Expert Systems With Applications 260 (2025) 125385

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

SoyaTrans: A novel transformer model for fine-grained visual classification of soybean leaf disease diagnosis

Vivek Sharma [a], Ashish Kumar Tripathi [a,*], Himanshu Mittal [b], Lewis Nkenyereye [c]

[a] Malaviya National Institute of Technology (MNIT), Jaipur, India
[b] Indira Gandhi Delhi Technical University For Women, Delhi, India
[c] Sejong University, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Soybean plant disease
Convolutional neural network
Random shifting
Swin transformer

ABSTRACT

Plant leaf disease detection has a considerable influence on the safety of crop quality. However, distinguishing different symptoms in leaves is a challenging task. Newfangled CNN architectures have a "moving window" feature that extracts local information of the image only and fails to capture global features. Additionally, CNN architectures take the whole image as an input which lacks in identifying the small lesions that result in poor classification. Therefore, a novel robust model, the SoyaTrans network, is designed by pairing CNN architecture with swin transformers which efficiently work on real field images. In this, a new random shifting is introduced in comparison to cyclic shift which enhances the classification performance while reducing the computational complexity. Moreover, the proposed SoyaTrans model ensembles the capabilities of conventional CNN with a swin transformer network that efficiently detects diseases on different types of crops. Furthermore, this paper presents a new soybean plant leaf disease dataset that is collected from real fields to overcome the challenge of a limited soybean leaf dataset. Experimental results of the proposed model are compared against the ten state-of-the-art methods in terms of five parameters, namely parameters, accuracy, precision, recall, and F1-score. In addition, the efficacy of the proposed model is validated on four publicly available datasets namely, Embrapa, Plant Village, AI2018, and PlantDoc. The proposed model surpassed all the ten state-of-the-art models, even under complicated backdrops, with an accuracy of 98.00%, 97.00%, 76.00%, and 92.00% on plantvillage, AI2018, PlantDoc, and Embrapa dataset with the least computational complexity of 5.2 million parameters. Lastly, the proposed model achieved 94.00% accuracy on the newly presented soybean leaf dataset.

1. Introduction

Soybean (Glycine max L. Merrill) is one of the imperative grain and edible oil crop produced throughout the world. This crop holds a prominent position of the world's most important seed legumes, contributing 25% of the global oil production (Agarwal, Billore, Sharma, Dupare, & Srivastava, 2013). Among the world's largest producers of soybeans, India ranks fifth after the United States, Brazil, Argentina, and China. Over two decades, soybeans have played a pivotal role in this scenario and will continue to do so as per the statistics. According to the Federation of Indian Chambers of Commerce & Industry (FICCI), 89% of the total soybean production in India comes from regions of Maharashtra and Madhya Pradesh. As per the report, the yield per hectare of the soybean crop in Madhya Pradesh region is 62.605 and produces 59.475 lakh MT. However, in the last few years, the average crop production has decreased due to the occurrence of severe diseases which are taking a great toll on yield during the growing process of soybean crop. Different types of diseases in soybean crop sorely affect the quality and production of food, biofuel, and fiber in the crop. These losses are severely incessant or disastrous. Therefore, it is of great significance that these diseases have severely impacted the quality and quantity of soybean. Moreover, timely treatment and effective identification of soybean diseases should be carried out to protect the crops from several diseases and ensure the sustainability of agroecosystems. The emergence of cutting-edge technologies, including artificial intelligence, machine learning, and deep learning, has significantly propelled advancements in the realm of crop disease detection. In traditional approaches, feature selection is done manually or relies on trained experts, which leads to improper identification, time taking, and high cost. With the rapid development of technologies automatic disease detection has witnessed significant performance using deep learning models. In the new era of technology, researchers did a comprehensive evaluation of different deep learning models widely used in several applications (Kaur et al., 2022; Liu, Wang, Wang, Chen, &

* Corresponding author.
E-mail address: ashish.cse@mnit.ac.in (A.K. Tripathi).

https://doi.org/10.1016/j.eswa.2024.125385
Received 8 November 2023; Received in revised form 28 August 2024; Accepted 10 September 2024
Available online 12 September 2024

Gao, 2023; Mostafa et al., 2022; Paymode & Malode, 2022; Turkoglu, Yanikoğlu, & Hanbay, 2022; Zhao, Sun, Xu & Chen, 2022), for the automated diagnosis of crop leaf disease. The specific characteristics of these architectures are to reduce the model complexity and increase accuracy. In recent years, convolutional neural networks (CNN) have revolutionized the computer vision field, which has inspired new approaches to identify early plant leaf diseases for intelligent agriculture. Despite the success of the aforementioned techniques, different issues still remain such as an imbalanced dataset, over-fitting problem, and lack of training images. Furthermore, the deep learning models are data hungry which requires a massive amount of data, while lightweight models fail in discriminating the efficient lesion features for all types of plant disease. As a result of its common convolution kernel parameters, redundant computations are eliminated and efficiency is enhanced. Besides this, the CNN convolution kernels have a "moving window" feature which is limited to extracting only the local information of the image and fails to capture global features. Therefore, aiming at the above issues, the introduction of the transformer has opened up new dimensions in the field of image identification (Dosovitskiy et al., 2020), image segmentation (Wu et al., 2022), and object detection (Li, Li, Qiao & Zhang, 2022). In the current scenario, the vision transformer (ViT) has achieved significant classification performance containing global information of the image and also performs better on irregular inputs. However, it requires a massive amount of data to converge as ViT emphasizes long-distance feature dependencies and fails to capture local features of images efficiently. On the same footprints, a novel shifted window hierarchical swin transformer (ST) is proposed (Liu et al., 2021). This transformer significantly simplifies and minimizes vision-related complexities. However, it fails to achieve the desired results in the images that are captured in the real fields with complicated backgrounds and noise. Therefore, a novel robust network of soyatrans is designed, by pairing CNN with swin transformers which efficiently work on real field images. In addition, a novel random shifting is proposed, to ensure attention with a distinct image position on each epoch. In addition, random shifting also reduces the overall computational complexity as compared to cyclic shifting. Further, adequate training images are required to achieve high generalization capabilities for transformer models. Insufficient training images have adverse effects on the identification accuracy of soybean leaf diseases. To address these challenges, real-field soybean images are collected from the real field. However, the dataset collected is limited in size. Therefore, common data augmentation techniques were employed for data size expansion such as flipping, zoom, rotation, and translation which improves the classification performance. To the best of our knowledge, this marks the initial implementation of the concept of random shifting in leaf disease identification.

The paper's collective contributions are succinctly summarized as follows:

- A novel Swin-enabled model with random window shifts has been developed for the efficient classification of crop leaf diseases. The computational cost is optimized through a reduction in the parameter count.
- A novel shifting criteria, random shifting is proposed which improves the classification performance over the original swin transformer cyclic shifting.
- An annotated soybean curated dataset is presented which is collected from different natural scenes, under complicated backdrops.
- The performance of the proposed SoyaTrans has been validated on the real field collected dataset as well as on four public datasets namely, Embrapa, Plant Village, AI2018, and PlantDoc-dataset

Finally, the paper is structured in different Sections Section 2 provides an in-depth exploration of related work in the field. Section 3 delineates the proposed model, presenting its key components and architecture. Section 4 comprehensively discusses the results obtained from simulations, offering insights and analysis. Lastly, Section 5 encapsulates the paper's conclusions, summarizing the findings and their implications.

2. Related work

This section demonstrates the review of different recent studies of CNN, and transformer networks in plant leaf diagnosis. These studies employ both types of networks, as elaborated in the following sections.

2.1. Convolutional neural network & attention based methods

With the advancement of deep learning techniques, numerous research works have been introduced in the literature that exhibit excellent performance in the plant disease diagnosis task. Owing to their evolution over time, convolutional neural networks (CNNs) have garnered considerable prominence in numerous domains. Therefore, many proficient researchers employ CNNs as the backbone network architecture, subsequently optimizing the model to enhance its suitability for the precise detection of plant diseases. Besides this, there are instances where background characteristics exhibit more prominent features than the leaf and diseased regions in the foreground, significantly diminishing the model's performance (Ferentinos, 2018).

Furthermore, integrating the attention mechanism into CNN, the model will focus on salient attributes, rather than predominantly emphasizing global features (Woo, Park, Lee, & Kweon, 2018). In a pioneering study, Zeng and Li developed a streamlined CNN architecture by incorporating a self-attention module for diverse crop leaf disease identification tasks and their model achieved 95.33% classification accuracy (Zeng & Li, 2020). In another network, Gao et al. devised an efficient dual-branch channel attention mechanism specifically tailored for crop disease detection. The proposed model has been verified on different datasets namely the PlantVillage dataset, AI Challenger 2018 dataset, and real field cucumber dataset (Gao, Wang, Feng, Li, & Wu, 2021). Yu et al. introduced a novel attention method for the early identification of leaf spots in apple leaf disease . In this two subnetworks are developed namely feature segmentation which provides detailed insights into feature maps for a particular spot and leaf area. Another one is a spot-aware network which is used to increase classification accuracy (Yu & Son, 2020). For instance, Qian et al. proposed CNN based self-attention network for early identification of maize leaf disease lesion information in real field images (Qian, Zhang, Chen, & Li, 2022). Pandey et al. developed a deep attention CNN model for efficient classification of 10,851 real-world RGB leaf images of 17 plant species captured from smartphones (Pandey & Jain, 2022). Zeng et al. deployed deep CNN and developed a network with the integration of multi-scale dilated convolution and cross-scale attention for rubber crop disease detection (Zeng et al., 2022). Zhao et al. have employed a channel-based attention module (CBAM)with residual and inception structure for better classification of tomato, corn, and potato leaf disease (Zhao, Sun, Xu & Chen, 2022). In yet another network, Zhao et al. introduced an architecture with the fusion of dual transfer learning with ResNet50 and CBAM for the classification of 4 different datasets namely, cucumber, tomato, AI challenger 2018, and self-built dataset (Zhao, Li, Li, Ma & Zhang, 2022).

2.2. Transformer based methods

The initial investigation on crop leaf identification involves the application of CNN and transfer learning methodologies. The introduction of transformers in natural language processing offers new vistas in the field of image processing (Vaswani et al., 2017). Li et al., developed a lightweight hybrid model SLViT using shuffle convolution, and vision

transformers (ViT). The efficacy of the proposed model has been evaluated initially on publicly available datasets and then applied to the real field of sugarcane leaf disease for efficient classification (Li et al., 2022). Li et al. proposed a novel vision transformer-based architecture ConViT for kiwifruit disease identification in a real-time environment. In this transformer network, is used to identify the local features, and convolution operations are used for global feature extraction which helps in efficient leaf disease classification (Li, Chen, Yang & Li, 2022). Yu et al. proposed a novel transformer, Mix-ViT with the combination of ultrafine-grained visual categorization with vision transformers which enhances the classification performance (Yu, Wang, Zhao, & Gao, 2022). Patil et al. developed a novel rice transformer model that uses a cross-attention mechanism for efficient rice disease classification of the samples collected in a real-time environment (Patil & Kumar, 2022).

Despite, the success of vision transformers, it requires a massive amount of datasets to be trained to get efficient results. To overcome the challenge of fewer datasets and for better visualization swin transformer has been the game changer in the field of computer vision. Wang et al. designed a novel method that uses an improved swin transformer as the backbone architecture for the identification of cucumber leaf spots in complex backgrounds (Wang et al., 2022). Guo et al. developed a novel architecture convolutional swin transformer (CST) for the identification of degrees and types of diseases of different crops namely cucumber, banana, potato, and tomato Guo, Lan, and Chen (2022). Bi et al. proposed an improved swin transformer that focuses on multi-scale features to accurately identify the different corn seed varieties (Bi et al., 2022). Zheng et al., presented a model Swin-MLP, where swin transformer is the backbone architecture combining multi-layer perceptron (MLP) for the appearance quality recognition of strawberries (Zheng, Wang, & Li, 2022). Chang et al. developed a novel hybrid method which identifies the fine edge features of plant leaf diseases (Chang, Wang, Zhao, Li, & Yuan, 2024). It is clear from the aforementioned literature that the integration of CNN with a transformer can be an alternative approach for efficient model development in this field (Jin, Chu, Qi, Feng, & Mu, 2024; Pacal, 2024). However, the classification accuracy is affected on a small dataset. To address this concern and enhance interpret-ability, it is crucial to design a model that demonstrates superior performance across diverse datasets. Therefore, a novel Swin-enabled CNN model, SoyaTrans is developed which significantly enhances the classification of leaf disease across diverse datasets. Furthermore, the model architecture is detailed in subsequent sections.

3. Proposed model

This section outlines the methodological approach employed for the development of the model. In this, a novel Swin-enabled CNN model, SoyaTrans, is developed for early disease detection. In addition, a brief introduction of the original swin transformer is presented initially for comprehensive understanding.

3.1. Swin transformer based architecture

Recent advancements in transformers within the realm of deep learning have sparked a revolution, ushering in new dimensions in the fields of image processing and computer vision (Dosovitskiy et al., 2020). Liu et al. introduced a novel hierarchical shifted window transformer network that significantly reduces the computational complexity, and improves the classification performance. The Swin architecture consists of patch partition, linear embedding, patch merging layer, and swin transformer block in different stages. In a patch partition, the RGB image is split into non-overlapping patches or tokens. Moreover, a linear embedding is executed on the raw features to project them into an arbitrary dimension C. In addition, the transformer block maintains the number of tokens, which the combination with linear embedding is denoted as stage 1. In the 2 stages, the block contains a pair of patch merging and swin transformer blocks. This process is repeated two times referred to as stage 3, and stage 4 respectively. Swin transformer block (SWB) comes in pairs, containing layer normalization (LN), and multi-layer perception (MLP). The first SWB contains window-based multi-head self-attention (W-MSA), with GELu as non-linearity between the layers. The Layernorm is used before W-MSA, and the MLP module and residual connections are applied after every module. Whereas the second (SWB) contains shifted window-based multi-head self-attention (SW-MSA), and Layernorm is used before WMSA and MLP module (Liu et al., 2021). In an original transformer, global self-attention is performed which shows the relationship between different tokens computed overall. However, this attention mechanism results in quadratic complexity in relation to a number of tokens, resulting in being unsuitable for different computer vision problems where a large set of tokens is required for dense predictions or to characterize the high-resolution images. Moreover, self-attention within the local window is computed for efficient modeling. Here, self-attention is arranged in a non-overlapping manner. For instance, if a window consists of $M \times M$ patches. The global multi-head self-attention (MSA) and window-based self-attention on images of $h \times w$ patches are stated in equations.

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{1}$$

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hwC \tag{2}$$

Where the former is quadratic to patch number hw, and the latter is linear when M is fixed (set to 7 by default). In this, the computations of window-based self-attention are scalable, whereas, computations of global MSA are commonly outrageous for large hw. Besides, the W-BSA module lacks connection across the window, limiting its modeling power. To mitigate the above challenge, a novel shifted window partitioning strategy has been introduced which switches between two partitioning configurations in successive Swin Transformer blocks while preserving the effectiveness of non-overlapping window computation.

To enhance interactivity with other windows, an attention network based on shifted windows is introduced in this study. Here a particular design study is followed. Firstly, the images is divided into nine parts, as illustrated in Fig. 1. Subsequently, region A is allocated to these segments. Similarly, regions B and C are relocated using the same method. Finally, the reconstituted images are partitioned into four parts for the computation of local attention.

Furthermore, residual learning was devised to streamline network training processes and facilitate deeper network architectures. The residual learning of the shifting window partitioning method computing consecutive Swin Transformer blocks as

$$\begin{aligned} \hat{z}^l &= W-MSA(LN(z^{l-1})) + z^{l-1} \\ z^l &= MLP(LN(\hat{z}^l)) + \hat{z}^l \end{aligned} \tag{3}$$

$$\begin{aligned} \hat{z}^{l+1} &= SW-MSA(LN(z^l)) + z^l \\ z^{l+1} &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \tag{4}$$

In this z is the feature map of each layer after processing, where $\hat{z}^l$ and zl denote the output features of the (S)WMSA module and the MLP module for block l, respectively. While W-MSA and SW-MSA denote window-based multi-head self-attention using regular and shifted window partitioning configurations, respectively.

3.2. Framework of the proposed SoyaTrans

In this study, a groundbreaking model for early disease identification of soybean leaves is proposed. The proposed SoyaTrans model combines the capabilities of Swin transformer (Liu et al., 2021) and CNN for early plant leaf identification. In this network, convolution blocks are used to extract the local features of the images, whereas, transformers blocks are used to extract the global features. This strategic design greatly enhances the ability to diagnose intricate deep lesion features

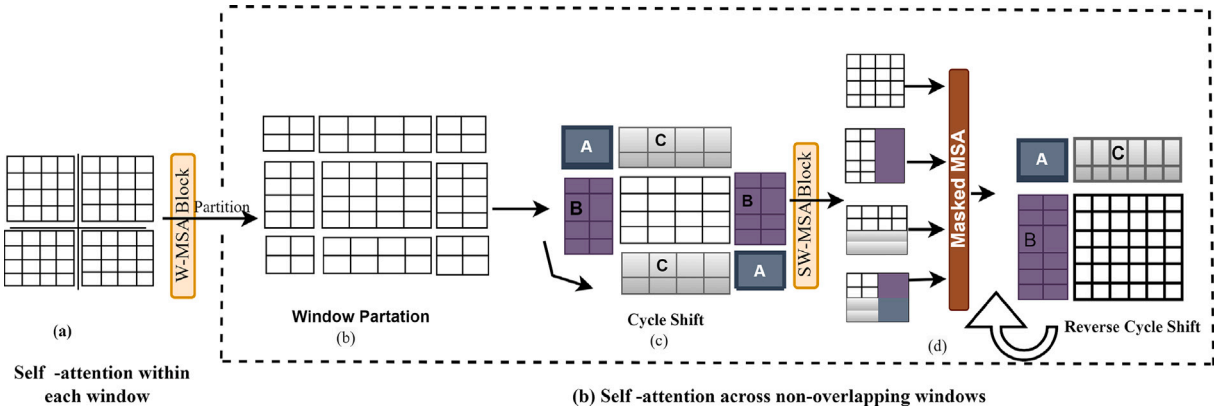

Fig. 1. Illustration of W-MSA & SW-MSA.

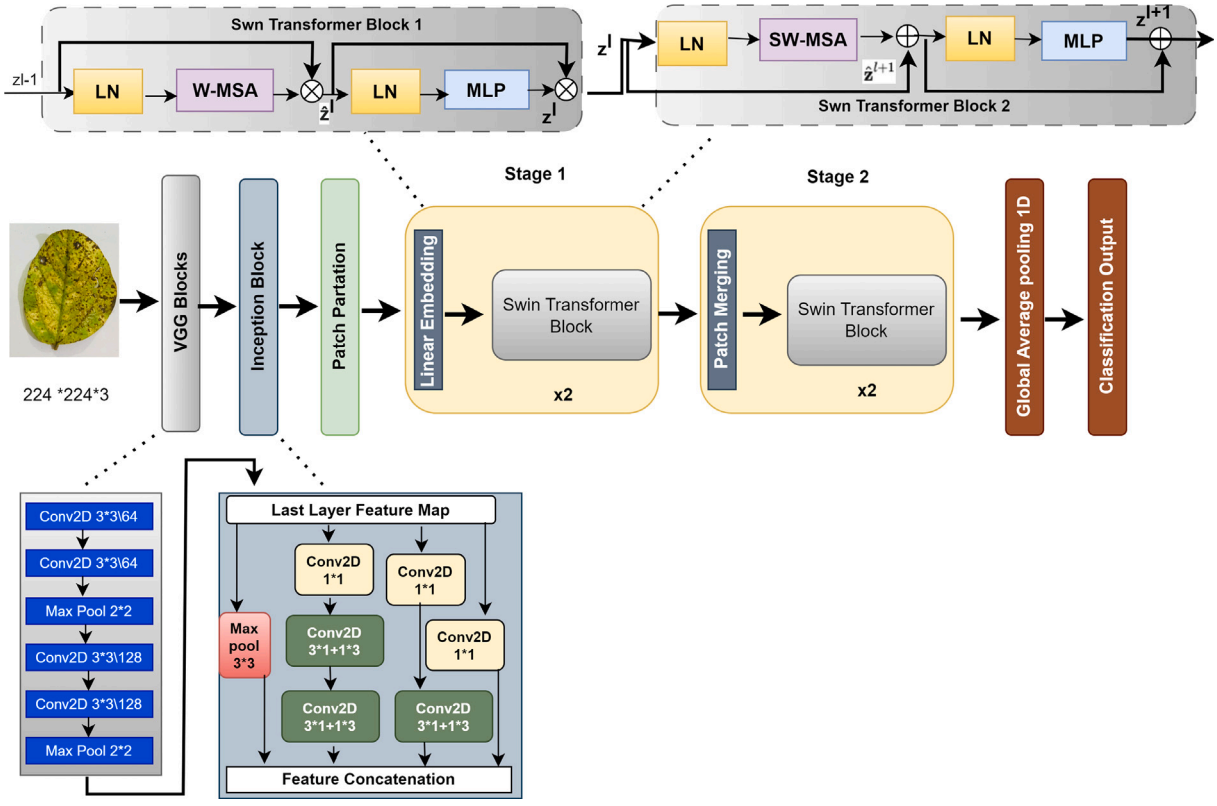

Fig. 2. Architectural outline of the Proposed Network.

associated with plant diseases. The architectural outline of the SoyaTrans is illustrated in Fig. 2. As demonstrated in the figure, the SoyaTrans network consists of VGG16, Inception v7, and Swin transformer components such as patch partition, linear embedding, two blocks containing patch merging, and swin transformer blocks incorporating LN, W-MSA, SW-MSA, and MLP. The SoyaTrans model is strengthened with the pre-trained networks such as VGG16 which performs better initialization of parameters. Whereas, the Inception v7 block enhances feature extraction by incorporating diverse convolutional filter sizes.

Furthermore, the swin transformer block with W-MSA and SW-MSA comprises two consecutive blocks crucial for extracting profound salient features within the patches. The SoyaTrans model takes an input image of 224 × 224. This image is passed into the VGG16 network two convolution blocks with each block containing convolution layers, and max pooling layer resulting in the 56 × 56 × 128 output. This output is further fed into the Inception v7 Conv block containing a multi-level feature extraction block. The multi-level feature extraction block improves the learnability of local features of the model and generates an output of size 56 × 56 × 512 after concatenating the feature maps generated by different Convolution layers. Further, the feature map undergoes a transformation into non-overlapping patches through the patch partition module. Here, patches are represented as tokens, and features are concatenations of RGB pixels. In the proposed network patch size is taken as 4 × 4, and the dimension of features in each patch is $4 \times 4 \times 3 = 48$. In addition, this is fed into a stack of two swin transformer blocks for feature extraction. In the first block, a linear embedding layer is enforced on the raw-valued feature which projects to an arbitrary dimension value denoted as C (embedding dimension). Further, these are passed through, two successive swin transformer blocks incorporating W-MSA and SW-MSA. Here, the first block *i.e.* Window based multiheaded self-attention uses a regular window partitioning strategy. The second swin transformer block contains Shifted window-based multiheaded self-attention for better interaction with other windows. Further masking is performed to each sub-area

because non-adjacent sub-areas do not exchange information during self-attention is denoted as stage 1. In stage 2, the patch merging layer is enforced to reduce the no of tokens as the network goes deeper which transforms the features, in addition to the swin transformer block of stage 1 capturing the long-term dependencies. Lastly, the global average pooling layer is incorporated which converts the final output of the Swin transformer block into a 1-D dimensional vector. Here softmax activation function is used as the dataset contains multiple classes. The pseudo-code of the proposed SoyaNet is depicted in 1.

Algorithm 1 :Proposed Model

```
Input: Dataset
Output: Optimal Classification
Set N_epochs as Number of Epochs
for epoch = 1,2,...,N_epochs do
    Set previous state ← image pixels position
    for batch = 1,2,....T do
        Collect x ∈ P(x) size(224, 224, 3) from dataset to train the model
        current state ← previous state
        window size ← 7x7
        array b have (224,224) position ← 50176
        n_windows ← 1024
        for window =1,2,....n_windows do
            array_c ← elements of current window
            array_b ← array_b -array_c
            array_d ← random 49(7X7) elements from remaining array_b to select window
            elements and do attention
            array_b ← array_b - array_d
            array_b ← array_b + array_c
        end for
    end for
end for
Save model after training
```

3.3. Shifting operation comparisons

In this shifting operation, consider a 4×4 image grid where positions are numbered from 1 to 16. In the context of W-MSA, attention is performed between every pixel within each window. There are 4 windows of size 2×2. Consider the set S containing the elements $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16$. In the first window, choose 4 pixels from the remaining positions: $3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16$ since attention has already been applied to the pixels in the window with positions $1, 2, 5, 6$. Consequently, the first window will consist of pixels $3, 9, 13, 15$. In the second window, select 4 pixels from the remaining 9-pixel positions $1, 2, 5, 6, 10, 11, 12, 14, 16$ as attention is already happened in $3, 4, 7, 8$ and pixels positions $3, 9, 13, 15$ are already fixed. Therefore, the second window will consist of pixels $2, 5, 11, 14$. Likewise, for the third window, choose 4 pixels from the remaining 7 positions: $1, 4, 6, 7, 8, 12, 16$ because attention already happened in 9,10,13,14 and pixels positions $2, 5, 11, 14$ $3, 9, 13, 15$ are already fixed. Therefore, the third window will consist of pixels $16, 1, 4, 12$. Finally, for the fourth window, choose four pixels from the remaining 4 positions: $6, 7, 8, 10$. Since, attention has already been applied to the pixels in the previous windows $11, 12, 15, 16$, and the positions $2, 5, 11, 14$, $3, 9, 13, 15$, and $16, 1, 4, 12$ have already been determined. Therefore, the fourth window will consist of pixels $6, 8, 10, 7$. Furthermore, in comparison to cyclic shifting *i.e.* the original shift, the proposed random shifting performs better. Fig. 3 illustrates the proposed shifting operations in different steps numerically.

In the case of random shifting, it is demonstrated that attention occurs with a unique image position each time. Further, using probability it is evident that the likelihood of the same window configuration occurring during random shifting is calculated as the product of decreasing probabilities $1/49 \times 1/48 \cdots \times 1/1$, which approaches an extremely small value close to 0. Therefore, random shifting serves as a viable alternative to cyclic shifting, especially considering the 7×7 window size, encompassing 49 positions. As every time attention occurs in distinct window elements. Hence, the overall computational complexity will be lower in random shifting compared to cyclic shifting. Therefore, this results in the reduction of overall computational complexity. Fig. 4 and Fig. 5 depict the base and proposed shifting criteria using leaf image. Further, the configuration details of the proposed model is depicted in Table 1.

3.4. Label smoothing cross entropy

In this study, label smoothing is used as a regularization technique, which outperforms well on both the training and testing datasets for better classification (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). In general, label smoothing is a technique that converts hard labels into soft labels, which improves network optimization, of confident training samples, and enhances the generalization ability of the model. The mathematical formulation of label smoothing is depicted as

$$I_k^{LS} = (1 - \alpha) I_k + \frac{\alpha}{K} \tag{5}$$

Where K represents the different labeled classes, alpha shows the hyperparameter value ranging between 0–1 which influences the smoothing level. Furthermore, techniques are applied to the crop disease dataset consisting of multi-labeled noise in the soybean leaf disease dataset resulting in a decrement in model performance. To improve the classification performance label smoothing cross-entropy(LSCE) loss function is applied to enhance the network accuracy (Müller, Kornblith, & Hinton, 2019).

The mathematical formulation of the LSCE loss function which is articulated as follows:

$$L = -\frac{1}{N} \sum_{i=1}^{N} I_k^{LS} . log(p_i) \tag{6}$$

In this, I_k, is the one-hot encoded sample of sample i representing 0 for negative, and 1 for positive class. Whereas, p_i is the probability of sample i which predicts the positive class. Conclusively, applying cross-entropy not only enhances the network performance in multi-labeled noise images but also prevents the model from over-fitting during the model training. Moreover, the fusion of cross-entropy and label-smoothing regularization techniques not only streamlines computational complexity but also enhances the accuracy of classification. In this study, the LSCE loss function has been incorporated to compute the final output which is defined in Eq. (6).

4. Performance analysis

In this section, the efficacy of the proposed SoyaTrans model has been evaluated on real-field soybean collected images. To assess the efficacy of the proposed model, it is also applied on four publicly available benchmark datasets namely plant village, AI challenge dataset, PlantDoc, and Embarapa dataset. In addition, the proposed SoyaTrans model is compared against ten state-of-the-art CNN and transformer-based models. These include different hybrid CNN & transformer-based, and transformer-based models. Further, the model performance is evaluated on five performance statistics. For better visualization, the proposed model is also validated using T-SNE, confusion matrix, and Local Interpretable Model Agnostic Explanation (LIME) and Grad-CAMs (gradient weighted class activation maps), and Grad-CAMs (gradient weighted class activation maps).

4.1. Experiment setup

This section demonstrates the experimental configuration employed in this study. The experiments were conducted on an Ubuntu 20.04 server equipped with a CPU E5-2650 of the fourth version, featuring 2.20 GHz and 16 cores. The system was enhanced with a P100 PCI-E GPU boasting 4000 CUDA cores and 16 GB of memory, manufactured

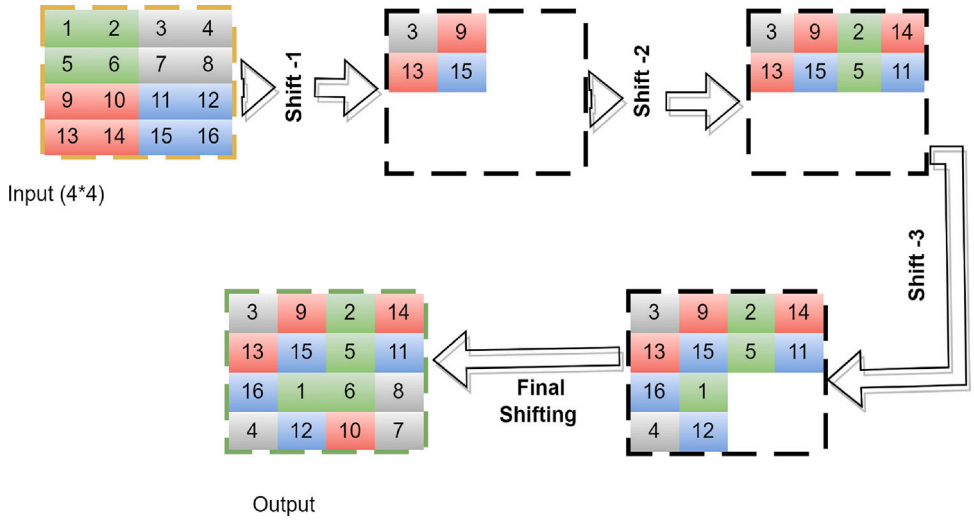

Fig. 3. Proposed Shifting operations.

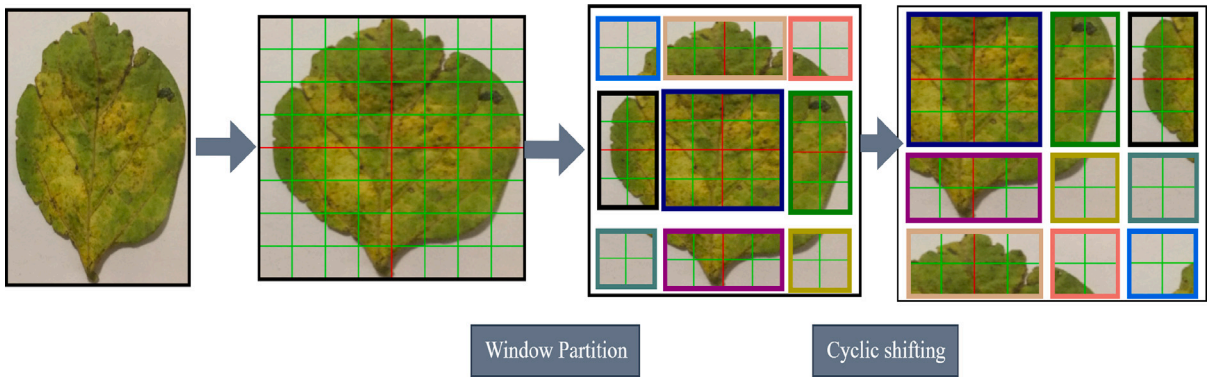

Fig. 4. Swin base leaf cyclic shifting.

Table 1
Layered description of the proposed model.

Type of Layer	Output Vector	Configurations
Input Layer	(None, 224, 224, 3)	Image
VGG Block	(None, 56, 56, 128)	Conv2D (3 × 3) (stride-1), BN, ReLU, MaxPooling2D (2 × 2) (stride-2)
Inception Block	(None, 56, 56, 512)	Conv2D (diff-diff kernel size and stride), BN, ReLU, MaxPooling2D (2 × 2) (stride-1)
Stage 1	(None, 14, 14, 512)	Conv2D (4 × 4) (stride-4), LayerNormalisation, W-MSA, MLP, Proposedshifting-MSA
Stage 2	(None, 7, 7, 512)	LayerNormalisation, W-MSA, MLP, Proposedshifting-MSA
Classification	(None, 6)	LayerNormalisation, AvgPooling, Linear

by NVIDIA. Additionally, the PyTorch platform was utilized for the implementation of the models.

In addition, for building the network different hyperparameters are used which significantly enhances the model performance. This involves the activation function, loss function, fine-tuning the learning rate, dropout rate, batch size, and optimizer. In this study, grid search technique is applied which find the best hyperparameters set from the search space of probability range from 0.3 to 0.7, and learning rate range from 0.0001 to 0.1. Furthermore, different optimizers SGD (Stochastic Gradient Descent), Rmsprop (Root Mean Square Propagation), and Ada-grad has been applied.

Furthermore, the results demonstrates that the different optimal parameters were taken into the account. To improve the efficacy of the model, batch size '32' is selected. This is due to the fact that increased batch size affects classification accuracy resulting in the

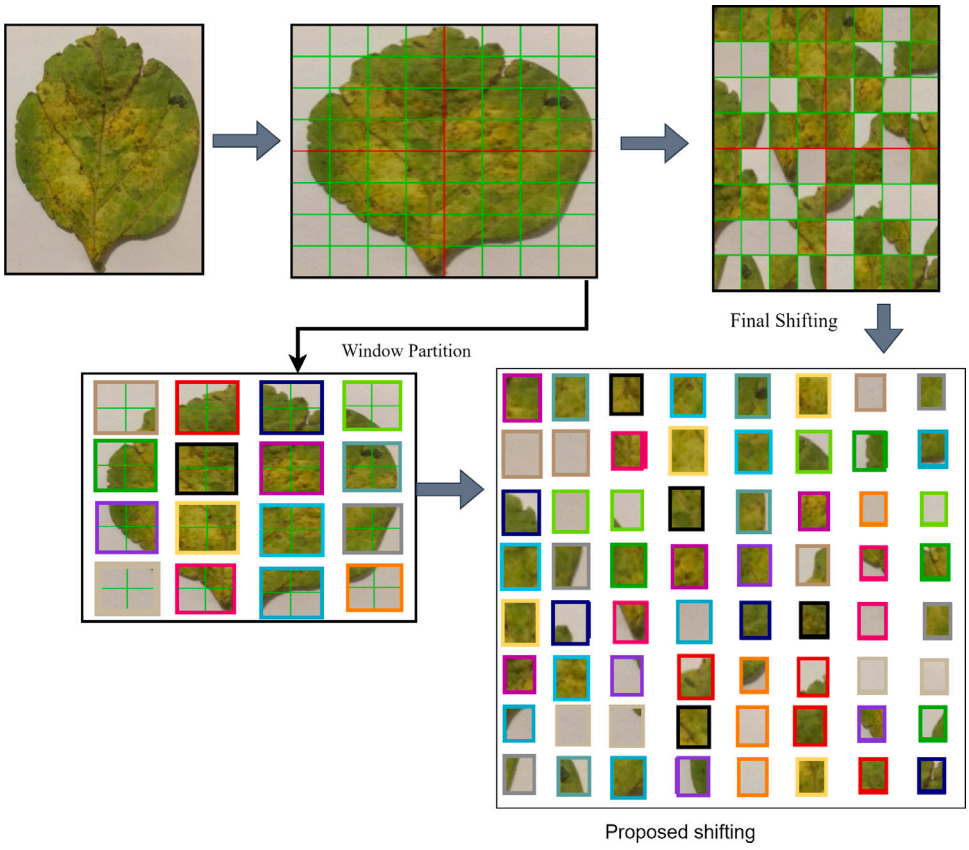

Fig. 5. Proposed leaf random shifting.

Table 2
Hyperparameter Constants.

Type of Layer	Search Space	Optimal Parameter
Learning Rate	[0.0001, 0.001, 0.01, 0.1]	0.0001
Optimizer	[SGD, RmsProp, Adam]	Ada-grad
Batch Size	[32, 64]	32
Dropout Rate	0.3, 0.4, 0.5,0.6, 0.7	0.5
Image Size	224 × 224	224 × 224
Activation Function	Relu, Leaky ReLU	Leaky ReLU
Loss Function	Label smoothing	Label smoothing cross entropy

decrement of learning rate. Besides this, the aforementioned parameters of the different models adopted for the experimentation study are set to default. In the training procedure, the training dataset is subdivided into batches and systematically processed in iterations. The frequency of weight updates in the neural network iterations escalates in correspondence with the advancement of epochs. As the number of epochs increases, the training curve gradually converges towards the optimal state, reaching a point of potential overfitting by the 50th epoch. Hence, we set the number of epochs at 50, maintaining a fixed learning rate of 0.0001, and automatic model saving during training. In addition to this, models are trained on different learning rate, which deteriorated the model performance. To improve the effectiveness of the model different optimizer were tested such as SGD, RmsProp, and Adam. The Adam optimizer outperforms on better in different datasets. Therefore, Adam optimizer is chosen, by its ability to handle non-stationary targets and sparse gradients. Moreover, there is a dynamic adjustment in the learning rate during model training and shown their better convergence at 0.0001. Additionally, sparse categorical cross-entropy is employed for the classification of multiple classes. Besides, this dropout rate is applied in every dropout layer with probability values ranging from 0.3 to 0.7. To mitigate over-fitting and enhance model performance, a dropout rate of 0.5 is chosen. Conclusively, Leaky Relu activation function followed by batch norm is applied after each layer. The integration of these techniques results in improved classification. During the evolution of the model, various combinations are explored to determine the optimal selection. Further, Sparse categorical cross-entropy is applied for classifying the multiple class. Conclusively, when these are amalgamated, classification results are improved. Throughout the model evolution, several combinations are applied for the optimal selection. Table 2 demonstrate the different hyperparameters considered during the model training.

4.2. Dataset description

In this study, the soybean leaf dataset comprises a of total 2890 raw images which were captured between August 2021 & 2022. The images were captured around 11:30 a.m. and 3:30 p.m. during ideal illumination and mild temperatures for data curation. The total number of raw photographs collected is uploaded from the Sony cyber shot camera & smartphone to the computer, and diagnosed into 06 different categories. During the data collection process, the leaves are not plucked out in the growing state. This is due to the fact that in its initial plugging, it affects the overall growth of the plant leaves. Soybean leaf consist of four growing stages to reach the flowering namely the initial pod, and full pod, initial stage to the full stage of seeding. Following image acquisition, a meticulous manual inspection was conducted to eliminate duplicates, low-quality, and damaged images, as well as those rendered unrecognizable due to severe disease attacks. Finally, we were left with 2500 images ready. At the outset, the captured images have dimensions of 4000 × 3000 × 3 pixels. Handling these size images is extremely difficult and necessitates a properly geared computing facility. Furthermore, an initial step involves cropping certain sections of the image backgrounds, resulting in images of size 3000 × 3000 × 3 pixels. Subsequently, the images are resized to dimensions of 224 × 224 × 3

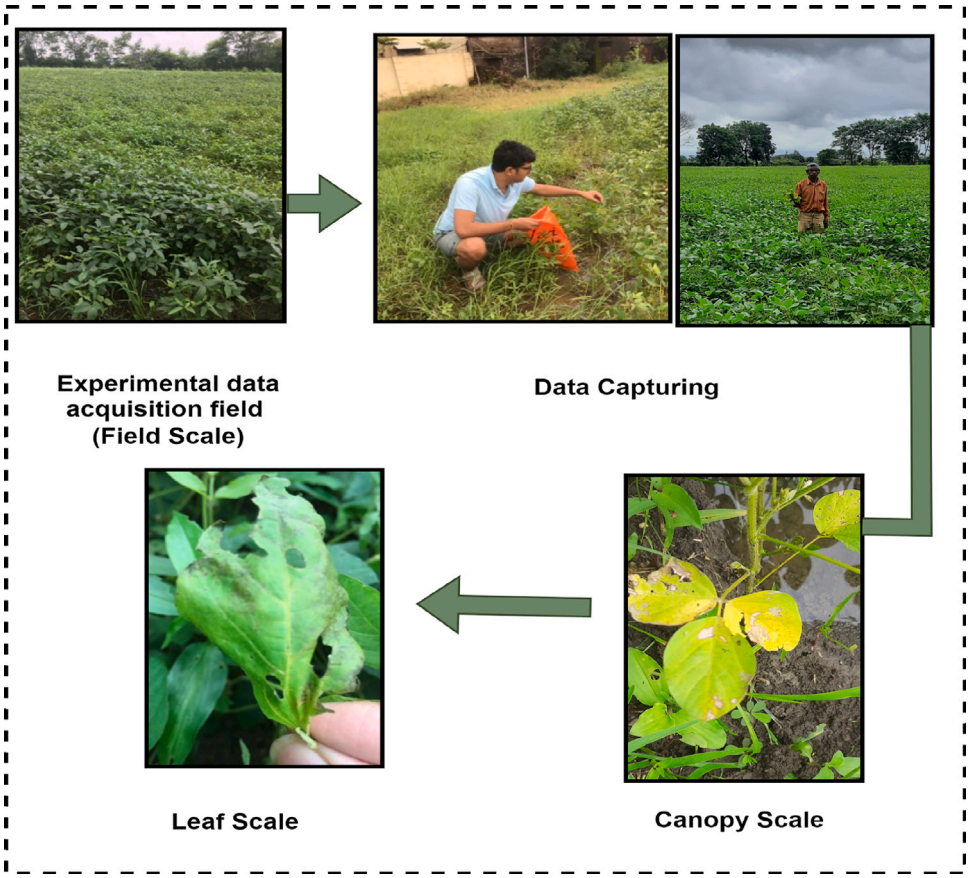

Fig. 6. Data collection framework.

pixels. Fig. 6 illustrates the overall data acquisition collected from the real field scenario. Besides this, the overall data pre-processing methodology is depicted in Fig. 8.

In this study, the images of soybean leaf diseases were acquired by using a digital camera (SONY's DSLR-A350, Japan). The exposure speed of the camera was set at 1/1000 s, and the aperture value was F8.0. During the shooting process, the lens was kept perpendicular to the blade, and the shooting distance was always kept at 40 cm. To show the growth state of infected leaves of soybean plants under natural conditions, the weather condition of a windless and sunny day was selected to complete the data acquisition. During the whole acquisition experiment, soybean leaves did not need to be picked to avoid adverse effects on soybean plant growth. In the four growth stages of soybean, including full flowering stage, beginning pod stage, full pod stage, beginning seed stage and full seed stage, 133 images of soybean brown leaf spot (SBLS), 277 images of soybean frogeye leaf spot (SFLS) and 113 images of soybean phyllosticta leaf spot (SPLS) were collected, with a total of 523 image samples, as shown in Fig. 7. Further, the dataset collected from the real field is preprocessed further using the following steps. Taking the soybean disease images as the object, the OTSU was used to separate the soybean disease leaves from the background. In addition, the region calibration was based on the segmented image, and soybean disease image with single leaf was acquired by mouse point selection and image clipping. Finally, the image size was scaled to 224 × 224, and then dark change, mirror transform and rotation were used to expand the data set image.

4.2.1. Region labeling

After the data curation, the next step involves the process of image labeling. The image acquired from the field contains different diseased leaves. Therefore, leaves are segmented using a region calibration technique that follows a defined procedure. Firstly, the image with a segmented background is processed. Secondly, calibration is performed on the defined set of leaves. Third, the image calibration coordination method is applied, which define the direction of images on both axis. In this, M and N axis direction is from top to bottom and left to right. Fig. 8 illustrates the rectangular area selected is removed from each leaf bunch. This process is performed repeatedly on the overall sample images.

4.2.2. Dataset expansion and partitioning

In this section dataset expansion, partitioning is performed on the images after scaling which is of size 224 × 224. Besides this, to obtain a single leaf, a region calibration method is applied. In addition, after applying these data augmentation techniques the overall image count is 4829 of six different leaf classes. Table 4 and 5 illustrates the different image datasets. Furthermore, it is also evident from the table that the soybean dataset is imbalanced which affects the classification performance and results in an over-fitting problem. To overcome these challenges, different augmentation techniques such as flipping, zoom, rotation, and translation have been applied to balance the dataset. Furthermore, to access the model generalization capabilities *i.e* the model performance on unseen datasets the data set is divided into the ratio of 80% in training samples, 20% testing sample, and 70% in training samples, 30% testing samples and depicted in Tables 6 and 7. Besides, this the dataset details of the images collected is depicted in Table 3.

4.3. Experimental results

In this study, two experiments were conducted for soybean crop disease. Primarily, the object detection task is performed which is used

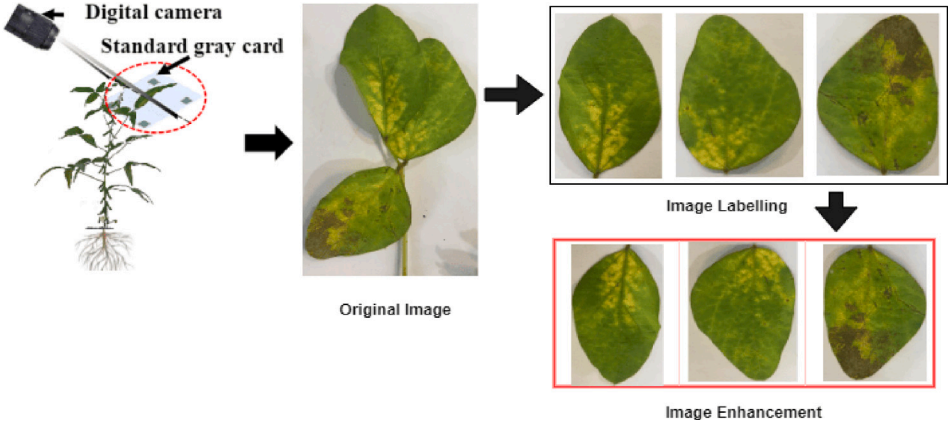

Fig. 7. Data processing methodology.

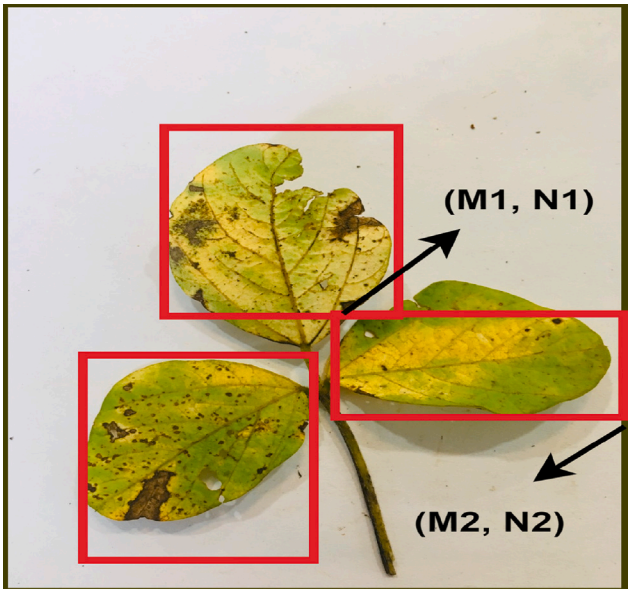

Fig. 8. Leaf calibration.

Table 3
Soybean Real Field Dataset.

Subject	Information
Type of data	Raw data
How data were acquired	Data curation performed with the help of local farmers, experts, and pathologists using different devices. The Sony Cyber-shot DSC-300 camera, and smartphones, including the Samsung Galaxy M31 equipped with a 64-megapixel camera and an Apple iPhone with a 12-megapixel camera.
Data format	The images are in JPEG format with a standard size of 224 × 224 pixels.
Description for data collection	The images were captured around 11:30 a.m. and 3:30 p.m. during ideal illumination and mild temperatures for data curation. captured primarily concentrate on the upper regions of soybean leaves affected by insect infestation, as well as leaves that exhibit normal, healthy conditions.
Data source location	Season 1 is collected from Vidisha district, whereas season 2 is from the guna district of Madhya Pradesh, India.

Table 4
Statistical data of real field collected soybean dataset (80:20).

Classes	Original dataset	Common Augmented	Training Samples (% 80)	Testing Samples (% 20)
Bacterial Blight	88	801	640	161
Brown Spot	81	801	640	161
Black Dot	381	803	640	163
Ferrugem Rust	352	802	640	162
Mild Infected	800	800	640	160
Severe Infected	822	822	640	160
Overall Images	2524	4829	3840	967

Table 5
Statistical data of real field collected soybean dataset (70:30).

Classes	Original dataset	Common Augmented	Training Samples (% 70)	Testing Samples (% 30)
Bacterial Blight	88	801	560	241
Brown Spot	81	801	560	241
Black Dot	381	803	561	242
Ferrugem Rust	352	802	560	242
Mild Infected	800	800	560	240
Severe Infected	822	822	575	247
Overall Images	2524	4829	3376	1453

to determine the leaf identification which shows the bounding boxes in the diseased leaf. Here, the recently developed YoLoV8 object detection model, is chosen which shows the excellent evaluation in identifying the damaged plant leafs. The YOLOv8 incorporates unique module which is capable of detecting the object in real time scenario, which drastically improves the information retention. Moreover, when different factors are combined it improves the object detection in the real field collected dataset without loosing the internal details of the image. Fig. 9 illustrates the bounding boxes which shows the confidence level of each predicted level. It is clearly vindicated that from the figure that the large rectangular boxes areas with the upper quartile edge, indicates the higher prediction accuracy for black dot, mild infected, rust, and severe infected. However, when the sample image dataset is insufficient as per the model requirements the prediction accuracy is slightly deteriorated. Secondly, the classification task is performed. For experimental analysis, the proposed SoyaTrans model is deployed and tested on the original and augmented dataset as shown in Table 5. The efficacy of the proposed SoyaTrans model is validated on real-field collected soybean leaf dataset. In addition, the proposed model is applied on different benchmark datasets namely, Embrapa, Plant Village, AI2018, and PlantDoc. The experimental findings of SoyaTrans are compared to the ten state-of-the-art methods in terms of five parameters, namely parameters, accuracy, precision, recall, and F1-score. The proposed model surpassed all the ten state-of-the-art considered models, such as Ghost-convolution enlightened Transformer (Yu, Xie, & Huang, 2023),

PlantXVIT (Singh Thakur, Khanna, Sheorey, & Ojha, 2022), Convolutional Swin Transformer (Guo et al., 2022), Inception convolutional transformers (ICVT) (Lu et al., 2022), Former-Leaf (Thai, Le, & Nguyen, 2023), Con-Vit (Li, Chen, Yang & Li, 2022), RIC-Net (Zhao, Sun, Xu & Chen, 2022), MobileNet-V2 (Chen, Zhang, Suzauddola, & Zeb, 2021), MFSwin Trans (Bi et al., 2022), and EfficientNet (Feng, Ong, Teh, & Zhang, 2024). Table 6 and 7 depicts the quantitative results of different CNN, attention, and transformer-based models on the soybean dataset. The high accuracy reported on multiple datasets suggests the model might be too closely fitted to the test conditions, and may not generalize well to unseen data, which is a common issue with deep learning models. For fair comparison the soybean dataset is divided different training and testing dataset. The SoyaTrans model has achieved 94.00% accuracy when the model is trained and tested on (80:20) ratio. Furthermore, is also inferred from the table that Inception convolutional transformers (ICVT) have achieved a competitive accuracy of 92.00% on the soybean dataset. Moreover, the accuracy is slightly degraded and achieved 92.00% when trained and tested on (70:30) ratio soybean dataset. It is also evident from the table that RIC-Net, convolution swin tansformer, and RIC-Net is competitive. Therefore, it is clearly evident from the results that the SoyaTrans network surpassed the state-of-the-art models in terms of accuracy. Further, the experimental results also affirm that the SoyaTrans model has outperformed all other methods in terms of accuracy, F1 score, recall, precision, memory usage and Flops. Whereas, in terms of inference time PlantXvit model is better than proposed SoyaTrans model with 13.01 ms and 14.50 ms in both ratio of the dataset. However, the proposed SoyaTrans achieves better accuracy than plantxvit. In addition, the classification performance of the proposed model has been applied to the different benchmark datasets namely, Plant villages (Hughes, Salathé, et al., 2015), AI2018 Challenger (Wu et al., 2017) PlantDoc, (Singh et al., 2020) and Embrapa dataset (Barbedo et al., 2018) as shown in Table 8. The four benchmark datasets consist of a total 76 crops with 220 classes, and a total number 139,112 images respectively. Table 9 illustrates the accuracy comparison of the proposed and state-of-the-art models on different datasets. In this, the models are trained and tested on two state-of-the-art datasets, namely plant village, and AI challenger comprising 54,306, and 35,861 images respectively. Furthermore, the models are tested on two state-of-the-art datasets plantdoc, and embrapa dataset comprising 2569, and 46,376 images captured from different demographic regions across the globe. The experimental results witness that the developed model is able to perform on a diverse range of datasets which confirms the global applicability and reproducibility of the developed model. It is clearly evident from the table that the proposed SoyaTrans has achieved the highest accuracy of 98.00%, 97.00%, 76.00%, and 92.00% on plantvillage, AI2018, PlantDoc, and Embrapa datasets. While, MobileNet V2 is the second runner method with 96%, in the plant village dataset. In addition, in the AI2018 dataset PlantXvit, and MobileNet v2 share a common position as compared to a proposed model with 90% accuracy. Besides this, in the PlantDoc dataset MFSwin Trans, and MobileNet V2 show competitive performance with 77% accuracy. Finally, in the Embrapa dataset, the RIC-Net, EfficientNet, and MFSwin Transformer network beholds the runner-up method with 90.00% accuracy. However, the accuracy of the proposed model is slightly deteriorated on the datasets where only testing is performed namely Plant-Doc and Embrapa. In addition, to this the performance of the proposed model surpassed the state-of-the-art model in terms of accuracy.

Furthermore, to demonstrate the efficacy of the SoyaTrans model, the T-SNE method is applied to get a better view of right similar and dissimilar samples of different classes in the dataset (Van der Maaten & Hinton, 2008). The learning features of the proposed SoyaNet model and other competing networks for the soybean dataset are plotted on 2D- plane using this method. Fig. 13, and 14 depicts the visualization results of different models on the soybean dataset. The final reshape or convolutional layer is responsible for the visualization of feature maps. It is clearly depicted from the figure that the different colors are indicative of distinct soybean classes. In the case of ghost, convolution enlightened transformer network the feature produced by different disease categories is not appropriate. In this, the disease category highlighted in pink *i.e.* ferrugum rust performs better results. Whereas other disease categories are not identified. This could be due to the presence of various types of diseases within a single leaf. Moreover, LIME is another explanation model for better interpreting of results in a faith-full manner that works on a linear approximation of model (Ribeiro, Singh, & Guestrin, 2016). Fig. 10 depicts the comparative analysis of different models. The proposed model is trained on the real field and collected a dataset of six different classes. It is evident from the results the proposed model identifies the disease portion carefully and understands the silent features in bacterial blight, brown spot, black dot, ferrugem rust, and severe infected. However, in mild infected disease, the model is focused partially and the results are not explainable.

Furthermore, Grad-CAM analysis of the developed model is also presented is another to explain and understand the model's behavior for the classification. Fig. 11 depicts the comparative analysis of different models in low intensity as well as the high intensity. Moreover, grad-cam gradient weighted class activation maps is another explanation model for better interpreting of results in a faith-full manner that works on a linear approximation of model. The proposed model is trained on the real field collected a dataset of six different classes. It is worth noting that the proposed soytrans model correctly identifies the diseased portion in severe infected as compared to mild infected. Moreover, in case of bacterial blight the results witnesses that it low intensity the illness portion of the disease is correctly identified in contrast wit the high intensity. It is evident from the results the proposed model identifies the disease portion carefully and understands the silent features in blackdot, brown spot, black dot, and mild infected . However, in rust disease, the model is focused partially and the results are not explainable.

Besides this, Fig. 12 integrated gradient AI explain-ability techniques is applied which shows some deeper insights of leaf image. However, the dataset collected from real field, which is effected from different environmental affects, performs the competitive results. This technique is applied over the six different classes of real field soybean collected a dataset. The results signifies that the integrated techniques is one of the prominent technique which signifies the deep insights of disease leaf images. As per our visual inception, specifically where the dataset is appropriate as per the model requirement namely black dot, rust, severe infected, mild infected performs well. Whereas, in the other classes such as bacterial blight, and brown spot results where images are less performance is deteriorated and not explainable. Sundararajan, Taly, and Yan (2017).

In addition, the Confusion matrices returned by the different models on the soybean dataset are depicted in Fig. 15. The confusion matrices for the soybean dataset are provided here. However, for other benchmark datasets with numerous large classes, it is not feasible to display them in this context. In the confusion matrix, the diagonal elements represent the correctly classified classes and others that are not classified properly. It can be observed that the proposed SoyaTrans model correctly classifies the majority of the classes of soybean namely bacterial blight, brown spot, black dot, ferrugem rust, mild infected, and severe infected. In addition, the clarification of brown spots is accurate as compared to other disease classes.

4.4. Discussion

This study aims to develop an efficient model which incorporates the capabilities of CNN and transformer model extracting the local and global feature of image of the dataset collected from the real field environment. Since, the deep learning models are data hungry which requires a massive amount of data, while lightweight models fail in discriminating the efficient lesion features for all types of plant disease. Furthermore, efforts has been made by different researches to overcome

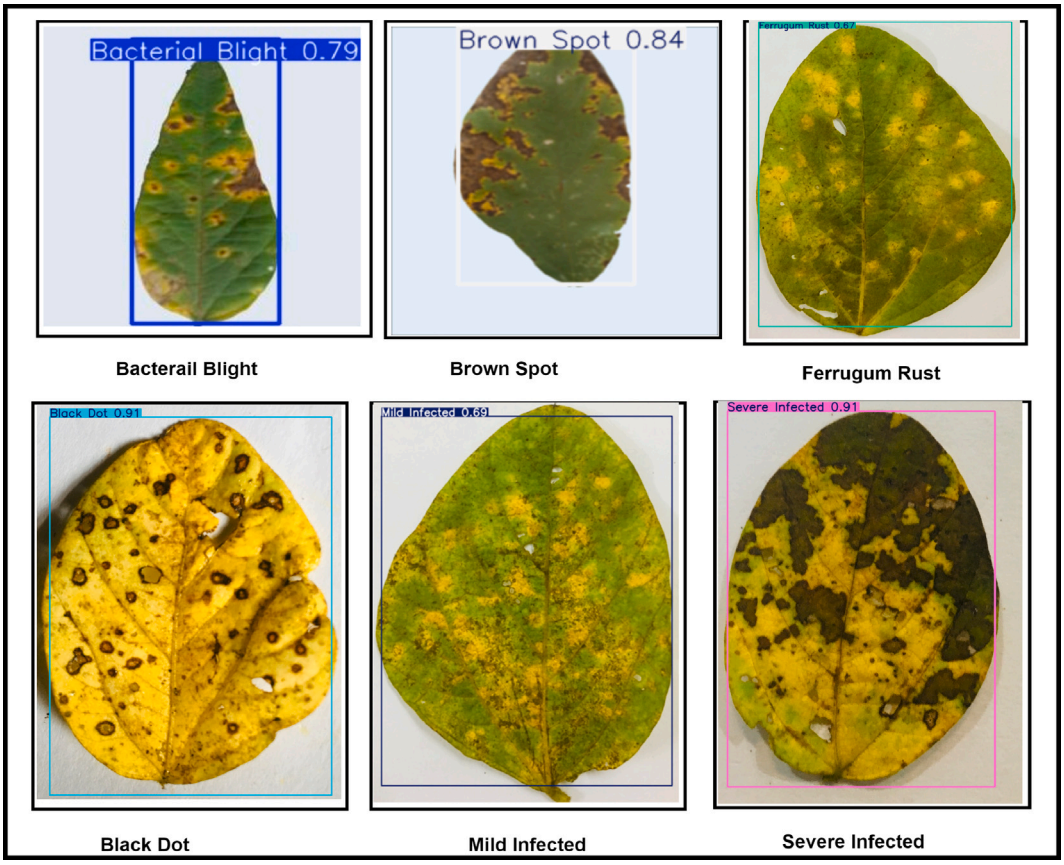

Fig. 9. Object detection results: Ground truth predicted bounding boxes.

Table 6
Comparison of the accuracy, precision, recall, FLOPS, memory usage, inference time, and number of parameters on (80:20).

Classification Models	Accuracy	Precision	Recall	F-1 score	Flops-(G)	Infer-ence Time (ms)	Mem-ory Usage (MB)	Parameters (M)
Ghost-convolution enlightened Transformer (Yu et al., 2023)	0.67	0.68	0.67	0.66	4.20	20.55	95.6	25.39
PlantXVIT (Singh Thakur et al., 2022)	0.78	0.80	0.78	0.77	1.41	13.01	23.9	6.40
Convolutional Swin Transformer (Guo et al., 2022)	0.89	0.90	0.88	0.89	4.35	22.31	103.2	27.47
Inception convolutional transformers (ICVT) (Lu et al., 2022)	0.92	0.96	0.95	0.95	1.69	17.22	46.7	11.16
Former-Leaf (Thai et al., 2023)	0.74	0.76	0.74	0.74	8.92	36.73	215.6	60.00
Con-Vit (Li, Chen, Yang & Li, 2022)	0.73	0.75	0.73	0.75	5.96	29.01	145.3	38.00
RIC-Net (Zhao, Sun, Xu & Chen, 2022)	0.88	0.89	0.87	0.88	1.23	15.02	25.8	6.71
MobileNet-V2 (Chen et al., 2021)	0.80	0.85	0.76	0.80	1.08	16.23	21.6	**5.32**
EfficientNet (Feng et al., 2024)	**0.89**	**0.88**	**0.89**	**0.90**	**2.84**	**29.03**	**117.8**	**24.00**
MFSwin Trans (Bi et al., 2022)	0.91	0.90	0.89	0.90	1.74	19.03	47.8	12.00
Proposed	**0.94**	**0.93**	**0.94**	**0.92**	**1.01**	**14**	**20.9**	**5.20**

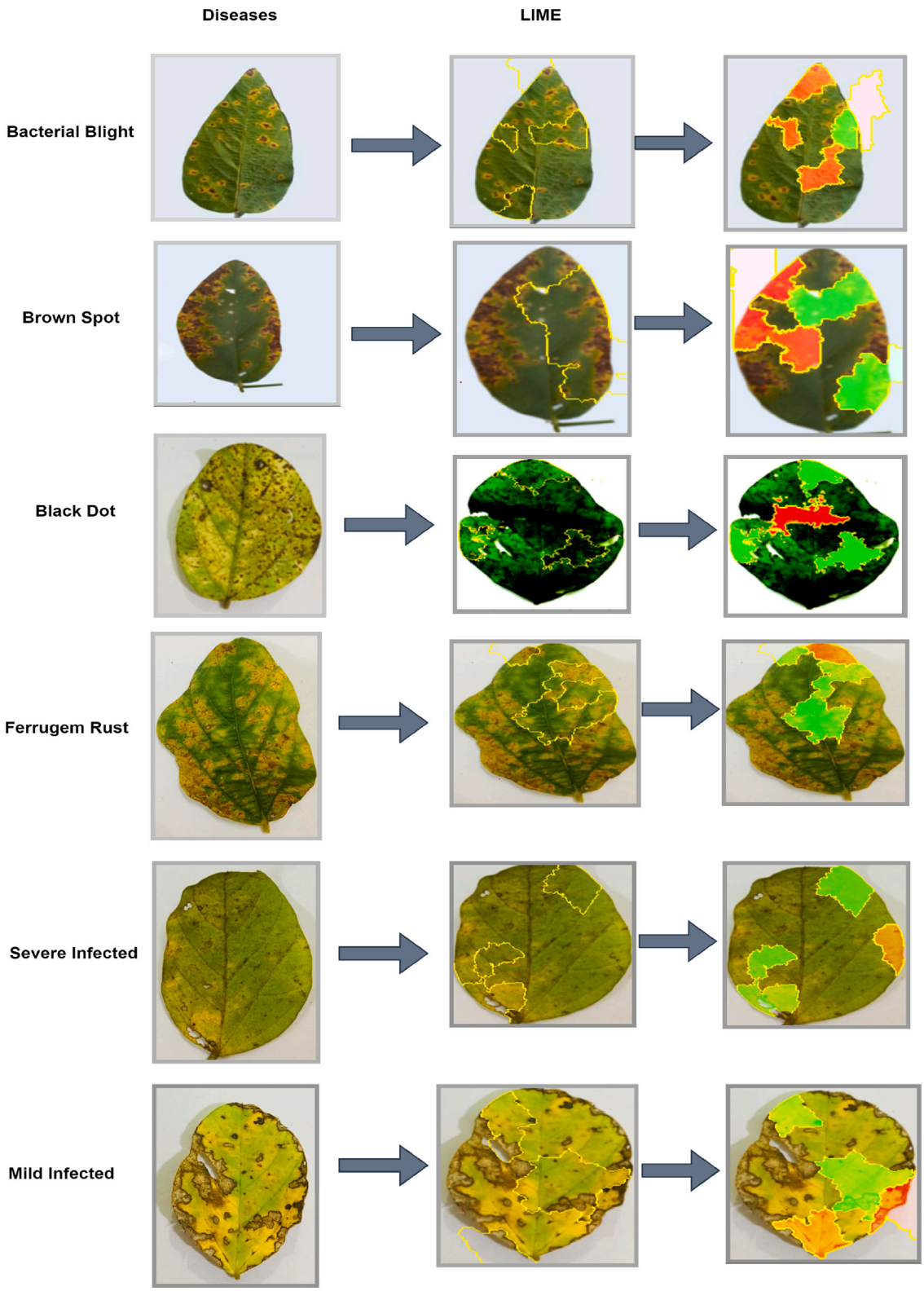

Fig. 10. Feature visualization of trained SoyTrans leaf images.

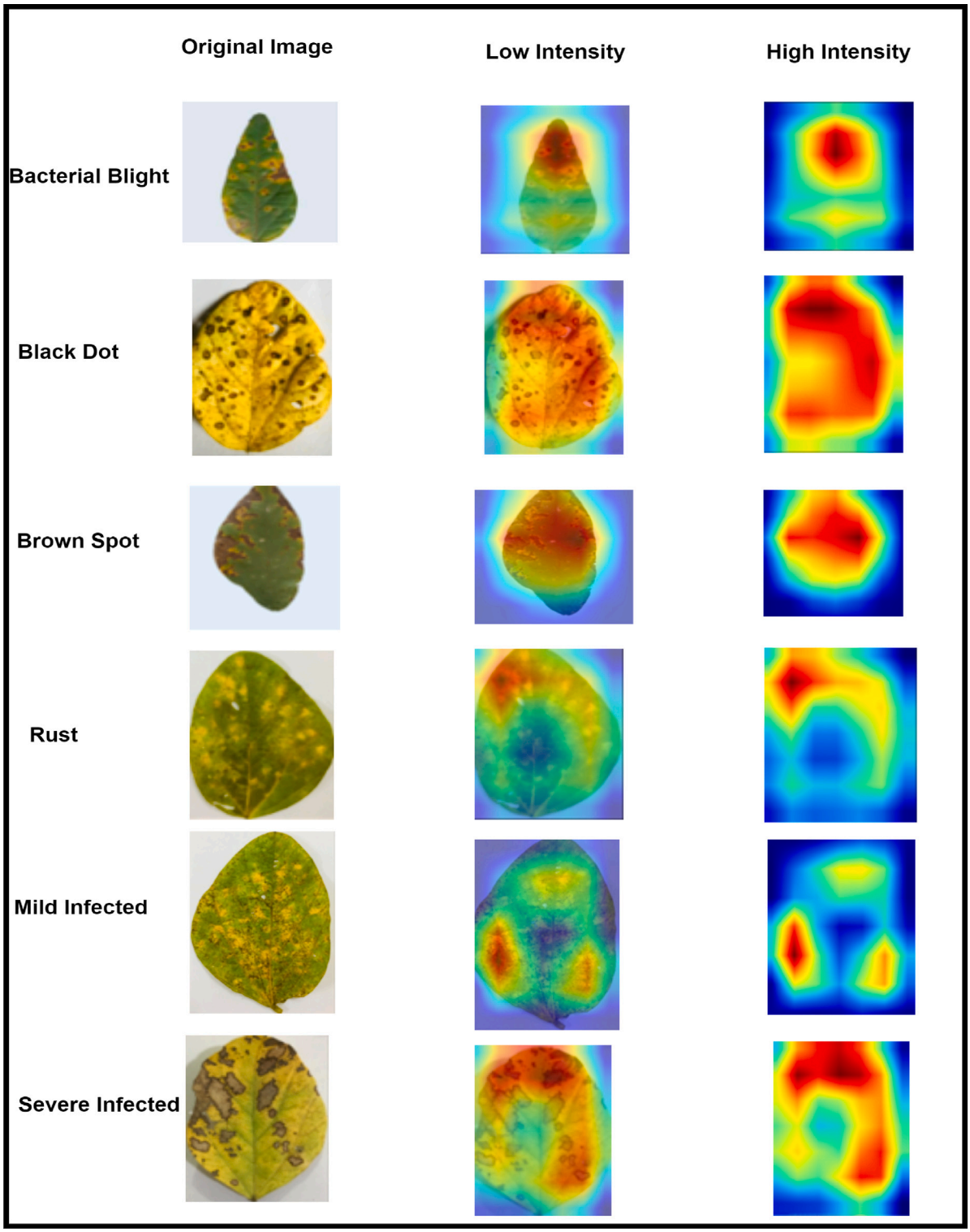

Fig. 11. Feature visualization results of gradcam trained on Soybean leaf images.

these challenges. Therefore, a novel soyatrans model is designed in such a way that, it can work on different types of crops disease identification with minimum training parameters, FLOPs, memory usage, and inference time.

Table 6 and 7 presents the comparison of the accuracy, precision, recall, FLOPS, memory usage, inference time, and number of parameters. It is noted that the proposed model has the minimum number of parameters, Flops, and memory usage. In addition, to this combining swin transformer block with the inception blocks performs better multi-scale feature which deepens the network's learning and strengthens its pattern. Moreover, the accuracy of mobileNet-v2 (Chen et al., 2021) is very quite closer to the proposed model. However, the above model achieves lower accuracy and also under performs in other metrics. Experimental results indicate that the model is lightweight, which limits its ability to accurately detect leaf diseases. Therefore, when selecting a model, it is crucial to consider both the model size and performance, especially when dealing with small datasets. Furthermore,

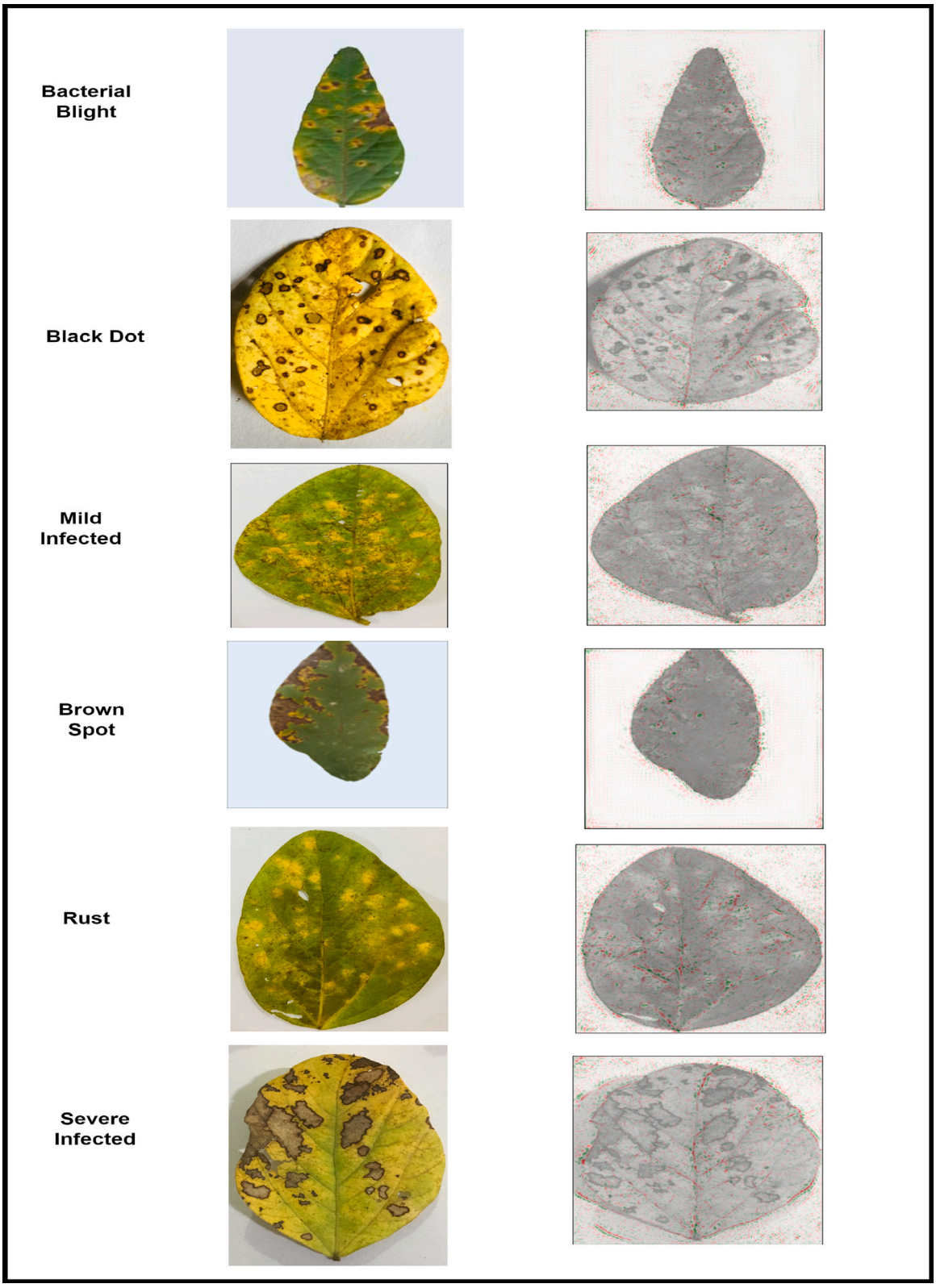

Fig. 12. Feature visualization results of integrated gradients trained on Soybean leaf images.

to deploy the model in real field scenario depends upon the different factor. In addition, the device contain the low memory foot prints, which are measured in FLOPs. One of the major advantage of the proposed soyatrans model is the lower FLOPs, which is due to different operation involved in due to the proposed novel random shifting. The proposed random shifting are crucial for enhanced interpretability, necessitating low computational resources as compared to state-of-the-art shifting technique. Further, the inference time of the proposed model is competitive, second after the plantxvit (Singh Thakur et al., 2022). Moreover, two quantitative measures LIME and GRAD-CAM affirms that the combination of VGG16, inception V7, new shifting criteria, with two blocks of swin transformer improves the classification

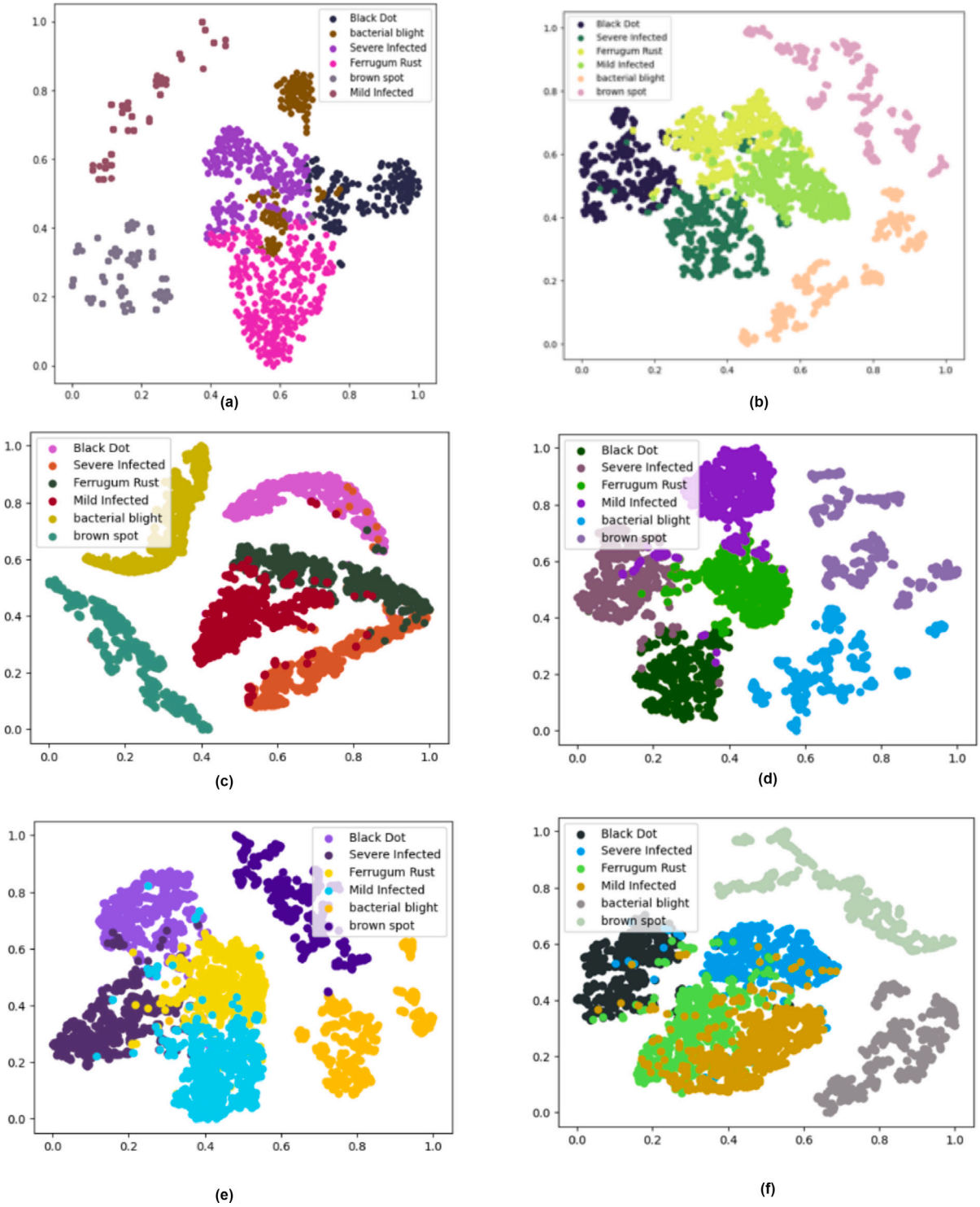

Fig. 13. TSNE Plots for soybean dataset (a) Ghost convolution enlightened transformer, (b) PlantXVIT, (c) Convolutional Swin Transformer, (d) Inception convolutional trans formers (ICVT), (e) Former-Leaf, (f) Proposed SoyaTrans.

of lesion features in plant leaves. The model proposed by Bi et al. has developed as swin transformer based model the less number of trainable parameters as compare to others and very fewer inference time and flops . However, the accuracy has drastically deteriorated. This is due to the fact that transformers fail to captures the local features, due to the small sample size (Bi et al., 2022).

Considering the inference time of the different models, the average time followed by plantxvit (Singh Thakur et al., 2022), RIC-Net (Zhao, Sun, Xu & Chen, 2022), MobileNet-V2 (Chen et al., 2021), Inception convolutional transformers (ICVT) (Yu et al., 2023), MFSwin Trans (Bi et al., 2022), and EfficientNet (Feng et al., 2024). The results witnessed that the proposed model has attained the moderate inference time in comparison to the other state-of-the-art models. In addition to this, the moderate flops count, which enhance the higher inference time. In contrast, the work is novel, though the different combination has been proposed with ViT. It is worth mentioning that the with introducing a novel random shifting criteria for the first time has significantly improved the leaf disease identification. Another important factor is the

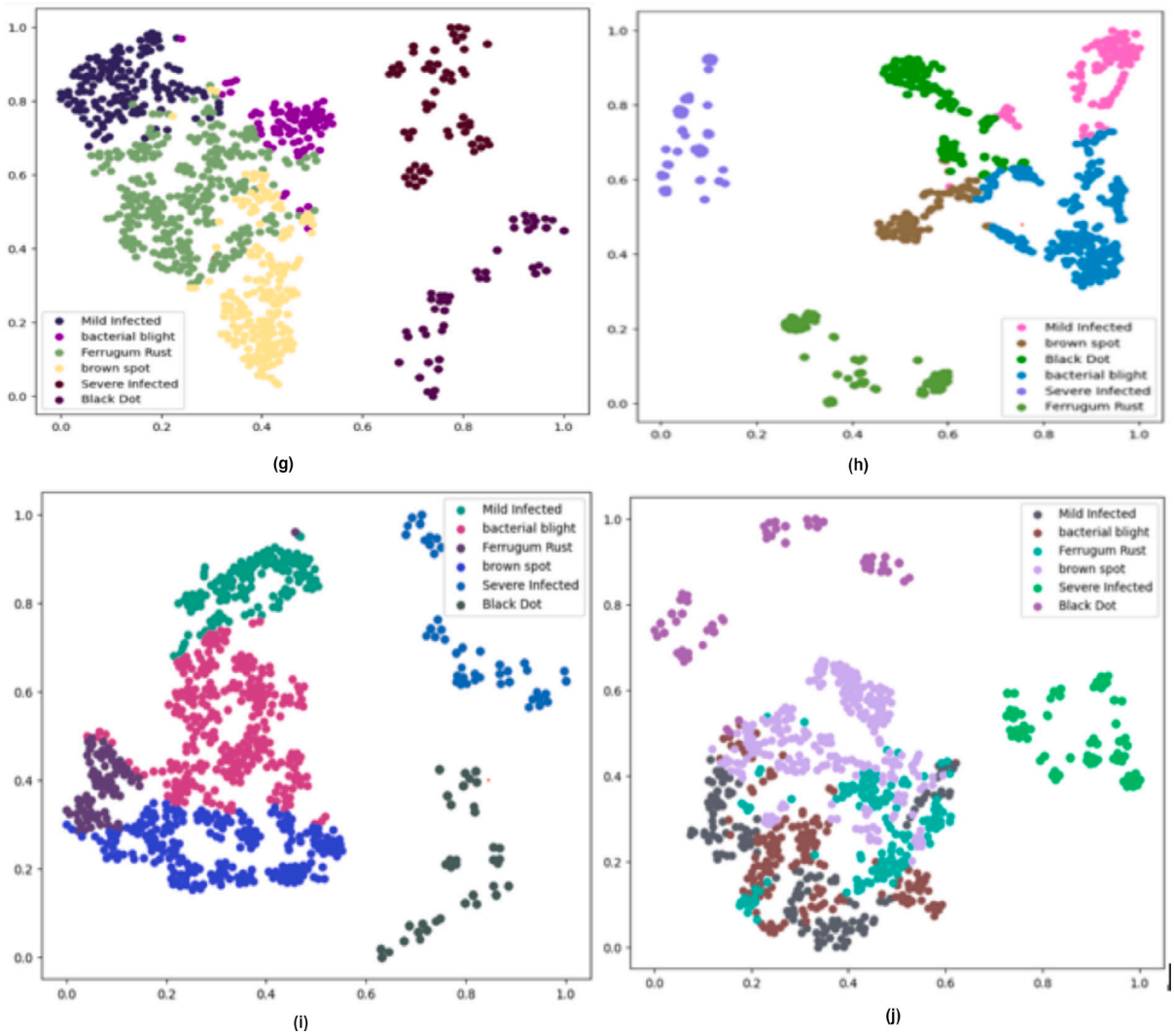

Fig. 14. TSNE Plots (g) RIC-Net, (h) MobileNet-V2, (i) MFSwin Trans, and (j) Con-Vit.

generalizability. To assess this, the proposed model is validated on real field soybean dataset, and four openly accessible datasets, specifically Embrapa, Plant Village, AI2018, and PlantDoc. The results are shown in Table 9.

The results clearly affirm that, the performance of MFSwin Trans model produces the competitive results (Bi et al., 2022). However, this model has high number of parameters in comparison to the proposed soyatrans model. Furthermore, key limitation which affects the model performance is the data captured from the real world environment. The dataset collected from real world consist of complex background, with multiple leaves in a single leaf affects degrades the model performance. Therefore, in the current study, the proposed model has been applied on four benchmark datasets. The results witnesses that the proposed model can be deplorable in real world with IoT applications as compared to the state-of-the-art models. However, there are several setbacks to look upon namely, early identification of similarity leafs disease symptoms during in its initial phase of development which is the major challenge in deploying the AI solutions. Another crucial factor is environmental changes, which is the major factor which shows the evolution of new variety of leaf disease. Conclusively, their is still a scope of improvement, where more robust techniques can be developed to overcome these challenges.

5. Conclusion

Plant diseases pose a substantial economic loss to agricultural productivity, constituting a significant threat to the sustainability of this sector. Therefore, this paper presents a novel SoyaTrans model for efficient classification of the real-field soybean dataset and other benchmark datasets. In the proposed SoyaTrans model, a Swin-enabled convolutional neural network is proposed for crop disease identification. The proposed model blends the benefits of the transformer to mine the global features with the CNN module to mine the local features of the image. In addition, a new random shifting criterion is proposed which not only improves the classification performance by increasing the discriminative features of the image but also exhibits lower computational complexity in comparison to the conventional cyclic shift criterion. Moreover, to overcome the challenge of limited training samples, a standard soybean plant leaf disease dataset, collected from real fields, is presented in this paper. In contrast to other networks, the proposed SoyaTrans network attains the best accuracy with 94.00% and 92.00% on two different dataset ratio of the newly presented soybean leaf dataset. Moreover, the proposed model is also evaluated on the performance indicators, and outperform on namely, number of parameters, flops, and memory usage except inference time. The minor hindrance is in reducing the inference time. Furthermore, the proposed

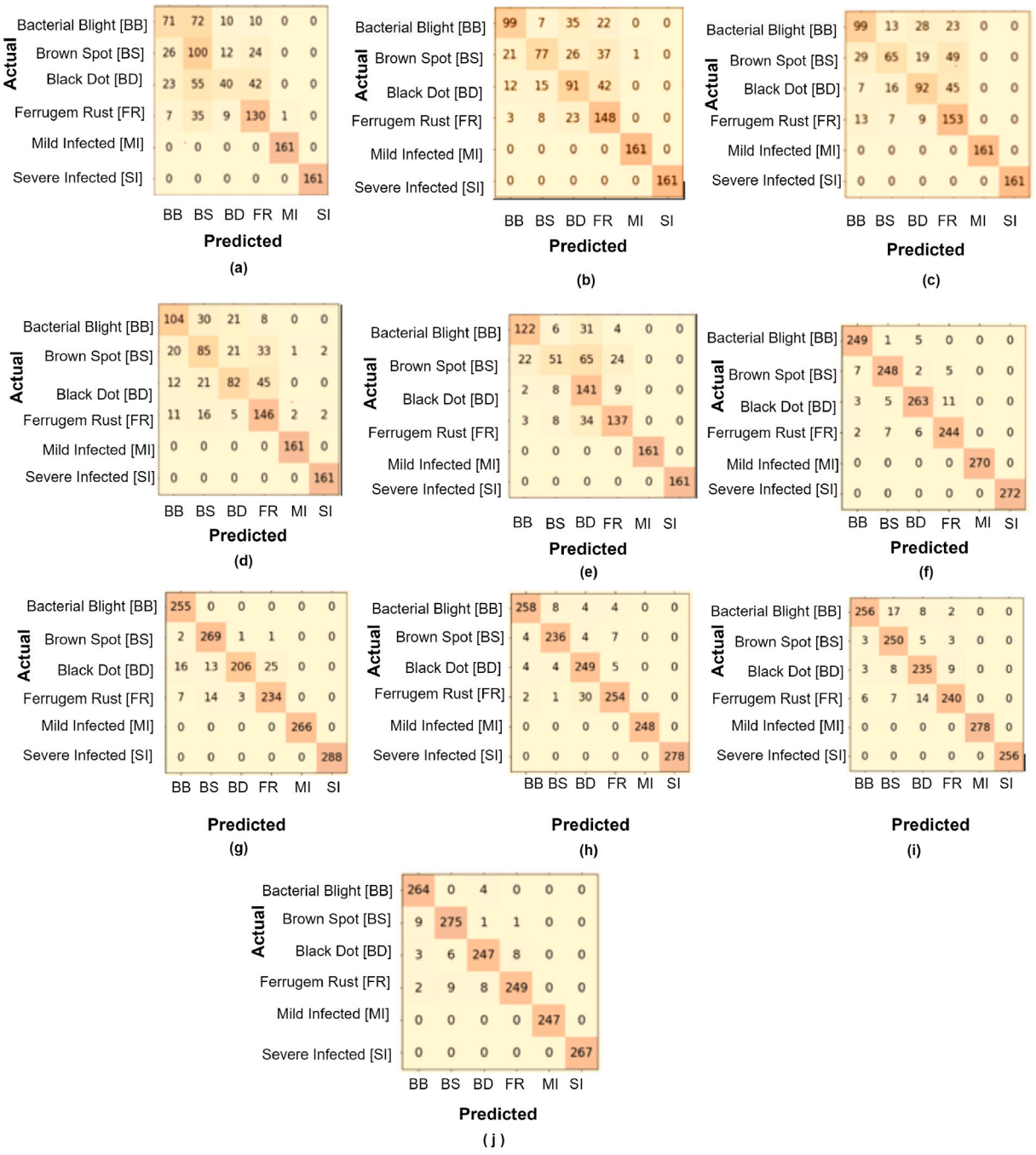

Fig. 15. Confusion Matrix for soybean dataset (a) Ghost convolution enlightened transformer, (b) PlantXVIT, (c) Convolutional Swin Transformer, (d) Inception convolutional trans formers (ICVT), (e) Former-Leaf, (f) Con-Vit, (g) RIC-Net, (h) MobileNet-V2, (i) MFSwin Trans, and (j) Proposed SoyaTrans.

model is validated across multiple openly accessible datasets, specifically Embrapa, Plant Village, AI2018, and PlantDoc. Experimental results of the proposed model are compared against the ten state-of-the-art methods in terms of eight parameters, namely parameters, accuracy, precision, recall, F1-score, inference time, memory usage, and FLOPS. The proposed model surpassed all the ten state-of-the-art models, even under complicated backdrops, with an accuracy of 98.00%, 97.00%, 76.00%, and 92.00% on plantvillage, AI2018, PlantDoc, and Embrapa dataset with comparatively lowest computational complexity *i.e.*, 5.2 million parameters. The results witnesses that the proposed soyatrans model cost and size is competitive as compare to CNN, which can boost the generalizability. For better interpretability of the results, the models are evaluated using the LIME tool, GRAD-CAM, T-SNE plots, and confusion matrices to show the distinct features of the considered models. Conclusively, this research presents the first pioneering and practical solution for using random shifting in transformer networks. In future, the proposed model may be further optimized to capture the dynamic situations such as mutation of the pathogens in the real field environment. The techniques such as continual learning may be adopted to mitigate this concern. Although the proposed model reduces the computational cost and number of parameters, different networks can be designed that reduce the inference without compromising their performance when deployed in real world applications. The findings of this research provides tangible benefits to the farmers which facilitates decision-making for early identification of leaf disease identification, enhances the economic, and improves the productivity of crops.

Table 7
Comparison of the accuracy, precision, recall, FLOPS, memory usage, inference time, and number of parameters on (70:30).

Classification Models	Accuracy	Precision	Recall	F-1 score	Flops-(G)	Infer-ence Time (ms)	Mem-ory Usage (MB)	Parameters (M)
Ghost-convolution enlightened Transformer (Yu et al., 2023)	0.65	0.66	0.64	0.65	4.00	22.09	95.9	25.39
PlantXVIT (Singh Thakur et al., 2022)	0.79	0.80	0.79	0.78	1.61	14.50	23.9	6.40
Convolutional Swin Transformer (Guo et al., 2022)	0.84	0.83	0.84	0.85	5.35	24.06	105.2	27.47
Inception convolutional transformers (ICVT) (Lu et al., 2022)	0.89	0.86	0.44	0.91	1.74	18.61	49.4	11.16
Former-Leaf (Thai et al., 2023)	0.69	0.70	0.70	0.73	10.85	39.56	224.6	60.00
Con-Vit (Li, Chen, Yang & Li, 2022)	0.70	0.71	0.73	0.74	6.91	32.30	153.4	38.00
RIC-Net (Zhao, Sun, Xu & Chen, 2022)	0.84	0.85	0.85	0.86	1.76	16.34	28.8	6.71
MobileNet-V2 (Chen et al., 2021)	0.77	0.79	0.78	0.79	1.37	19.64	24.6	**5.32**
EfficientNet (Feng et al., 2024)	**0.85**	**0.83**	**0.83**	**0.85**	**3.72**	**28.03**	**112.8**	**24.00**
MFSwin Trans (Bi et al., 2022)	0.89	0.89	0.88	0.90	1.74	19.03	47.8	12.00
Proposed	**0.92**	**0.91**	**0.91**	**0.92**	**1.30**	**17.34**	**22.9**	**5.20**

Table 8
Benchmark Datasets used in experiments.

Dataset	No of Crops	Classes	Total images
Plant villages (Hughes et al., 2015)	14	38	54,306
AI2018 Challenger (Wu et al., 2017)	10	59	35,861
PlantDoc (Singh et al., 2020)	13	30	2,569
Embrapa (Barbedo et al., 2018)	39	93	46,376
Total	76	220	139,112

Table 9
Average accuracy comparison on benchmark datasets.

Classification Models	Plant Village	AI2018	Plant-Doc	Embrapa
Inception convolutional transformers (ICVT) (Yu et al., 2023)	0.94	0.89	0.67	0.80
PlantXVIT (Singh Thakur et al., 2022)	0.93	0.90	0.71	0.83
Convolutional Swin Transformer (Guo et al., 2022)	0.92	0.86	0.74	0.83
Ghost-convolution enlightened Transformer (Lu et al., 2022)	0.96	0.85	0.76	0.91
Former-Leaf (Thai et al., 2023)	0.91	0.88	0.72	0.87
Con-Vit (Li, Chen, Yang & Li, 2022)	0.93	0.87	0.76	0.88
RIC-Net (Zhao, Sun, Xu & Chen, 2022)	0.95	0.89	0.75	0.90
MobileNet-V2 (Chen et al., 2021)	0.96	0.90	0.77	0.89
MFSwin Trans (Bi et al., 2022)	0.94	0.89	0.77	0.90
EfficientNet (Feng et al., 2024)	**0.92**	**0.89**	**0.77**	**0.90**
Proposed	**0.98**	**0.97**	**0.79**	**0.92**

6. Declarations

Ethical approval

Not applicable

CRediT authorship contribution statement

Vivek Sharma: Conceptualization, Methodology, Software, Writing – original draft, Visualization, Validation. **Ashish Kumar Tripathi:** Supervision, Data curation, Formal analysis, Writing – original draft, Software. **Himanshu Mittal:** Writing – review & editing, Supervision, Resources, Validation. **Lewis Nkenyereye:** Visualization, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors express their gratitude for the resources and support received from Dr. K C Sharma principal scientist at ICAR-Indian Agricultural Research Institute, for providing valuable suggestions, leaf data annotation, and technical expertise.

References

Agarwal, D. K., Billore, S., Sharma, A., Dupare, B., & Srivastava, S. (2013). Soybean: introduction, improvement, and utilization in India—problems and prospects. *Agricultural Research*, *2*(4), 293–300.

Barbedo, J. G. A., Koenigkan, L. V., Halfeld-Vieira, B. A., Costa, R. V., Nechet, K. L., Godoy, C. V., et al. (2018). Annotated plant pathology databases for image-based detection and recognition of diseases. *IEEE Latin America Transactions*, *16*(6), 1749–1757.

Bi, C., Hu, N., Zou, Y., Zhang, S., Xu, S., & Yu, H. (2022). Development of deep learning methodology for maize seed variety recognition based on improved swin transformer. *Agronomy*, *12*(8), 1843.

Chang, B., Wang, Y., Zhao, X., Li, G., & Yuan, P. (2024). A general-purpose edge-feature guidance module to enhance vision transformers for plant disease identification. *Expert Systems with Applications*, *237*, Article 121638.

Chen, J., Zhang, D., Suzauddola, M., & Zeb, A. (2021). Identifying crop diseases using attention embedded MobileNet-V2 model. *Applied Soft Computing*, *113*, Article 107901.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Feng, J., Ong, W. E., Teh, W. C., & Zhang, R. (2024). Enhanced crop disease detection with EfficientNet convolutional group-wise transformer. *IEEE Access*, *12*, 44147–44162.

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, *145*, 311–318.

Gao, R., Wang, R., Feng, L., Li, Q., & Wu, H. (2021). Dual-branch, efficient, channel attention-based crop disease identification. *Computers and Electronics in Agriculture*, *190*, Article 106410.

Guo, Y., Lan, Y., & Chen, X. (2022). CST: Convolutional swin transformer for detecting the degree and types of plant diseases. *Computers and Electronics in Agriculture*, *202*, Article 107407.

Hughes, D., Salathé, M., et al. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv preprint arXiv:1511.08060.

Jin, H., Chu, X., Qi, J., Feng, J., & Mu, W. (2024). Learning multiple attention transformer super-resolution method for grape disease recognition. *Expert Systems with Applications*, *241*, Article 122717.

Kaur, P., Harnal, S., Tiwari, R., Upadhyay, S., Bhatia, S., Mashat, A., et al. (2022). Recognition of leaf disease using hybrid convolutional neural network by applying feature reduction. *Sensors*, *22*(2), 575.

Li, X., Chen, X., Yang, J., & Li, S. (2022). Transformer helps identify kiwifruit diseases in complex natural environments. *Computers and Electronics in Agriculture*, *200*, Article 107258.

Li, S., Li, K., Qiao, Y., & Zhang, L. (2022). A multi-scale cucumber disease detection method in natural scenes based on YOLOv5. *Computers and Electronics in Agriculture*, *202*, Article 107363.

Li, X., Li, X., Zhang, S., Zhang, G., Zhang, M., & Shang, H. (2022). SLViT: Shuffle-convolution-based lightweight vision transformer for effective diagnosis of sugarcane leaf diseases. *Journal of King Saud University-Computer and Information Sciences*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Liu, Y., Wang, Z., Wang, R., Chen, J., & Gao, H. (2023). Flooding-based MobileNet to identify cucumber diseases from leaf images in natural scenes. *Computers and Electronics in Agriculture*, *213*, Article 108166.

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*(5), 1755–1767.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Mostafa, A. M., Kumar, S. A., Meraj, T., Rauf, H. T., Alnuaim, A. A., & Alkhayyal, M. A. (2022). Guava disease detection using deep convolutional neural networks: A case study of guava plants. *Applied Sciences*, *12*(1), 239.

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? *Advances in Neural Information Processing Systems*, *32*.

Pacal, I. (2024). Enhancing crop productivity and sustainability through disease identification in maize leaves: Exploiting a large dataset with an advanced vision transformer model. *Expert Systems with Applications*, *238*, Article 122099.

Pandey, A., & Jain, K. (2022). A robust deep attention dense convolutional neural network for plant leaf disease identification and classification from smart phone captured real world images. *Ecological Informatics*, *70*, Article 101725.

Patil, R. R., & Kumar, S. (2022). Rice transformer: A novel integrated management system for controlling rice diseases. *IEEE Access*, *10*, 87698–87714.

Paymode, A. S., & Malode, V. B. (2022). Transfer learning for multi-crop leaf disease image classification using convolutional neural networks VGG. *Artificial Intelligence in Agriculture*.

Qian, X., Zhang, C., Chen, L., & Li, K. (2022). Deep learning-based identification of maize leaf diseases is improved by an attention mechanism: Self-attention. *Frontiers in Plant Science*, 1154.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" explaining the predictions of any classifier. *11*, In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., & Batra, N. (2020). PlantDoc: A dataset for visual plant disease detection. *vol. 10*, In *Proceedings of the 7th ACM IKDD coDS and 25th COMAD* (pp. 249–253).

Singh Thakur, P., Khanna, P., Sheorey, T., & Ojha, A. (2022). Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT. arXiv e-prints, arXiv–2207.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319–3328). PMLR.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Thai, H.-T., Le, K.-H., & Nguyen, N. L.-T. (2023). FormerLeaf: An efficient vision transformer for cassava leaf disease detection. *Computers and Electronics in Agriculture*, *204*, Article 107518.

Turkoglu, M., Yanikoğlu, B., & Hanbay, D. (2022). PlantDiseaseNet: Convolutional neural network ensemble for plant disease and pest detection. *Signal, Image and Video Processing*, *16*(2), 301–309.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, F., Rao, Y., Luo, Q., Jin, X., Jiang, Z., Zhang, W., et al. (2022). Practical cucumber leaf disease recognition using improved swin transformer and small sample size. *Computers and Electronics in Agriculture*, *199*, Article 107163.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision* (pp. 3–19).

Wu, J., Wen, C., Chen, H., Ma, Z., Zhang, T., Su, H., et al. (2022). DS-DETR: A model for tomato leaf disease segmentation and damage evaluation. *Agronomy*, *12*(9), 2023.

Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., et al. (2017). Ai challenger: A large-scale dataset for going deeper in image understanding. *20*, (8), (pp. 1049–1056). arXiv preprint arXiv:1711.06475.

Yu, H.-J., & Son, C.-H. (2020). Leaf spot attention network for apple leaf disease identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 52–53).

Yu, X., Wang, J., Zhao, Y., & Gao, Y. (2022). Mix-ViT: Mixing attentive vision transformer for ultra-fine-grained visual categorization. *Pattern Recognition*, Article 109131.

Yu, S., Xie, L., & Huang, Q. (2023). Inception convolutional vision transformers for plant disease identification. *Internet of Things*, *21*, Article 100650.

Zeng, W., & Li, M. (2020). Crop leaf disease recognition based on self-attention convolutional neural network. *Computers and Electronics in Agriculture*, *172*, Article 105341.

Zeng, T., Li, C., Zhang, B., Wang, R., Fu, W., Wang, J., et al. (2022). Rubber leaf disease recognition based on improved deep convolutional neural networks with an cross-scale attention mechanism. *Frontiers in Plant Science*, 274.

Zhao, X., Li, K., Li, Y., Ma, J., & Zhang, L. (2022). Identification method of vegetable diseases based on transfer learning and attention mechanism. *Computers and Electronics in Agriculture*, *193*, Article 106703.

Zhao, Y., Sun, C., Xu, X., & Chen, J. (2022). RIC-net: A plant disease classification model based on the fusion of inception and residual structure and embedded attention mechanism. *Computers and Electronics in Agriculture*, *193*, Article 106644.

Zheng, H., Wang, G., & Li, X. (2022). Swin-MLP: a strawberry appearance quality identification method by swin transformer and multi-layer perceptron. *Journal of Food Measurement and Characterization*, 1–12.