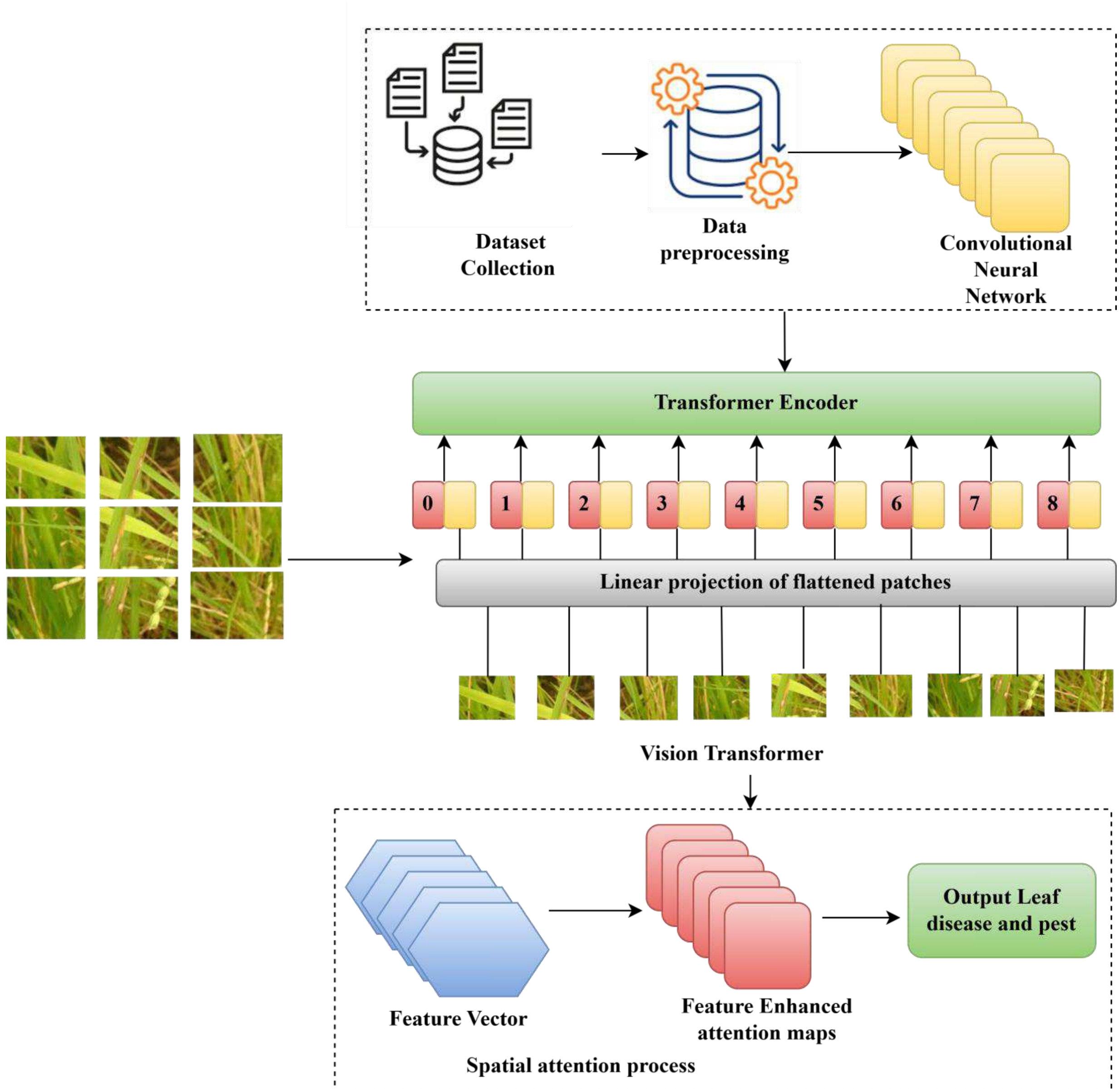

Figure 12: Overall architecture of rice leaf disease detection

The entire architecture of the proposed frame work is displayed in figure 12. The model performs well throughout testing, demonstrating high accuracy and durability in challenging scenarios like occlusion and shifting light. Convolutional Vision Transformer with SA is very useful for drone-based or mobile disease monitoring because of its ability to concentrate on the crucial portions of image, allowing for precise disease identification and localization in real time. Ideal for drones and mobile platforms, its lightweight design ensures efficient operation on devices with limited resources. Spatial Attention mechanism promotes adaptability to the challenging environmental conditions such as varied lighting and occlusions. Algorithm 1 describes the procedures involved in Convolutional Vision Transformers with Spatial Attention.

Algorithm 1: Convolutional Vision Transformers with Spatial Attention

Procedures:

Step 1: Preprocessing Phase

Input Image Preparation: Load the image I of size $H \times W \times C$. Where, H and W are height and width and C is the number of channels

Image Normalization**:** Normalize pixel values to the range [0, 1] for consistent feature scaling:

$I\ (norm) = I(x,y,c)\ \frac{I(x,y,c)-\mu}{\sigma}$

Data Augmentation**:** Apply transformations like rotation (θ), flipping and cropping to create diverse training examples.

Step 2: Feature Extraction using Convolutional Layers

Apply Convolutional Layers**:**

Extract feature maps F from *I(norm):*

$F = Conv\ (I\ (norm):\ W)$

Where, W is the convolution kernel and F is of size $H' \times W' \times C'$

ReLU Activation**:** $F' = ReLU(F)$

Step 3: Transformer Encoding

Flatten and Tokenize Feature Maps: Convert feature maps F''into a sequence of tokens T:

$T = Flatten\ (F'')\ + Positional\ Embedding$

Multi-Head Spatial-Attention: Perform attention on tokens T:

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

Step 4: SA Mechanism

Compute SA Map: Aggregate channel-wise feature responses to highlight regions of interest

$$M_s(x,y) = \frac{\sum_c |F_c(x,y)|}{\sum_{x,y,c} |F_c(x,y)|}$$

Enhance features with spatial weights

$F(x,y,c) = F_c(x,y) \cdot M_s(x,y)$

Step 5: Disease Classification and Localization

Global Average Pooling:Compress feature maps to reduce dimensionality:

G=GlobalAvgPool(F)

Fully Connected Layers for Classification:Pass G through dense layers to predict disease class P:

P=Softmax(W_{fc}*G+*b_{fc}*)*

Bounding Box Prediction:

Regression is used to predict the bounding box coordinates (x1, y1, x2, y2). x1, y1 are the coordinates of the top-left corner of the bounding box. X2, y2 are the coordinates of the bottom-right corner

B=W_r*G+*b_r

Step 6: Loss Function

Classification Loss:Cross-entropy loss for disease classification:

$$L_c = -\sum_i y_i \log(p_i)$$

$y_i$denotes the ground truth labelwhile the predicted probability is indicated as p_i.

Localization Loss**:** Mean Squared Error (MSE) for bounding box prediction:

$$L_l = \frac{1}{4}\sum_{j=1}^{4}(B_j - B_j^{true})^2$$

Combined Loss**:**

$$L = L_c + \lambda L_l$$

Combined loss L is the weighted combination of L_c andL_l, with the goal of maximizing bounding box accuracy and disease prediction. L_c ensures that the model properly identifies the disease by using cross-entropy, which penalizes incorrect predictions based on the probability assigned to right class. $L_l$decreases the similarities betwixt predicted as well asreal bounding box coordinates by using Mean Squared difference, which measures the bounding box's positional accuracy. A weight factor λ ensures that the model appropriately prioritizes categorization and localization during training and balances their relevance. This harmony is essential for effective diseased leaf identification and localization.

V RESULTS AND DISCUSSIONS

In this paper,Python is used for implementation. The proposed CVT with SA mechanism detects the rice leaf disease accurately than existing methods. This suggested hybrid method which extracts features from visual data using CNN, allowing for the collection of local patterns such as textures and edges while distinguishing between various rice leaf diseases.Vision

Transformer is used for modeling global dependencies in images by dividing into patches and applying spatial-attention mechanisms and improving rice leaf detection accuracy. By focusing on relevant spots in images, the model improves its ability to identify and rank contaminated areas by using SA to give more weight in geographical information. This improves disease detection accuracy by reducing distractions from irrelevant background information in image. This proposed method has enhanced theleaf disease detection using these combined methods. Convolutional Vision Transformer with SA achieves 98.12% accuracy in feature extraction and 98.56% accuracy in classification using dataset1. In dataset 2, this method achieves98.26% accuracy in feature extraction and 98.67% accuracy in classification. The proposed method outperforms than the methods.

5.1 Performance Metrics for rice plant disease detection

5.1.1 Accuracy

In predictive modeling, accuracy is the measure of closest model's projections is to real-world outcomes. It evaluates the model because many choices and forecasts rely on its accuracy and dependability.

T-True, F-False, P-Positive, N-Negative

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{ ------- } (10)$$

5.1.2 Precision

In predictive modeling, accuracy is the proportion of total expected positive observations to correctly forecast positive observationsit demonstrates the model'ssuccessful reduction of false positives, guaranteeing that the positive forecasts it delivered are accurate and reliable by extension and error reduction in many other domains.

$$Precision = \frac{TP}{TP+FP} \text{ ------ } (11)$$

5.1.3 Recall

Recall in predictive modeling is the fraction of real positive instances properly detected in the model. In sectors like disease detection, identifying all positives is critical since it shows the efficient detection of all relevant instances in a particular class.

$$Recall = \frac{TP}{TP+FN} \text{ ------ } (12)$$

5.1.4 F-measure

F-measure, which determines the harmonic mean of recall and accuracy, is a strong all-around measurement of the efficient performance in model that is necessary for preventing both false positives and false negatives.

$$F - measure = 2 \times \frac{Precision \times recall}{precision + recall} \text{------- (13)}$$

Table 2: Performance comparison of feature extraction using dataset 1

Feature Extractionusing Dataset 1				
Algorithm /Metrics	**Accuracy %**	**Precision %**	**Recall %**	**F-measure %**
CNN [33]	95.14	94.67	94.96	94.20
ViT [18]	95.87	95.12	95.45	95.02
SA [21]	96.45	96.21	96.32	96.01
CVT [53]	97.75	96.67	97.23	96.54
CVT with SA	98.12	97.85	98.02	97.67

Table 2 illustrates the comparison of feature extraction for CNN, ViT, SA, CVTand CVT with SA using dataset 1 on the performance metrics. Including SA allows the proposed CVT to outperform the existing approaches for disease detection in rice leaf. It has performed with 98.12% accuracy using dataset 1.

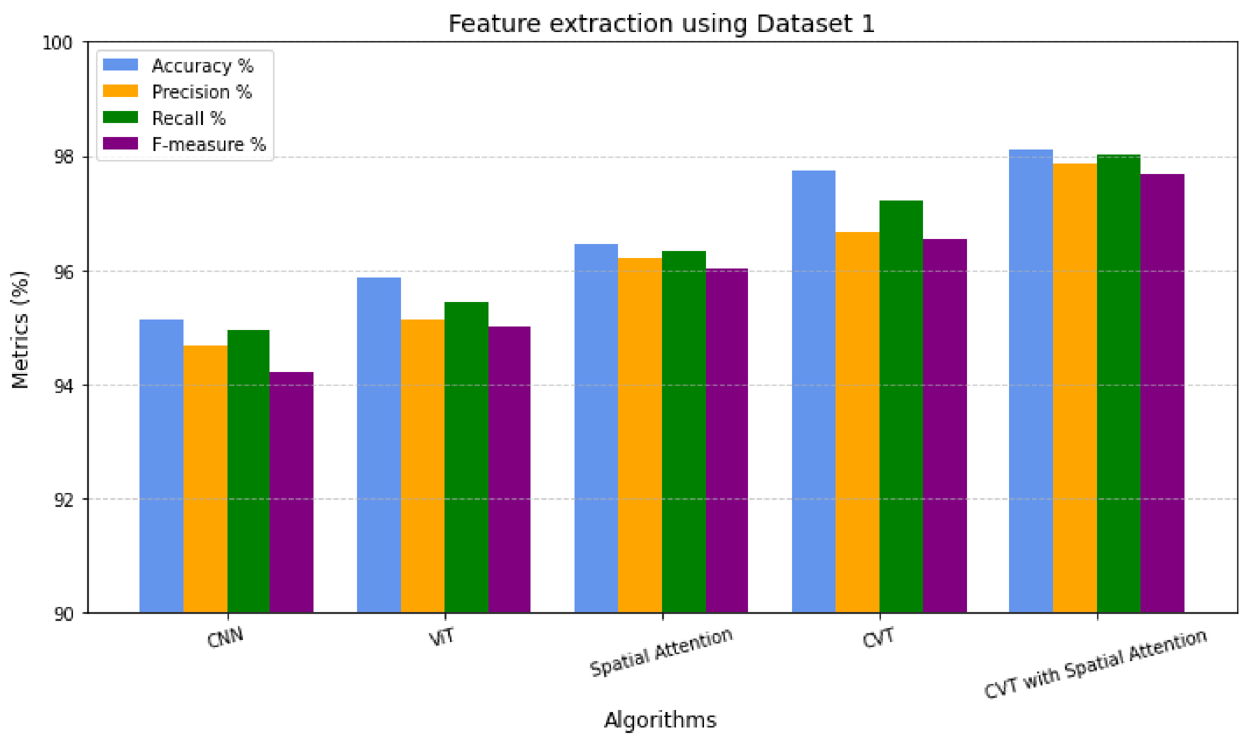

Figure 13: Comparison flow chart of feature extraction using dataset 1

Figure 13 shows CNN, ViT, SA, CVT and CVT with SA on dataset 1, together with other performance metrics such as recall, accuracy and precision. The proposed CVT with SA outperforms in rice leaf disease detection using dataset 1 compared to the other existing methods. While the inscription of metrics percentage on the y-axis of the chart, the methodologies currently in use and explored in this work are represented on the x-axis.

Table 3: Comparison on classification using dataset 1

Classification using Dataset 1				
Algorithm /Metrics	**Accuracy %**	**Precision %**	**Recall %**	**F-measure %**
CNN [33]	95.80	95.01	95.76	94.98
ViT[18]	96.74	95.67	96.64	95.23
SA [21]	97.56	96.96	97.45	96.45
CVT [53]	98.12	97.54	97.89	97.04
CVT with Spatial Attention	98.56	97.78	98.34	97.64

The performance metrics of CNN, ViT, SA, CVT and CVT with SA is shown in table 3 for the classification on dataset 1. Regarding disease identification in rice leaves, the suggested CVT with SA outperforms current approaches in terms of accuracy. On dataset 1, it has achieved 98.56% accuracy.

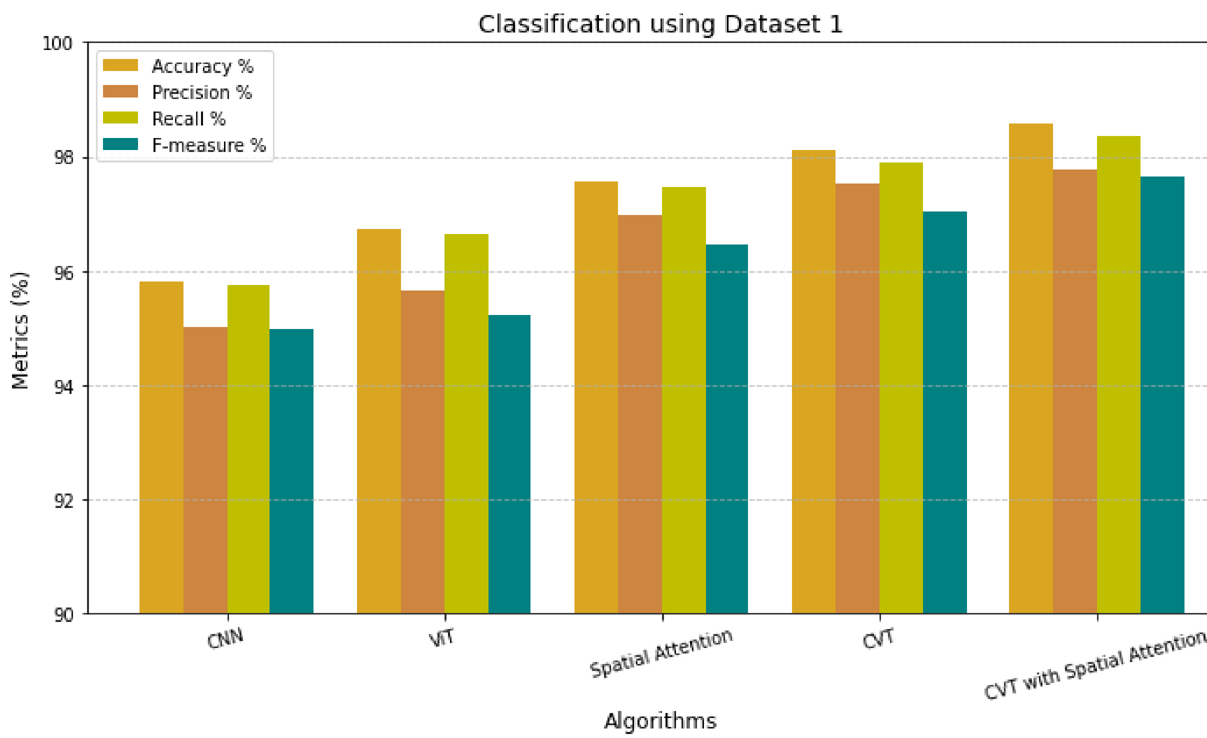

Figure 14: Comparison chart of classification using dataset 1

Figure 14 compares the algorithms like CNN, ViT, SA, CVT and CVT with SA using dataset 1. The proposed CVT with SA outperforms in rice leaf disease detection using dataset 1 compared to the other existing methods in classification. The y-axis of graph shows the metricspercentage, while the x-axis shows the techniques that are currently in use and described in this study.

Table 4: Comparison on feature extraction using dataset 2

Feature Extraction using Dataset 2				
Algorithm /Metrics	**Accuracy %**	**Precision %**	**Recall %**	**F-measure %**
CNN [33]	95.69	94.90	95.29	94.76

ViT [18]	96.87	95.89	96.34	95.75
SA [21]	97.13	96.57	96.89	96.23
CVT [53]	97.87	97.20	97.55	97.11
CVT with SA	98.26	97.78	98.11	97.54

Table 4 compares the performance metrics of the following algorithms, CNN, ViT, SA, CVT and CVT with SA across dataset 2. In terms of accuracy, the presented CVT with SA outperforms the existing approaches for disease detection in rice leaves. During testing on dataset 2, it has achieved 98.26% accuracy.

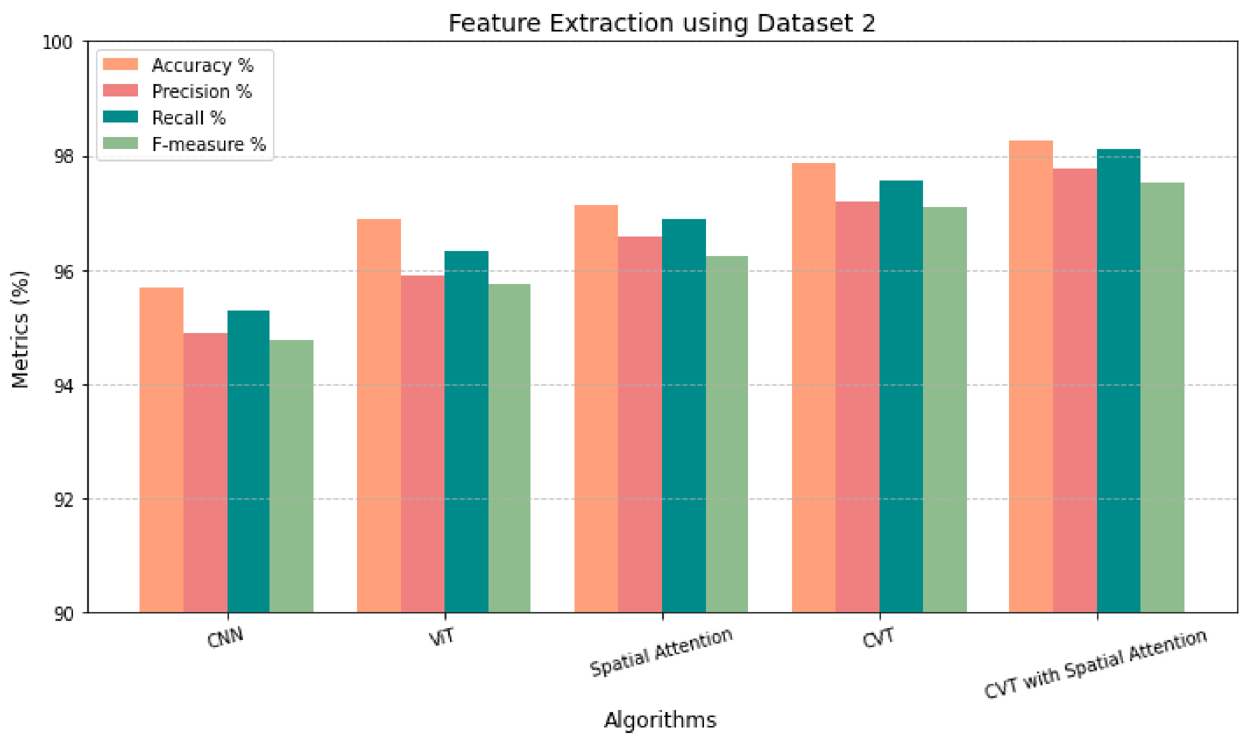

Figure 15: Comparison chart of Feature extraction using dataset 2

Figure 15 depicts CNN, ViT, SA, CVT and CVT with SA together with additional performance metrics such as recall, accuracy and precision for dataset 2. The suggested CVT with SA has outperformed competing feature extraction approaches for disease identification in rice leaves on dataset 2. The y-axis ofgraph shows the metric values, whereas the x-axis shows the techniques that are currently in use and those described in this study.

Table 5: Comparison on feature extraction using dataset 2

Classification using Dataset 2				
Algorithm /Metrics	**Accuracy %**	**Precision %**	**Recall %**	**F-measure %**
CNN [33]	95.87	94.89	95.43	94.56
ViT [18]	96.89	95.35	96.43	95.10
SA [21]	97.87	96.78	97.45	96.53
CVT [53]	98.12	97.46	98.00	97.11
CVT with SA	98.67	97.98	98.45	97.61

Table 5 uses dataset 2 and compares the accuracy, precision, recall, and f-measures of CNN, ViT, SA, CVT and CVT with SA. Regarding disease identification in rice leaves, the suggested CVT with SAhas performed better than the existing methods in terms of accuracy. It achieves 98.67% accuracy by using dataset 2.

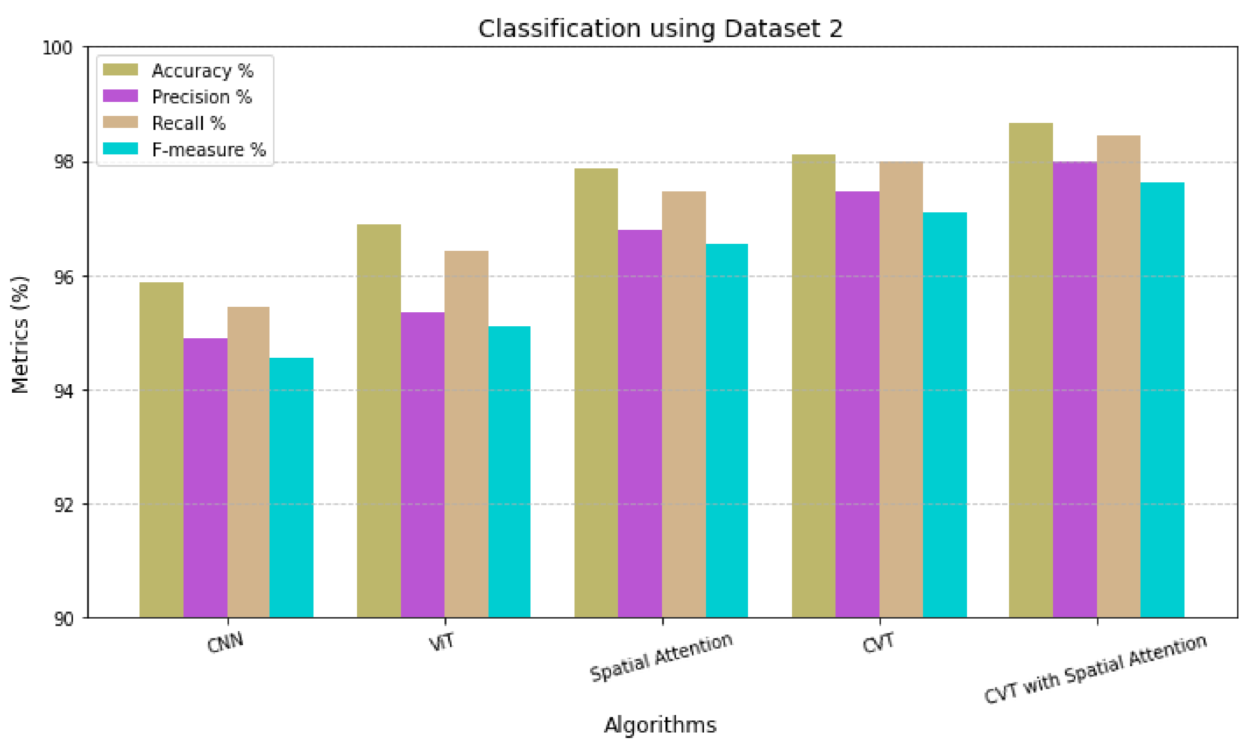

Figure 16: Comparison chart of classification using dataset 2

Figure 16 shows the various performance metrics such as accuracy, precision, recall, and f-measure comparison for CNN, ViT, SA, CVT and CVT with SAusing dataset 2. The proposed CVT with SAhas outperformed in rice leaf disease detection using dataset 2, compared to the other existing methods in classification. In this chart, x-axis shows the various existing and proposed methods used in this paper and the y-axis shows the various metrics for evaluating the performance of algorithms.

VI CONCLUSION

Finally, the proposed Hybrid Convolutional Vision Transformer (CVT) model exhibits high accuracy and reliability on recognizing the various rice leaf diseases. Thismethod outperforms CNN and ViTbased approaches by combining the advantages of ViT for capturing global dependencies and Convolutional Neural Networks (CNN) for local feature extraction. A Spatial Attention module refines the presented model by focusing on disease-prone locations.Hence, classification accuracy and model interpretability are improved. The experimental findings on rice leaf disease dataset show above 98.12% feature extraction accuracy, 98.56% classification accuracy in dataset 1, 98.26% feature extraction accuracy and 98.67% classification accuracy in dataset 2, demonstrate the utility of hybrid CVT in real-world applications. This approach has significant scaling potential for the application in agricultural systems, particularly for drone-based or mobile disease monitoring. This work aims for the future research and expand the dataset to include more rice leaf diseases, enhance the SAmechanism for real-time deployment in automated leaf detection. Furthermore, exploring the hybrid CVT model's adoption in other agricultural domains has lead the way for broader acceptance of AI-powered disease control technologies.

Funding - This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Author Contributions - Pushpa Athisaya Sakila Rani - Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Writing – Original Draft Preparation, Writing – Review & Editing.

Conflict of interest - The authors declare that they have no conflicts of interest.

Availability of Data and Material - The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code Availability - The code used in this study is available from the corresponding author on reasonable request.

Clinical Trial Number - Not applicable.

Consent to Participate - Not applicable.

Consent to Publish - Not applicable.

Ethics Approval - Not applicable.

Corresponding Author- A. Pushpa Athisaya Sakila Rani

Email - pushpasakila@gmail.com

References

1. Aggarwal, S., Suchithra, M., Chandramouli, N., Sarada, M., Verma, A., Vetrithangam, D., ... &Ambachew Adugna, B. (2022). Rice Disease Detection Using Artificial Intelligence and Machine Learning Techniques to Improvise Agro-Business. *Scientific Programming*, *2022*(1), 1757888.
2. Bari, B. S., Islam, M. N., Rashid, M., Hasan, M. J., Razman, M. A. M., Musa, R. M., ... & Majeed, A. P. A. (2021). A real-time approach of diagnosing rice leaf disease using deep learning-based faster R-CNN framework. *PeerJ Computer Science*, *7*, e432.
3. Bhowmik, A. C., Ahad, M. T., Emon, Y. R., Ahmed, F., Song, B., & Li, Y. (2024). A customised Vision Transformer for accurate detection and classification of Java Plum leaf disease. *Smart Agricultural Technology*, *8*, 100500.
4. Brown, D., & De Silva, M. Plant Disease Detection on Multispectral Images using Vision Transformers. In *Proceedings of the 25th Irish Machine Vision and Image Processing Conference (IMVIP), Galway, Ireland* (Vol. 30).
5. Mukherjee, R., Ghosh, A., Chakraborty, C., De, J. N., & Mishra, D. P. (2025). Rice leaf disease identification and classification using machine learning techniques: A comprehensive review. *Engineering Applications of Artificial Intelligence*, *139*, 109639.
6. Deng, J., Yang, C., Huang, K., Lei, L., Ye, J., Zeng, W., ... & Zhang, Y. (2023). Deep-learning-based rice disease and insect pest detection on a mobile phone. *Agronomy*, *13*(8), 2139.
7. Kashyap, B., & Kumar, R. (2021). Sensing methodologies in agriculture for monitoring biotic stress in plants due to pathogens and pests. *Inventions*, *6*(2), 29.

8. 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.
9. Devi, R. S., Kumar, V. R., & Sivakumar, P. (2023). EfficientNetV2 Model for Plant Disease Classification and Pest Recognition. *Computer Systems Science & Engineering*, *45*(2).
10. Ding, B. (2023). LENet: Lightweight and efficient LiDAR semantic segmentation using multi-scale convolution attention. *arXiv preprint arXiv:2301.04275*.
11. Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
12. Fang, L., Wu, Y., Li, Y., Guo, H., Zhang, H., Wang, X., ... & Hou, J. (2021). Using channel and network layer pruning based on deep learning for real-time detection of ginger images. *Agriculture*, *11*(12), 1190.
13. Hasan, M. M., Rahman, T., Uddin, A. S., Galib, S. M., Akhond, M. R., Uddin, M. J., & Hossain, M. A. (2023). Enhancing rice crop management: Disease classification using convolutional neural networks and mobile application integration. *Agriculture*, *13*(8), 1549.
14. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
15. Yang, L., Yu, X., Zhang, S., Long, H., Zhang, H., Xu, S., & Liao, Y. (2023). GoogLeNet based on residual network and attention mechanism identification of rice leaf diseases. *Computers and Electronics in Agriculture*, *204*, 107543.
16. Islam, M. A., Shuvo, M. N. R., Shamsojjaman, M., Hasan, S., Hossain, M. S., & Khatun, T. (2021). An automated convolutional neural network based approach for paddy leaf disease detection. *International Journal of Advanced Computer Science and Applications*, *12*(1).
17. Jain, S., Sahni, R., Khargonkar, T., Gupta, H., Verma, O. P., Sharma, T. K., ... & Kim, H. (2022). Automatic rice disease detection and assistance framework using deep learning and a Chatbot. *Electronics*, *11*(14), 2110.

18. Thakur, P. S., Chaturvedi, S., Khanna, P., Sheorey, T., & Ojha, A. (2023). Vision transformer meets convolutional neural network for plant disease classification. *Ecological Informatics*, *77*, 102245.
19. Pongnumkul, S., Chaovalit, P., &Surasvadi, N. (2015). Applications of smartphone-based sensors in agriculture: a systematic review of research. *Journal of Sensors*, *2015*(1), 195308.
20. Joshi, P., Das, D., Udutalapally, V., Pradhan, M. K., & Misra, S. (2022). Ricebios: Identification of biotic stress in rice crops using edge-as-a-service. *IEEE Sensors Journal*, *22*(5), 4616-4624.
21. Kong, J., Wang, H., Yang, C., Jin, X., Zuo, M., & Zhang, X. (2022). A spatial feature-enhanced attention neural network with high-order pooling representation for application in pest and disease recognition. *Agriculture*, *12*(4), 500.
22. Latif, G., Abdelhamid, S. E., Mallouhy, R. E., Alghazo, J., & Kazimi, Z. A. (2022). Deep learning utilization in agriculture: Detection of rice plant diseases using an improved CNN model. *Plants*, *11*(17), 2230.
23. Mique Jr, E. L., &Palaoag, T. D. (2018, April). Rice pest and disease detection using convolutional neural network. In *Proceedings of the 1st international conference on information science and systems* (pp. 147-151).
24. Lin, S., Xiu, Y., Kong, J., Yang, C., & Zhao, C. (2023). An effective pyramid neural network based on graph-related attentions structure for fine-grained disease and pest identification in intelligent agriculture. *Agriculture*, *13*(3), 567.
25. Ullah, W., Javed, K., Khan, M. A., Alghayadh, F. Y., Bhatt, M. W., Al Naimi, I. S., & Ofori, I. (2024). Efficient identification and classification of apple leaf diseases using lightweight vision transformer (ViT). *Discover Sustainability*, *5*(1), 116.
26. Arinichev, I. V., Polyanskikh, S. V., Volkova, G. V., &Arinicheva, I. V. (2021). Rice fungal diseases recognition using modern computer vision techniques. *International Journal of Fuzzy Logic and Intelligent Systems*, *21*(1), 1-11.
27. Lv, P., Xu, H., Zhang, Y., Zhang, Q., Pan, Q., Qin, Y., ... & Chen, C. (2024). An Improved Multi-Scale Feature Extraction Network for Rice Disease and Pest Recognition. *Insects*, *15*(11), 827.

28. Tabbakh, A., &Barpanda, S. S. (2023). A deep features extraction model based on the transfer learning model and vision transformer “tlmvit” for plant disease classification. *IEEE Access*, *11*, 45377-45392.
29. Jhatial, M. J., Shaikh, R. A., Shaikh, N. A., Rajper, S., Arain, R. H., Chandio, G. H., ... & Shaikh, K. H. (2022). Deep learning-based rice leaf diseases detection using Yolov5. *Sukkur IBA Journal of Computing and Mathematical Sciences*, *6*(1), 49-61.
30. Wijayanto, A. K., Prasetyo, L. B., Hudjimartsu, S. A., Sigit, G., & Hongo, C. (2024). Textural features for BLB disease damage assessment in paddy fields using drone data and machine learning: Enhancing disease detection accuracy. *Smart Agricultural Technology*, *8*, 100498.
31. Liu, T., Chen, W., Wu, W., Sun, C., Guo, W., & Zhu, X. (2016). Detection of aphids in wheat fields using a computer vision technique. *Biosystems Engineering*, *141*, 82-93.
32. Daniya, T., & Vigneshwari, S. (2019). A review on machine learning techniques for rice plant disease detection in agricultural research. *system*, *28*(13), 49-62.
33. Li, D., Wang, R., Xie, C., Liu, L., Zhang, J., Li, R., ... & Liu, W. (2020). A recognition method for rice plant diseases and pests video detection based on deep convolutional neural network. *Sensors*, *20*(3), 578.
34. Jia, L., Wang, T., Chen, Y., Zang, Y., Li, X., Shi, H., & Gao, L. (2023). MobileNet-CA-YOLO: An improved YOLOv7 based on the MobileNetV3 and attention mechanism for Rice pests and diseases detection. *Agriculture*, *13*(7), 1285.
35. Patil, R. R., & Kumar, S. (2022). Rice transformer: A novel integrated management system for controlling rice diseases. *IEEE Access*, *10*, 87698-87714.
36. Mondal, D., Kar, P., Roy, K., Kole, D. K., & Roy, S. K. (2023). X-ResFormer: A Model to Detect Infestation of Pest and Diseases on Crops. *SN Computer Science*, *5*(1), 86.
37. Parez, S., Dilshad, N., Alghamdi, N. S., Alanazi, T. M., & Lee, J. W. (2023). Visual intelligence in precision agriculture: Exploring plant disease detection via efficient vision transformers. *Sensors*, *23*(15), 6949.
38. Mukherjee, R., Ghosh, A., Chakraborty, C., De, J. N., & Mishra, D. P. (2025). Rice leaf disease identification and classification using machine learning techniques: A comprehensive review. *Engineering Applications of Artificial Intelligence*, *139*, 109639.

39. 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.

40. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211-252.

41. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

42. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

43. 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.

44. WAHYUN, W. C., & Sitio, A. S. (2020). Pest Detection Expert System And Method Using Bayes Rice Diseases. *Journal Of Computer Networks, Architecture and High Performance Computing*, *2*(2), 313-319.

45. Wiatowski, T., & Bölcskei, H. (2017). A mathematical theory of deep convolutional neural networks for feature extraction. *IEEE Transactions on Information Theory*, *64*(3), 1845-1866.

46. Pallathadka, H., Ravipati, P., Sajja, G. S., Phasinam, K., Kassanuk, T., Sanchez, D. T., & Prabhu, P. (2022). Application of machine learning techniques in rice leaf disease detection. *Materials Today: Proceedings*, *51*, 2277-2280.

47. Vasantha, S. V., Kiranmai, B., & Krishna, S. R. (2021). Techniques for rice leaf disease detection using machine learning algorithms. *Int. J. Eng. Res. Technol*, *9*(8), 162-166.

48. 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.

49. Chung, C. L., Huang, K. J., Chen, S. Y., Lai, M. H., Chen, Y. C., & Kuo, Y. F. (2016). Detecting Bakanae disease in rice seedlings by machine vision. *Computers and electronics in agriculture*, *121*, 404-411.
50. 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.
51. Salamai, A. A., Ajabnoor, N., Khalid, W. E., Ali, M. M., &Murayr, A. A. (2023). Lesion-aware visual transformer network for Paddy diseases detection in precision agriculture. *European Journal of Agronomy*, *148*, 126884.
52. Radenović, F., Tolias, G., & Chum, O. (2018). Fine-tuning CNN image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, *41*(7), 1655-1668.
53. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22-31).