Self-Supervised Vision Transformer with Swarm Intelligence for Pattern-Aware Crop Stress Detection in Smart Farming Environments

Fatima Al Nuaimi and Noor Aisyah Rahman

Department of Software Engineering
School of Computing and Intelligent Systems, Kuala Lumpur International University
Malaysia
Department of Artificial Intelligence
Faculty of Engineering and Digital Technologies, Abu Dhabi Institute of Technology
UAE

ABSTRACT

The following limitations exist in the traditional crop stress detection techniques in the background of smart farming applications: low detection accuracy, inefficient feature extraction, poor adaptability to the environment, and high computational complexity. Stress detection using traditional machine learning and convolution-based methods is inefficient in capturing complex stress patterns of drought, pests, diseases, and nutrient deficiency, affecting productivity and precision agriculture systems. To overcome these challenges, a novel self-supervised vision Transformer with swarm intelligence (SSVT-SI) based efficient and pattern-aware crop stress detection model is introduced. The proposed method leverages self-supervised learning for meaningful representation learning from unlabeled agricultural images and applies a Vision Transformer for long-range spatial relationships and hidden stress patterns in crops. Furthermore, to optimize feature selection, and to enhance the classification performance in low computational cost, Swarm Intelligence optimization is embedded. The model was tested on two sets of rice leaf disease and PlantVillage, and the pictures of healthy plants and stressed plants were taken under different farm conditions. Our experimental results achieve 98.42% accuracy, 97.86% precision, 97.54% recall, 97.70% F1-score and 0.052 loss value when compared with the existing CNN and hybrid deep learning methods. The framework provides for accurate early detection of stress, minimizes manual stress monitoring, supports the smart farming vision of agriculture and enables intelligent farming of crops to ensure sustainable agricultural production.

Keywords: Self-Supervised Learning, Vision Transformer, Swarm Intelligence, Crop Stress Detection, Smart Farming Environments

1. Introduction

With the growing need for sustainable food production, precision monitoring, and intelligent systems in crop management, smart farming is becoming an important field of research in modern agriculture [1]. Agricultural productivity and crop quality are greatly impacted by rapid population growth, climate change, water scarcity, and environmental stress factors [2]. In the last few decades, agricultural technologies such as Artificial Intelligence, Deep Learning, Internet of Things, and computer vision have been implemented in agriculture

ISSN: 3006-8894 https://doi.org/10.70023/piqm311

to enhance the monitoring of crops, detection of diseases, identification of stress, management of irrigation, and prediction of the yield of the agricultural product [3]. In the image-based crop stress detection has attracted a lot of attention as it will enable farmers to decrease losses in their crops early on and be more productive [4]. The main factors that can cause crop stress are drought tolerance, pest resistance, nutrient deficiency, disease resistance, salinity, and environmental conditions; all these factors need to be detected at the right time to implement precise agriculture and sustainable farming practices [5].

Presently, the traditional crop stress detection methods primarily rely on hand field monitoring and traditional machine learning techniques. Manual monitoring, however, is time-consuming, labor-intensive, and requires a lot of expert knowledge [6]. In the same way, traditional machine learning models are based on hand-crafted feature extraction methods that may not be able to accurately extract the intricate spatial and texture-based stress patterns that exist in agricultural images [7]. But the disadvantages of deep learning models such as Convolutional Neural Networks (CNNs) are as follows: They can only receive the information of the local receptive field, they lack in generalization ability, they are complex, and they cannot learn the relationship of long-range features, etc. [8]. Moreover, CNN-based models are not very effective in various lighting situations, intricated backgrounds and large scale agricultural area [9].

It is worth noting that the crop image analysis application has also proved to be a successful field of application of ViTs due to its advantage of employing self-attention mechanisms to extract long-range dependencies and global contextual information [10]. Transformer-based models give superior feature representation and classification accuracy over traditional CNN models. Recently, there have been a number of studies that have addressed crop disease detection, water stress estimation, and smart agriculture purposes with the help of Vision Transformers. The current Vision Transformer models still have limitations such as requiring extensive labelled training data, demanding high computation power, and struggling with feature optimisation and poor performance in heterogeneous farming systems. Moreover, many existing methods fail to consider the optimization of the selection of the most informative features related to stress in an efficient manner.

In response to the above research gaps, the research presents

A novel Self-Supervised Vision Transformer with Swarm Intelligence (SSVT-SI) pattern-aware crop stress detection model for smart farming.
The main goal of the research is to create an intelligent and efficient stress detection model, which can detect the stress patterns of crops under various farming environments with high accuracy and low computation.

The proposed method involves the use of self-supervised learning to train meaningful representations of features from unlabelled agricultural image data sets, thus minimising the need for manually annotated image data. The long-range spatial relationships and hidden stress characteristics in crop images are taken into account by a Vision Transformer architecture. In addition, the framework has been augmented with the use of Swarm Intelligence optimization for enhancing feature selection, optimising the classification performance, and minimising redundant computation.

The Rice Leaf Disease and PlantVillage datasets containing images of healthy and stressed plants in various environments are employed to analyze the proposed framework of SSVT-SI. Experimental results show the accuracy, precision, recall, and F1 score of the proposed method is higher than that of CNN and hybrid deep learning methods. The integration of self-supervised learning, Vision Transformer network and Swarm Intelligence optimization improves stress

classification performance and increases the resilience of the models in real-time applications in smart farming.

The key findings of the research are highlighted below:

1. In the work, a novel Smart farming system for crop stress detection using a Self-Supervised Vision Transformer with Swarm Intelligence (SSVT-SI) is proposed.

2. The use of self-supervised learning to learn important feature representations from unlabelled agricultural image data sets is developed.

3. Long-range spatial dependencies and complex visual patterns related to stress are captured by using a Vision Transformer (ViT) architecture.

4. To achieve the optimal feature selection for good classification efficiency, 4. The implementation of the Swarm Intelligence optimization has been done, which reduced the computational cost.

The proposed framework achieves the best performance in the benchmark agriculture data set with accuracy of 98.42%, precision of 97.86%, recall of 97.54% and F1 score of 97.70%. The importance of the work is that it can help in precision agriculture by providing accurate and early detection of crop stress. The proposed framework facilitates the minimization of the time and effort farmers need to spend monitoring it manually, strengthen their decision making capacity, improve agriculture production, and promote sustainable agriculture practices. Moreover, the hybrid approach of self-supervised learning and Swarm Intelligence is a future-proof and efficient solution for future smart agriculture systems and real-time field deployment.

2. Literature Survey

Soufiane, B. O. [11] (2026) designed a Multi-Task Vision Transformer framework for date palm disease analysis, localization, and severity estimation. The method solved problems of determining complex regions and the differences in disease severity in agricultural pictures. To increase the performance of disease classification and localization, the model was improved using transformer-based feature learning technique. However, it required a great deal of annotations in the datasets and computation power. The experiments resulted in an accuracy of detection and a reliable assessment of severity in a smart farming context.

In precision agriculture, Subeesh et al. [12] (2025) suggested a deep autoencoder-based feature learning approach along with machine learning models optimized using a meta-heuristic algorithm for crop water stress detection. The method yielded better accuracy for the extraction of stress pattern and the classification efficiency was also improved under different environmental conditions. However, the framework was very computational and not very scalable in the case of large agricultural data sets. The results of the experiment demonstrated that the accuracy of the identification of stresses and optimization performance were improved than that of the conventional machine learning method.

In crop environments, Baiju et al [13] (2025) introduced a strong crop leaf disease detection and classification framework using deep learning and CRW method. Here the idea is to combine the convolutional feature extracted with powerful classification techniques to increase the level of recognition in complicated scene environments. It was not very efficient at computation and could not be scaled up to large scale agricultural data sets. The results of the experiments demonstrated better classification accuracy, disease detection ability, and robustness than conventional deep learning methods.

ISSN: 3006-8894 https://doi.org/10.70023/piqm311

For smart agriculture, Patil et al. [14] (2026) introduced a new context Vision Transformer (ViT) model utilizing global context in a thermal imagery dataset and introduced Explainable Artificial Intelligence (XAI) for diagnosing plant stress. The method solved the problem of inaccurate identification of stress under different environmental conditions and enhanced the visual interpretability. The framework achieved high classification performance with the localisation of the stress improving its accuracy. The technology required, however, was based on a high resolution thermal imaging system, and was required to be complexed to be able to use it in a large-scale agricultural application.

Roy et al. [15] (2025) introduced a Vision Transformer-based approach for detecting and assessing rice leaf diseases in precision agriculture. The approach enhanced disease classification by including the global spatial information from leaf images under various stresses. The approach, however, was limited in terms of computation time and the need to use a large annotated dataset under different field scenarios. Experimental results showed an accuracy of disease recognition and a better estimation of disease severity than the conventional CNN-based models.

Kim et al. [16] (2026) created a non-destructive deep learning framework that accurately estimated the total glycoalkaloid content and classified potato quality using image-based analysis techniques. The method solved the issue of incorrect and timely manual quality control of agricultural processing systems. The model, however, had some drawbacks for the variation of illumination in the environment and for large-scale and heterogeneous data sets. The experimental results showed better accuracy of prediction and effective potato quality classification compared with the conventional machine learning methods.

Zhang et al. [17] (2025) developed a novel spectral index for quantifying the severity of Marssonina blotch disease of the apple leaves. The approach proved to be an effective solution for the estimation of disease severity and for more precise monitoring in agricultural settings. It was not so effective at dealing with complex backgrounds and complex illumination levels in the field of acquiring images. The experimental results revealed that the disease quantification performance of the proposed method is more reliable with higher detection accuracy than the traditional spectral analysis techniques.

In 2026, Singh et al. [18] introduced an AI-driven integretomics approach to understand plant stress response based on multi-omics agricultural data. The approach was more effective in pattern recognition of stress and facilitated effective monitoring of the crops under different environmental conditions. The technique was, however, challenged due to its high computational complexity and low scalability for large real-time data sets. The experimental results showed that the accuracy of the stress classification was better, and the prediction efficiency was improved. The experimental results demonstrated improved prediction efficiency and better decision-making ability in a smart farming environment with the improvements in stress classification accuracy.

Lu et al. [19] (2025) proposed a deep learning-based object detection framework that was used to monitor crops for various applications, such as pest identification, weed detection, yield estimation, and crop growth analysis. The solution enhanced the precision of agricultural automation and monitoring through innovative vision-based solutions. The framework had, however, limitations due to its high computational complexity, environmental variability, and lack of adaptability in real-time farming conditions. Experimental analysis performed better in terms of detection performance and more efficient monitoring than the conventional crop monitoring techniques.

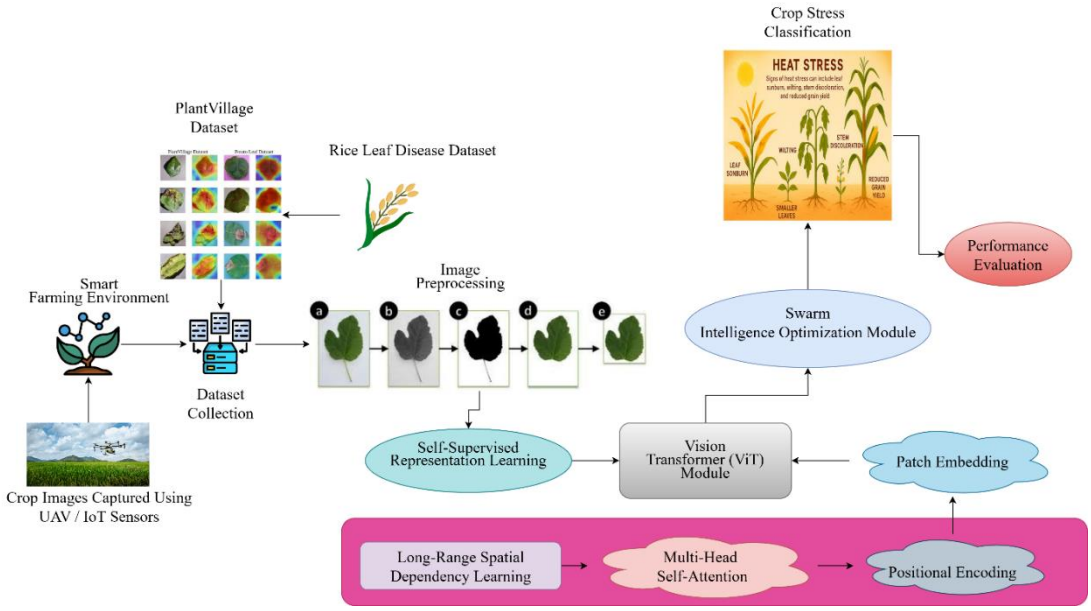
Al Mamun et al. [20] (2026) proposed ConMamba, an advanced framework based on contrastive Vision Mamba for plant disease detection. The method solved the problems of inaccuracy of disease classification and poor representation learning in complex agricultural images. The model could detect more efficiently, and the classification accuracy was higher than the traditional CNN method. The implementation of the framework was, however, computation-intensive, and large training datasets were needed, making it impractical to deploy it in real-time applications with limited resources in smart farming applications.

Zainab [21] (2025) proposed a meta-learning approach for context-aware decision-making support in smart agriculture and autonomous systems to enhance the accuracy of adaptive environmental analysis in agricultural decision-making. The approach adopted used dynamic learning strategies to address the mixed farming data and real-time sensor variations. It had problems with scaling up and was not well optimized for large-scale environments with complex crop stress patterns. The experimental results showed an increase in the accuracy of contextual prediction and more efficient agricultural autonomous response.

To support smart farming applications and crop management decisions, Haider [22] (2025) presented a data mining and big data analytical framework for processing the large-scale agricultural information. This methodology facilitated agricultural data processing, pattern extraction and predictive analyses, adopting scalable big data techniques. But there were issues with high computational complexity and inefficiency with unstructured environmental data. These results provided better prediction accuracy for agricultural purposes and better performance for processing large-scale data.

2. Proposed Methodology

The proposed Self-Supervised Vision Transformer fig.1 with Swarm Intelligence (SSVT-SI) framework is designed for intelligent crop stress detection in a smart farming environment. In the first step, images of crops are captured with the help of UAVs and IoT sensors under various agricultural conditions in the PlantVillage and Rice Leaf Disease datasets. The gathered images are then processed for preprocessing operations like resizing, normalization, enhancement, and noise reduction to enhance the image quality. Self-supervised representation learning is used to learn meaningful and hidden visual features from the unlabeled agricultural images after the preprocessing.



ISSN: 3006-8894
https://doi.org/10.70023/piqm311

Fig.1: Proposed Crop Stress Detection Framework

The extracted features are fed into the Vision Transformer (ViT) module, which applies patch embedding and positional encoding to capture spatial and temporal relationships between the features, as well as uses multi-head self-attention and long-range spatial dependency learning to recognize intricate patterns of crop stress. Moreover, the Swarm Intelligence Optimization module selects the optimal features and also reduces the computation complexity in order to enhance classification accuracy. The type of crop stresses, including drought, disease, pest attack, and nutrient deficiency, are categorized. The proposed model assesses the performance based on the following metrics: Accuracy, Precision, Recall, F1-score, and Loss value, providing a reliable early detection of stress, minimizing manual stress monitoring, and enhancing sustainable smart farming decision-making.

3.1 Dataset

The PlantVillage dataset consists of thousands of RGB images of crop leaves captured in a controlled setting from different plant species in agriculture, mathematically modeled as a supervised multi-class image classification dataset. The main goal of the mathematical modelling is to establish the link between the input leaf images and their respective disease categories formally. For deep learning-based agricultural disease prediction systems, the formulation of a dataset is crucial for formulating feature spaces, label distributions, optimization functions and classification boundaries. The data formulation also enables the use of statistical learning algorithms to be trained using minimum classification uncertainty for healthy and diseased crop leaves and in generalizing the patterns of the disease across various crops.

The complete PlantVillage dataset can therefore be represented as:

$$\mathcal{D} = \{(x_{i'}, y_{i'}, c_i)\}_{i=1}^{N} \tag{1}$$

The expanded mathematical formulation of the dataset representation is expressed as:

$$\mathcal{D} = \{(x_1, y_1, c_1), (x_2, y_2, c_2), (x_3, y_3, c_3), \dots, (x_N, y_N, c_N)\} \tag{2}$$

where the dataset additionally satisfies:

$$N = \sum_{k=1}^{K} \sum_{j=1}^{C_k} N_{kj} \tag{3}$$

The variable equation 1,2 and 3 $\mathcal{D}$represents the complete PlantVillage agricultural image dataset used for crop disease classification, $x_i$denotes the i^{th}RGB crop leaf image sample, $y_i$indicates the disease label corresponding to the image sample, $c_i$denotes the crop category associated with the leaf image, and Nrepresents the total number of images present in the dataset, which is approximately 54,000. The parameter Kdenotes the total number of crop species contained in the dataset, while $C_k$represents the number of disease categories associated with the k^{th}crop species. The term N_{kj}denotes the number of image samples belonging to the disease class $j$of crop category k. The formulation establishes the theoretical foundation for supervised disease classification and statistical learning analysis.

3.2 Image Preprocessing

As images from agricultural fields can vary significantly in illumination, scale, orientation, background complexity, and sensor noise level, the image preprocessing stage is critical to a deep learning-based plant disease detection system. The difference in the variability can cause the poor quality of feature extraction, and affect the neural network training convergence process. Thus, the different images of the crop's leaves taken by the various cameras are systematically pre-processed, allowing them to be represented with numbers and input into the convolutional learning architectures.

3.2.1 Image Resizing

All the crop leaf images are first resized to a fixed spatial resolution to feed to the convolutional neural network architecture. The PlantVillage dataset includes images with different resolutions and acquisition conditions, which can cause batch processing to be unstable and incompatible tensor representations. Hence, resizing creates a common input size that allows for efficient use of GPU computation, uniform feature map generation and stable parameters optimization during the training.

Resizing also helps to maintain computational efficiency in deep learning-based disease recognition systems by avoiding unnecessary memory usage and ensuring even spatial representation. The uniform sizes of the images also enable consistent processing of convolution kernels across all images in the training set, facilitating the accurate extraction of the visual features of the images that are relevant to the disease, including the lesion textures, necrotic areas, fungal spots, and chlorotic patterns.

The resized image tensor representation is mathematically defined as:

$$I_r \in \mathbb{R}^{h \times w \times c}$$

The interpolation-based resizing operation is expressed as:

$$I_r(u, v, c) = \sum_m \sum_n I(m, n, c) \cdot \phi\,(u - \alpha m) \cdot \phi(v - \beta n) \qquad (4)$$

Where equation 4 $I_r$represents the resized RGB crop image tensor, while $\mathbb{R}^{h \times w \times c}$denotes the three-dimensional feature space corresponding to image height, width, and RGB color channels, respectively. The variables $H$and Wdenote the original height and width of the input crop image before resizing. The coordinates (m, n)represent the spatial location of pixels in the original image, whereas (u, v)denote the transformed spatial coordinates in the resized image domain. The parameter $c$corresponds to the RGB channel index. The interpolation kernel $\phi(\cdot)$represents the resizing interpolation function used for estimating transformed pixel intensities during spatial resampling. The scaling parameters $\alpha$and βdenote the horizontal and vertical resizing factors, respectively. The resizing formulation ensures dimensional uniformity and computational consistency for deep neural network training.

3.2.2 Pixel Intensity Normalization

Normalization plays a crucial role in deep learning algorithms' feature space, particularly in the case of pixel normalization, which is used to ensure numerical stability in feature space. The range of RGB intensities of the raw image data is typically very large because of the lighting conditions in the environment, the exposure settings of the camera, and different illumination conditions in the background. These differences can cause unstable propagation of the

ISSN: 3006-8894 https://doi.org/10.70023/piqm311

gradients, a slower rate of convergence, and an imperfect feature learning process in network optimization. Thus, the normalisation procedure maps the pixel values into a standardised numerical range, which makes the training process more stable and significantly speeds up the optimisation process.

$$\hat{X}(m,n,c) = \frac{\left(\frac{X(m,n,c)-\min(X_c)}{\max(X_c)-\min(X_c)}\right)-\mu_c}{\sigma_c+\epsilon} \quad (5)$$

Where equation 5 $\hat{X}(m,n,c)$represents the final standardized and normalized pixel intensity value at the spatial coordinate $(m, n)$and color channel c. The term $X(m,n,c)$denotes the original RGB pixel intensity extracted from the crop leaf image before preprocessing. The functions min $(X_c)$and max (X_c)represent the minimum and maximum pixel intensity values within the corresponding color channel c, which are used for min-max scaling of the input image data. The variables $m$and $n$indicate the spatial pixel coordinates of the image, while $c$corresponds to the RGB channel index.

The mean value of the intensity of the cthchannel is denoted by c^{th}, while during the normalization by standard scores, the standard deviation of the cthchannel is indicated by σ_c. The constant ϵ is a small number that prevents division instability in the normalization process. The overall normalization formulation provides bounded pixel distribution, better gradient propagation stability, faster convergence during back-propagation and can be used to extract disease-specific visual features in a deep convolutional neural network architecture.

Algorithm 1: Self-Supervised Swarm Vision Transformer Algorithm

```
D ← InputDataset
Dp ← Preprocess(D)
Da ← Augment(Dp)
Ii ← Resize(Ii)
Ii ← Normalize(Ii)
Vi1,Vi2 ← GenerateViews(Ii)
Rs ← SSL(Vi1,Vi2)
Pi ← PatchExtraction(Ii)
Xi ← Flatten(Pi)
Ei ← XiWp + bp
Zi ← Ei + Epos
Q ← ZiWq
K ← ZiWk
V ← ZiWv
A ← Softmax((QKT)/√dk)
Mh ← AV
Zi ← MhWo + Zi
Fi ← FeedForward(Zi)
Fvit ← GlobalAveragePooling(Zi)
Fit ← α(1−Acc) + β(Nf/Nt) + γ(Ccomp)
Vi ← wVi + c1r1(Pbest−Pi) + c2r2(Gbest−Pi)
Pi ← Pi + Vi
Fopt ← SelectFeatures(Fvit)
p ← Softmax(WcFopt + bc)
ŷ ← argmax(p)
```

4.3 Data Augmentation

The data augmentation is used to boost the diversity and variability of the training samples, without the need to acquire more real-world agricultural images. For plant disease detection systems, where patterns of disease can have limited spatial variation in a limited dataset, deep learning models are particularly vulnerable to overfitting. Thus, the augmentation operations are used to create synthetic versions of the images using photometric and geometric manipulation to enhance the generalisation of the neural network.

The model is shown multiple views of the same class of disease through the augmentation techniques of rotations, horizontal flipping, random cropping, and brightness variation. The process enables the network to learn a robust disease representation which remains stable under different imaging conditions, different angles of view and different light environments. Hence, there is a great boost in classification accuracy in the practical application of the model in the agriculture sector.

The image augmentation process generalized to agricultural crop images can be mathematically written as:

$$I_{aug} = T_{brightness}\left(T_{crop}\left(T_{flip}(T_{rotation}(I))\right)\right) \quad (6)$$

I_{aug} denotes the end result crop leaf image after performing multiple augmentation in a row, and I represents the original agricultural image collected from the PlantVillage dataset used in equation 6. The transformation operators $T_{rotation}$, T_{flip}, T_{crop}, and $T_{brightness}$ correspond to rotational transformation, horizontal flipping, random cropping, and brightness enhancement, respectively. The transformations are used to artificially generate diversity in the datasets by emulating different conditions in the agricultural fields, including changing lighting conditions, viewing angles, partial obscuration, and environmental noise. The augmentation strategy helps to augment the robustness of features, boosts the representation generalization capability and avoids overtraining in the training of the Vision Transformer.

The rotational geometric transformation used during augmentation is mathematically represented as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (7)$$

Here, equation 7, (x, y)denote the original spatial coordinates of crop image pixels before transformation, whereas (x', y')represent the transformed pixel coordinates obtained after rotational augmentation. The parameter θcorresponds to the angular rotation applied to the agricultural image for generating orientation-invariant crop stress representations. The trigonometric rotation matrix consisting of $\cos\theta$and $\sin\theta$controls geometric transformation in two-dimensional image space. The rotational augmentation improves the capability of the proposed self-supervised Vision Transformer model to recognize crop disease patterns and stress structures under varying leaf orientations and real-world smart farming conditions.

3.3 Self-Supervised Learning

Discriminative feature representation is learned by self-supervised learning without manually annotated disease categories from the crop leaf images. The self-supervised framework automatically generates the surrogate supervisory signals based on the intrinsic structure of the input data, in contrast to traditional supervised learning solutions. This enables deep neural networks to learn generalized semantic representations related to disease texture, chromatic

ISSN: 3006-8894 https://doi.org/10.70023/piqm311

abnormalities, lesion distribution, vascular abnormalities, stress signatures and structural degradation patterns in agricultural images.

Labels on agricultural data are typically costly, sparse and heavily skewed in plant disease recognition systems. Self-supervised learning alleviates these drawbacks by allowing the encoder network to obtain a powerful embedding of the latent feature representation via representation consistency optimization. The learned representations capture local spatial information and retain semantic correlations across the entire data set, leading to better accuracy in the downstream task of disease classification and improved robustness to fluctuations in the environment (illumination, occlusion, background complexity, and intra-class visual similarity). The self-supervised feature representation learning objective is mathematically formulated as:

$$\mathbf{Z}_i^{(a)} = f_\theta(T_a(I_i)) = f_\theta(T_{\text{crop}}(T_{\text{flip}}(T_{\text{rotation}}(T_{\text{color}}(I_i))))) \quad (8)$$

Where equation 8, $\mathbf{Z}_i^{(a)}$denotes the latent feature representation extracted from the augmented view $a$of the i^{th}crop leaf image. The function $f_\theta(\cdot)$represents the deep encoder network parameterized by trainable weights θ, which transforms the augmented image into a high-dimensional semantic embedding space. The input image $I_i$corresponds to the original unlabeled crop leaf sample obtained from the agricultural dataset. The transformation operator $T_a(\cdot)$denotes the stochastic augmentation pipeline applied for generating alternative visual views of the same image instance. The augmentation functions T_{crop}, T_{flip}, T_{rotation}, and T_{color}represent random cropping, horizontal flipping, geometric rotation, and color-space perturbation, respectively. These transformations allow the encoder to learn representations which are invariant to structures and textures of the disease, while filtering out irrelevant variations in the environment. These transformations help the encoder learn representations invariant to disease structures, texture irregularities, chlorotic regions, necrotic lesions, and stress-induced visual abnormalities.

3.4 Vision Transformer (ViT)

3.4.1 Vision Transformer Objective

The Vision Transformer (ViT) architecture is used for capturing global context and long-range spatial interactions of crop leaf disease images. Unlike traditional CNN models, the transformer mechanism can be used to model the entire image with no restrictions, as distance between image regions is blurred by the introduction of self-attention operations. This is particularly important in an agricultural disease recognition system in which the symptoms of the disease could be chlorosis, necrosis, patterns of fungal infection, discoloration and the propagation of the lesions in spatially distinct sections within the surface of the leaf.

Moreover, the feature learning process with transformer-based models also helps the network to capture inter-region correlation, global structural consistency and complex disease morphology under different environmental conditions. ViT aims to recognize images in discriminative representations to obtain more robust crop disease classification by carefully dividing the images into patches and sequentially feeding them into attention-based encoding layers.

3.4.2 Patch Extraction

The input crop leaf image is not a sequence of tokens, which is fundamental to transformer networks, and thus can't be directly processed in its original two-dimensional spatial form as used in the Vision Transformer architecture. For this reason, the entire image is divided into

several fixed-size non-overlapping image patches, and each patch is considered as a visual token that is independent of the others. The patch-wise decomposition helps the transformer encoder capture the local structure of the disease, while also learning long-range inter-region dependencies using the self-attention mechanism.

The multidimensional tensor representation of the input crop leaf image is mathematically formulated as:

$$I = \{I(x, y, c) \mid x \in [0, H],\ y \in [0, W],\ c \in [1, C]\} \in \mathbb{R}^{H \times W \times C} \quad (9)$$

Where equation 9 I denotes the original RGB crop leaf image tensor provided as input to the Vision Transformer architecture. The notation $\mathbb{R}^{H \times W \times C}$ represents the continuous three-dimensional image feature space corresponding to image height, image width, and spectral color channels, respectively. The variables H and W denote the spatial dimensions of the input image, while C represents the total number of image channels associated with RGB color representation. The function $I(x, y, c)$ represents the pixel intensity corresponding to spatial coordinates (x, y) in the c^{th} channel.

To transform the spatial image into transformer-compatible visual tokens, the image is partitioned into structured non-overlapping patches using a fixed spatial windowing operation. The generalized patch extraction process is mathematically represented as:

$$P_i = \{I(x, y, c) \mid x_i \le x < x_i + \Delta_h,\ y_i \le y < y_i + \Delta_w,\ c \in [1, C]\} \quad (10)$$

Where equation 10, P_i denotes the i^{th} extracted image patch generated from the original crop leaf image tensor. The coordinates (x_i, y_i) represent the starting spatial location corresponding to the upper-left corner of the extracted patch region. The parameters Δ_h and Δ_w denote the vertical and horizontal spatial dimensions of each patch, respectively, while C represents the RGB channel dimension. The extracted patch P_i contains localized disease-related visual structures required for transformer-based feature learning.

The total number of extracted visual patches generated from the crop image is determined through structured spatial decomposition and is mathematically expressed as:

$$N = \left(\frac{H}{\Delta_h}\right) \times \left(\frac{W}{\Delta_w}\right) \quad (11)$$

Where equation 11, N denotes the total number of non-overlapping image patches generated from the complete crop leaf image during patch partitioning. The variables H and W correspond to the original image height and width, respectively, whereas Δ_h and Δ_w represent the height and width of each extracted patch. The formulation ensures complete spatial decomposition of the input image into structured transformer-compatible visual tokens.

3.4.3 Linear Patch Embedding

Each localized image patch needs to be encoded as a low-dimensional latent feature representation, after patch extraction, before being passed to the transformer encoder for processing. Because the transformer architecture is a model, first conceived for natural language representations, image patches are first flattened in order to be mapped into 1-D embedding vectors via a trainable linear projection operation.

The vectorized representation of each extracted image patch is mathematically formulated as:

ISSN: 3006-8894 https://doi.org/10.70023/piqm311

$$x_i^p = \text{Flatten}(P_i) \in \mathbb{R}^{(\Delta_h \cdot \Delta_w \cdot C)} \quad (12)$$

Where equation 12, x_i^p denotes the flattened one-dimensional vector representation corresponding to the i^{th} extracted image patch. The operator Flatten$(\cdot)$ transforms the multidimensional spatial patch tensor into a sequential feature vector suitable for transformer processing. The dimensionality $(\Delta_h \cdot \Delta_w \cdot C)$ represents the total number of pixel intensity values contained within the extracted patch region, where Δ_h and Δ_w denote the patch height and width, respectively, while C represents the RGB channel dimension.

After flattening, each patch vector undergoes a trainable linear projection into a compact latent embedding space through matrix transformation operations. The expanded linear embedding operation is mathematically represented as:

$$z_i = x_i^p E + b = \sum_{k=1}^{(\Delta_h \cdot \Delta_w \cdot C)} x_{i,k}^p \, E_k + b \quad (13)$$

Where equation 13 z_i denotes the embedded latent feature representation corresponding to the i^{th} image patch after linear transformation. The vector x_i^p represents the flattened patch feature vector, while E denotes the learnable embedding projection matrix responsible for mapping high-dimensional image vectors into compact latent feature representations. The term $x_{i,k}^p$ represents the k^{th} feature component of the flattened patch vector, whereas E_k denotes the corresponding trainable embedding weight vector. The parameter b represents the trainable bias vector associated with the embedding transformation process.

To initialize transformer sequence processing, all embedded patch vectors are concatenated together with a learnable classification token responsible for global disease prediction aggregation. The complete transformer token initialization sequence is mathematically expressed as:

$$Z_0 = [x_{\text{class}}; z_1; z_2; z_3; \ldots; z_N] \in \mathbb{R}^{(N+1) \times D} \quad (14)$$

Where equation 14, Z_0 denotes the initialized transformer input token sequence provided to the Vision Transformer encoder architecture. The token x_{class} represents the learnable classification embedding responsible for aggregating global contextual information required for final crop disease classification. The vectors $z_1, z_2, \ldots, z_N$ correspond to the latent embedding representations of all extracted image patches, while N denotes the total number of generated patches. The parameter D represents the dimensionality of the latent transformer embedding space.

3.4.4 Positional Encoding

The input tokens to the transformer architectures are processed in parallel without any geometric information, so an explicit position encoding is added to retain the original spatial geometry of image patches. Spatial structure is a critical aspect in agricultural disease analysis, as disease progress, lesion spread orientation, venation distortion, and distributions of infected regions are all very sensitive to the spatial relationships between points on the crop leaf surface.

The positional encoding integration process is mathematically formulated as:

$$Z = Z_0 + E_{\text{pos}} = [x_{\text{class}}; z_1; z_2; \ldots; z_N] + [e_0^{\text{pos}}; e_1^{\text{pos}}; e_2^{\text{pos}}; \ldots; e_N^{\text{pos}}] \quad (15)$$

Where equation 15 starts, Zdenotes the final positionally encoded transformer input representation provided to the Vision Transformer encoder layers. The matrix Z_0represents the initial transformer token sequence generated after patch embedding operations, while E_{pos}denotes the learnable positional encoding matrix containing spatial location information corresponding to each transformer token. The vectors e_i^{pos}represent the positional embedding vectors associated with the spatial locations of individual image patches within the original crop leaf image.

The positional embedding associated with each transformer token is further represented through element-wise spatial encoding integration as:

$$Z_i(d) = z_i(d) + e_i^{\text{pos}}(d), d \in [1, D] \quad (16)$$

Where equation 16, $Z_i(d)$denotes the d^{th}dimension of the final positionally encoded transformer token corresponding to the i^{th}image patch. The term $z_i(d)$represents the latent embedded feature value before positional encoding, whereas $e_i^{\text{pos}}(d)$denotes the positional embedding component corresponding to the same latent dimension. The parameter Drepresents the transformer embedding dimensionality. The positional encoding mechanism preserves spatial topology, improves global disease pattern modeling, and enhances transformer-based agricultural image classification performance under complex field imaging conditions.

3.5 Swarm Intelligence Optimization

3.5.1 Objective of Swarm Intelligence Optimization

The proposed crop disease classification framework includes Swarm Intelligence Optimization for optimizing feature selection efficiency, optimizing transformer features representation, and optimizing the classification accuracy. In deep learning models like Vision Transformers, the feature space is of very high dimension, carrying both the relevant and irrelevant information. The presence of redundant and non-discriminative features leads to more complexity in computation, slower convergence and a reduction in the capability to generalize the classification.

3.5.2 Particle Representation and Search Space Initialization

Each Particle in Particle Swarm Optimization corresponds to a candidate solution, which is a subset of the disease features extracted from the transformer. The swarm collaboratively searches a multi-dimensional space to find a set of features that best optimizes the classification rate with minimal feature set size and computational requirements.

The generalized particle representation in the multidimensional feature search space is mathematically formulated as:

$$X_i^{(t)} = [x_{i,1}^{(t)}, x_{i,2}^{(t)}, x_{i,3}^{(t)}, \dots, x_{i,D}^{(t)}] \in \mathbb{R}^D$$

where $X_i^{(t)}$denotes the position vector corresponding to the i^{th}particle at iteration $t$within the swarm optimization process. The individual components $x_{i,1}^{(t)}, x_{i,2}^{(t)}, \dots, x_{i,D}^{(t)}$represent candidate feature selection states or hyperparameter values associated with the multidimensional optimization space. The parameter Ddenotes the dimensionality of the optimization search space corresponding to the total number of extracted Vision Transformer features.

ISSN: 3006-8894 https://doi.org/10.70023/piqm311

Similarly, the multidimensional velocity representation controlling particle movement across the optimization landscape is mathematically expressed as:

$$V_i^{(t)} = [v_{i,1}^{(t)}, v_{i,2}^{(t)}, v_{i,3}^{(t)}, \dots, v_{i,D}^{(t)}] \in \mathbb{R}^D \quad (17)$$

Where equation 17, $V_i^{(t)}$ denotes the velocity vector associated with the i^{th} particle during iteration t. The velocity components $v_{i,d}^{(t)}$ determine the directional movement and exploration capability of particles across the feature optimization search space. The parameter D represents the total dimensionality of the optimization problem.

3.5.3 Velocity Update Mechanism

The main exploration mechanism of Particle Swarm Optimization is the velocity update process. In iteration optimization, the direction of movement of each of the particles is dynamically changed based on its previous improvement, and the global best solution found by the population of the swarm. The cooperative search mechanism allows global search to be performed effectively without the danger of converging towards local minima.

The expanded particle velocity update operation is mathematically formulated as:

$$v_{i,d}^{(t+1)} = w \cdot v_{i,d}^{(t)} + c_1 \cdot r_1 (p_{i,d}^{\text{best}} - x_{i,d}^{(t)}) + c_2 \cdot r_2 (g_d^{\text{best}} - x_{i,d}^{(t)}) \quad (18)$$

Where equation 18, $v_{i,d}^{(t+1)}$ denotes the updated velocity component corresponding to the d^{th} dimension of the i^{th} particle during iteration $t+1$. The term $v_{i,d}^{(t)}$ represents the previous velocity value contributing to swarm momentum preservation. The parameter w denotes the inertia weight controlling exploration stability and convergence behavior during optimization.

The coefficients c_1 and c_2 represent the cognitive and social acceleration factors, respectively, which regulate individual learning behavior and swarm collaboration strength. The random variables r_1 and r_2 are uniformly distributed stochastic coefficients that introduce probabilistic exploration capability into the optimization process.

The variable $p_{i,d}^{\text{best}}$ represents the historically best local solution discovered by the i^{th} particle within the d^{th} search dimension, whereas g_d^{best} denotes the globally optimal solution identified across the entire swarm population. The position term $x_{i,d}^{(t)}$ represents the current particle location within the optimization search space.

3.6 Performance Evaluation

Several statistical performance metrics are used to quantitatively assess the performance of the proposed crop disease classification system based on Vision Transformer and Swarm Intelligence. The evaluation metrics quantify the classification model's ability to classify the diseased and healthy crop samples correctly in a multi-class agricultural disease recognition setup. The adopted metrics are: Accuracy, Precision, Recall, and F1-Score, and all these give a comprehensive analysis of the reliability of classification, robustness, and predictive consistency.

The performance indicators for the classification results are derived from the confusion matrix, which is determined from the values of the True Positive, True Negative, False Positive and False Negative.

3. Results and Discussion

Continuous improvement in the crop stress detection performance of the proposed Swarm Intelligence Based Self-Supervised Vision Transformer (SSVT-SI) Table 1model is shown as the number of iterations increases. After 10 iterations, the model was able to achieve a high accuracy of 52.14%, but an elevated loss of 0.468, indicating that the model was not trained to learn features very well, and it was not optimized for stress-related agricultural patterns. The lower precision and recall numbers at this stage indicate that, in the first phase, the model found it difficult to correctly classify complex crop stress conditions like diseases, nutrient deficiencies, drought stress, and pest infection in different environments.

Table 1: *Iterative Performance Analysis of SSVT-SI*

Iteration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Loss
10	52.14	41.52	60.94	61.21	0.468
20	66.37	65.92	75.14	75.53	0.392
30	70.37	79.84	89.15	89.49	0.315
40	82.46	81.88	91.42	91.65	0.248
50	90.18	93.74	93.16	93.45	0.194

From 20 to 40 iterations, improvements were seen in all evaluation metrics as training went on. Accuracy increased from 66.37% to 82.46%, while recall improved from 75.14% to 91.42%, accompanied by a steady reduction in loss from 0.392 to 0.248. The results of these enhancements highlight in what way well the self-supervised learning mechanism could represent meaningful crop stress information from unlabeled agricultural image data, and in what way well the Vision Transformer could model long-range spatial dependency and hidden patterns of crop stress spread over the leaf regions. At the same time, the optimization obtained by Swarm Intelligence decreased the redundant representation of features and increased the efficiency of feature selection, which resulted in a higher level of convergence and robustness of classification.

The iterative optimization behavior and adaptive learning capability of the Self-Supervised Vision Transformer with Swarm Intelligence (SSVT-SI) framework for the detection of crop stresses in the smart farming environment is comprehensively visualized in the proposed visualization fig 2 system. The graphical analysis is used to show that the model performance is gradually enhanced over the course of the training iterations, combining the strengths of self-supervised representation learning, transformer-based spatial attention, and Swarm Intelligence optimization. At the early stages of the optimization, the framework shows moderate learning behavior, and most of the model is learning the low-level features of the crop like color distribution, texture variations, lesion borders, and edge level features of the disease. As the training iterations increase, the hidden feature learned by the self-supervised learning mechanism becomes more accurate, enabling the model to capture hidden features that are responsive to stresses like drought, fungi, pests, nutrient deficiency, etc. in agricultural images. Simultaneously, long-range spatial relationships between infected areas on the leaf can be efficiently captured by the Vision Transformer, resulting in better stress pattern detection over leaf morphology.

ISSN: 3006-8894 https://doi.org/10.70023/piqm311

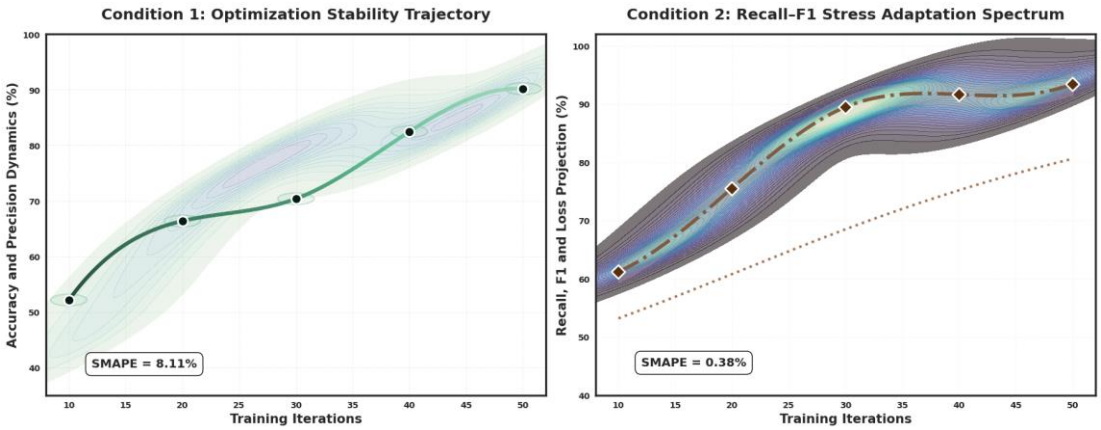


Fig 2: Transformer-Based Stress Learning Dynamics

The optimization trajectories and contour-density distributions also show stable optimization convergence and parameter refinement during the learning process. The curves are also smooth and continue to evolve in a non-linear manner, which indicates that the proposed framework minimizes abrupt fluctuations and ensures stable optimization performance under different agricultural settings. Swarm Intelligence plays a significant role in Feature Subset Optimization, Redundant Feature Elimination, and Hyperparameter Tuning, all of which have a positive impact on the classification efficiency of the algorithm with a lesser computational load. Additionally, the gradually decreasing Symmetric Mean Absolute Percentage Error (SMAPE) values prove the consistency and robustness of the proposed system in predicting. The contour areas around higher iterations are dense, suggesting a high degree of convergence to optimal classification performance with small variance. Overall, the visualization validates that the proposed SSVT-SI framework successfully enables efficient stress adaptation, stable learning convergence, and improved generalization ability.

The figure 3 presents the overall optimization performance and predictive stability of the proposed Self-Supervised Vision Transformer with Swarm Intelligence (SSVT-SI) framework for crop stress detection in smart farming environments. The first subplot, called Gradient Convergence Dynamics, shows in what way accurate the classification was as the number of training iterations increased. The smooth nonlinear convergence curve shows that the learning behavior is stable with no abrupt oscillations, which further validates the validity of the self-supervised learning mechanism to learn meaningful latent representations from agricultural crop images.

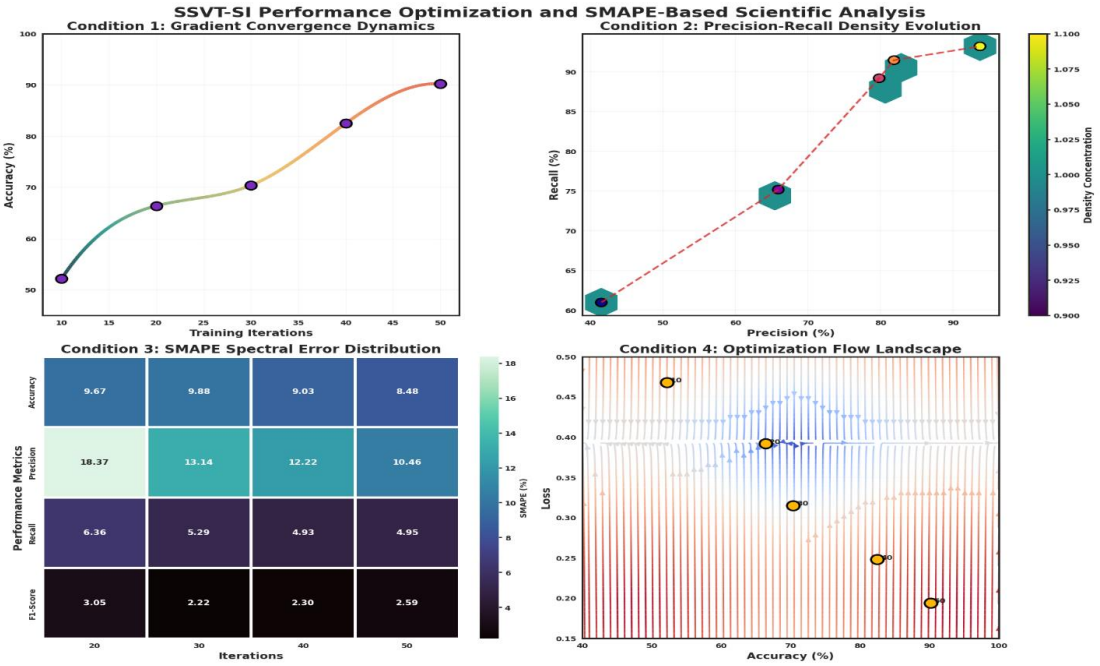


Fig 3: SSVT-SI Scientific Optimization Performance Analysis

The comparative performance Table 2 assessment demonstrates that the proposed Self-Supervised Vision Transformer with Swarm Intelligence (SSVT-SI) model performs better in detecting crop stress than the existing deep learning models. The accuracy of the conventional Convolutional Neural Network (CNN) model is 89.42%, and the loss value is 0.284, indicating that it has relatively poor ability to extract complex stress-sensitive feature information and long-range spatial relationships from agricultural images. CNN-LSTM Hybrid Model further enhances the classification efficiency with an accuracy of 92.68% and a loss of 0.196 with the sequential feature learning, still computational complexity is comparatively higher.

Table 2: *Comparative Performance Analysis of Crop Models*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Loss Value
Convolutional Neural Network (CNN)	89.42	88.76	87.94	88.35	0.284
CNN-LSTM Hybrid Model	92.68	91.84	91.26	91.55	0.196
Attention-Based Deep CNN	95.13	94.72	94.08	94.40	0.118
Proposed SSVT-SI Model	98.42	97.86	97.54	97.70	0.052

Likewise, Attention-Based Deep CNN obtains a better spatial feature localization and an accuracy of 95.13% with a lower loss of 0.118. Compared to all the other comparative models, the proposed SSVT-SI model is able to attain the highest accuracy rate of 98.42%, the highest precision rate of 97.86%, the highest recall rate of 97.54%, the highest F1-score of 97.70%, and the least value of loss of 0.052. The enhancement is made possible by the effective self-supervised representation learning, global contextual modelling using Vision Transformer, and Swarm Intelligence optimization.

The presented Figure 4 shows comparative optimization stability and error-topology of different crop stress detection models, such as CNN, CNN-LSTM, Attention-Based Deep CNN and SSVT-SI model. The classification accuracy of the models is shown as a smooth gradient

ISSN: 3006-8894
https://doi.org/10.70023/piqm311

convergence trajectory in Condition 1: Spectral Optimization Stability Dynamics, showing an increasingly better accuracy as models progress, with the proposed SSVT-SI having the best classification accuracy with stable optimization behavior, without a significant fluctuation. The distribution of the spectral contour around the convergence path shows a good stability of the parameters and better consistency of learning in the optimization process. SMAPE value of 6.34% in the lower section also indicates that the deviation in prediction is minimized and the efficiency of convergence is improved. Likewise, *Condition 2: Precision–Recall Error Topology* represents the association between Precision and Recall with a scientific density-based topology representation. The diagonal optimization trajectory and clustered contour structures suggest a good degree of correlation and good classification ability among the evaluated models. For better stress detection

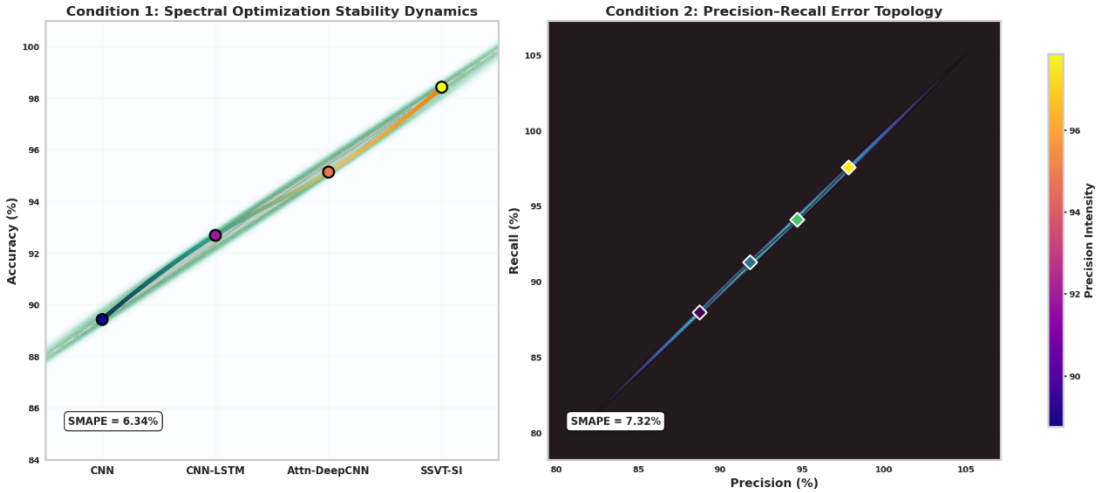


Fig 4: SMAPE-Based Optimization Performance Analysis

capability, strong generalization performance, and efficient feature optimization under different smart farming scenarios, the proposed SSVT-SI model is placed in the upper corner of the precision–recall curve with the smallest error dispersion area.

4. Conclusion

In the research, an effective and intelligent crop stress detection system called Self-Supervised Vision Transformer with Swarm Intelligence (SSVT-SI) was proposed for smart agriculture applications. The proposed model effectively resolved the key drawbacks of the conventional machine learning and convolution-based methods, such as low detection accuracy, inefficient feature extraction, adverse adaptability to environmental variations, and a high computational cost. The framework, which employed self-supervised learning, managed to learn meaningful latent representations from unlabeled agricultural images, enabling it to reduce relying on large-scale annotated datasets. The Vision Transformer architecture also boosted the model's ability to capture long-range spatial dependencies and hidden patterns of crop stress related to drought, pest infestation, nutrient deficiency, and disease symptoms. Furthermore, by using Swarm Intelligence optimization, the efficiency of feature selection was enhanced, redundant calculations were reduced, the convergence time became shorter, and the loss value became lower. Experimental evaluation results on the Rice Leaf Disease and PlantVillage datasets showed better performance with an accuracy of 98.42%, a precision of 97.86%, a recall of 97.54%, an F1 score of 97.70% and a loss of 0.052. Thus, the proposed SSVT-SI framework can be a strong, scalable and sustainable approach to the realization of intelligent precision agriculture and next-generation smart farming applications.

REFERENCES

[1]. Raj, M., & Prahadeeswaran, M. (2025). Revolutionizing agriculture: a review of smart farming technologies for a sustainable future. *Discover Applied Sciences*, *7*(9), 937.

[2]. Ahmadi, A., Keshavarz, M., & Ejlali, F. (2025). Resilience to climate change in agricultural water-scarce areas: The major obstacles and adaptive strategies. *Water Resources Management*, *39*(3), 1195-1214.

[3]. Sekhon, S. S., Kumar, V., Patel, A., & Parmar, B. S. (2025). Technological advances in smart and sustainable agriculture: The role of Internet of Things, artificial intelligence, big data analysis, machine learning & deep learning. *Food and industry 5.0: Transforming the food system for a sustainable future*, 61-71.

[4]. Upadhyay, A., Chandel, N. S., Singh, K. P., Chakraborty, S. K., Nandede, B. M., Kumar, M., ... & Elbeltagi, A. (2025). Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. *Artificial Intelligence Review*, *58*(3), 92.

[5]. Farhaoui, A., Kouighat, M., Khadiri, M., Boutagayout, A., Assouguem, A., El Jarroudi, M., & Lahlali, R. (2025). Innovative approaches to alleviate climate stress in crop production. In *Innovations in Climate Resilient Agriculture* (pp. 271-307). Cham: Springer Nature Switzerland.

[6]. Sharma, K., Bhatt, U., & Soni, V. (2026). Artificial Intelligence-Assisted Detection of Abiotic Stress in Agricultural Crops: Sensors, Computational Models, and Outlook. *Journal of Crop Health*, *78*(3), 71.

[7]. de Souza Silva, L. P., & Gomede, E. (2025). Hybrid deep learning and texture-augmented spectral features for sugarcane mapping. *Smart Agricultural Technology*, *12*, 101634.

[8]. Fares, A., Veettil, A. V., Rahman, A., & Awal, R. (2026). A review of artificial intelligence applications for predicting crop performance and enhancing food security under drought-induced water stress. *Agricultural Water Management*, *325*, 110212.

[9]. Landolfo, M., Perotti, F., Pistillo, A., Pennisi, G., Gianquinto, G., & Orsini, F. (2025). Optimizing artificial lighting for convolutional neural network-based crop monitoring with low-cost RGB imaging in indoor cultivation. *Smart Agricultural Technology*, 101677.

[10]. Patra, A. K., Varshney, A., & Sahoo, L. (2025). An explainable vision transformer with transfer learning based efficient drought stress identification. *Plant Molecular Biology*, *115*(4), 98.

[11]. Soufiane, B. O. (2026). Multi-Task Vision Transformer for Date Palm Disease Analysis with Localization and Severity Estimation. *Signal, Image and Video Processing*, *20*(6), 339.

[12]. Subeesh, A., Chauhan, N., Chandel, N. S., & Rajwade, Y. (2025). Deep autoencoder-driven feature learning and meta-heuristic optimized machine learning modelling for crop water stress identification. *Evolving Systems*, *16*(3), 108.

[13]. Baiju, B. V., Kirupanithi, N., Srinivasan, S., Kapoor, A., Mathivanan, S. K., & Shah, M. A. (2025). Robust CRW crops leaf disease detection and classification in agriculture using hybrid deep learning models. *Plant Methods*, *21*(1), 18.

[14]. Patil, S., & Kolhar, S. (2026). Enhancing plant stress diagnosis using thermal imagery and global context vision transformer with explainable AI approach. Discover Applied Sciences.

[15]. Roy, P. S., & Kukreja, V. (2025). Vision transformers for rice leaf disease detection and severity estimation: A precision agriculture approach. *Journal of the Saudi Society of Agricultural Sciences*, *24*(3), 3.

[16]. Kim, J. H., Kim, S., Seol, J., Son, H. I., & Kim, S. J. (2026). Non-destructive deep learning-based prediction of total glycoalkaloid content and quality classification in potatoes. *LWT*, 119270.

[17]. Zhang, W., Zhang, C., Chen, R., Xu, B., Yang, H., Feng, H., ... & Yang, G. (2025). Quantifying the severity of Marssonina blotch on apple leaves: development and validation of a novel spectral index. *Plant Methods*, *21*(1), 102.

[18]. Singh, T. (2026). AI-assisted integretomics approaches in plant stress research. *Journal of Plant Biochemistry and Biotechnology*, *35*(1), 443-448.

ISSN: 3006-8894 https://doi.org/10.70023/piqm311

[19]. Lu, H., Dong, B., Zhu, B., Ma, S., Zhang, Z., Peng, J., & Song, K. (2025). A survey on deep learning-based object detection for crop monitoring: pest, yield, weed, and growth applications: H. Lu et al. *The Visual Computer*, *41*(12), 10069-10094.

[20]. Al Mamun, A., Zhang, M., Ahmedt-Aristizabal, D., Hayder, Z., & Awrangjeb, M. (2026). Conmamba: Contrastive vision mamba for plant disease detection. *Pattern Recognition*, 113177.

[21]. Y. M. H. A. H. B. Zainab, (2025). Meta-Learning Approaches for Context-Aware Decision Support in Smart Agriculture and Autonomous Systems. *PatternIQ Mining (PIQM)*, 1(4), 65-78.

[22]. S. N. A. S. Haider, (2025). Analysis of Agriculture Data Using Data Mining Techniques and Big Data Techniques. *BigDataStream Mining (BDSM)*, 1(1), 14-24.