Asia Pacific Academy of Science Pte. Ltd.

Journal of Biological Regulators and Homeostatic Agents 2026, 40(3), 121.
https://doi.org/10.65746/jbrha121

Article

Enhanced hybrid vision transformer for zero-shot plant leaf disease classification

Pavithra Elangovan[1,*], A. M. J. MD. Zubair Rahman[2]

[1] Department of Computer Science and Engineering, AL-Ameen Engineering College (Autonomous), Erode 638104, Tamil Nadu, India
[2] Department of Electronics and Communication Engineering, AL-Ameen Engineering College (Autonomous), Erode 638104, Tamil Nadu, India
*** Corresponding author:** Pavithra Elangovan, pavithra.ameen@outlook.com

CITATION

Elangovan P, Rahman AMJMDZ. Enhanced hybrid vision transformer for zero-shot plant leaf disease classification. Journal of Biological Regulators and Homeostatic Agents. 2026; 40(3): 121. https://doi.org/10.65746/jbrha121

ARTICLE INFO

Received: 03 February 2026
Revised: 20 February 2026
Accepted: 06 March 2026
Available online: 25 June 2026

Abstract: Plant leaf disease classification is essential in the agriculture industry, and recent advances in deep learning (DL) and machine learning (ML) have resulted in numerous approaches for detecting and classifying diseases using plant images. However, traditional diagnosis methods rely on human expertise and remain time-consuming and labor-intensive. To address this issue, the proposed system introduces a novel approach, Enhanced Hybrid Vision Transformer with Zero-shot learning classification (EHVZSC), for plant species identification. The method combines the strengths of Vision Transformers (ViT) and Zero-Shot Learning, enabling accurate classification of unseen classes without additional training data. The proposed approach leverages ViT to learn robust image representations, which are then used to generate a set of prototypes for zero-shot classification. Evaluated on the Plant Village dataset The model was trained with 50 epochs, a batch size of 32, and a learning rate of 0.0001 because these parameters provided stable convergence without overfitting, the proposed EHVZSC model achieves state-of-the-art performance with reduced data requirements, improving testing accuracy by up to 15 %, sensitivity by up to 10 %, specificity by up to 10 %, F1-score by up to 12 %, and ROC performance by up to 25 % over existing methods, while attaining 95.7 % accuracy, 97.8 % sensitivity, 95.0 % specificity, and a 96.45 % F1-score, demonstrating its superior ability to capture fine-grained disease features through attention-enhanced representation learning and robust zero-shot adaptability.

Keywords: plant leaf disease classification, vision transformer (VIT), zero-shot learning, deep learning, image recognition, prototype-based classification

1. Introduction

India is a well-known country in the field of agriculture, where the vast majority of people work in the sector. Enhancing food quality and productivity at lower costs and greater profit is the goal of agricultural research. Agrochemicals, seeds, and soil interact in a complex way to produce the agriculture production model. The essential agricultural goods are fruits and vegetables. Transpiration, photosynthesis, fertilization, pollination, and germination are just a few of the vital processes that the illness signals as being compromised in a normal plant condition. Therefore, it is considered a significant responsibility to cure plant diseases early on.

Numerous diseases cause damage to plant leaves, which affects crop productivity. Finding leaf diseases is crucial. Maintaining plant leaves regularly is what makes agricultural products profitable. Due to their ignorance of leaf diseases, farmers produce insufficient yields. The diagnosis of leaf diseases is crucial since productivity determines profit and loss. Thus, in this case, deep learning techniques are used to identify diseases in the leaves of tomato, potato, grape, apple, and maize plants [1]. Twenty-four thousand photos of leaves are used, together with twenty-four different types of leaf diseases [2].

As such, having a system that can automatically identify indications of leaves disease as soon as feasible is crucial. Improved efficacy and efficiency in leaf disease diagnosis are now achievable because to recent developments in ML. Apple growers would avoid significant loss or harm if a ML-based approach proved effective in identifying problems in their fruit early [3].

Plant leaf disease diagnosis is frequently costly, time-consuming, prone to error, and requires specialized knowledge. In order to distinguish between diseases, recent methods rely on image-based analysis, in which algorithms identify patterns and anomalies in leaf photos. Because many plant leaf diseases have visually similar symptoms that make manual identification challenging, this line of research is crucial. For accurate disease classification, a combination of computer vision, ML, and image processing techniques is therefore usually needed for an effective diagnosis [4].

DL has achieved major progress in recent years, especially in image classification tasks for rapid and accurate plant disease identification [5]. However, a number of issues still need to be addressed in the classification of plant diseases, even with the advancement of ViT-based models and Transfer Learning (TL) [6]. Large, balanced datasets are still necessary for standard ViT architectures, and they have trouble capturing fine-grained local lesion patterns. Furthermore, when multiple diseases have similar visual symptoms, traditional TL-based classifiers provide limited inter-class separability and are unable to generalize to unseen disease categories [7]. These problems point to a glaring research gap in the creation of a model that combines enhanced discriminative representation, robust local–global feature learning, and zero-shot capability for the identification of unseen diseases. The contributions are:

- Proposed a novel Enhanced Hybrid Vision Transformer with Zero-Shot Classification (EHVZSC) framework for accurate plant leaf disease identification using attention-based deep feature learning.
- Integrated Vision Transformer (ViT) with Zero-Shot Learning to classify both seen and unseen plant disease classes without requiring additional labeled training samples.
- Improved disease detection performance on the Plant Village dataset by achieving 95.7 % accuracy, 97.8 % sensitivity, 95.0 % specificity, and 96.45 % F1-score with reduced data dependency.
- Demonstrated significant performance gains over existing methods, including up to 15 % higher accuracy, 12 % better F1-score, and 25 % improvement in ROC performance through robust attention-enhanced representation learning.

Although the suggested EHVZSC model performs well, there are a number of obstacles to its actual implementation in agricultural settings. The accuracy of classification may be impacted by noise, shifting illumination, complicated backgrounds, and obscured foliage in field photos. Detection is further complicated by early-stage disease symptoms, environmental factors, and different crop kinds. Furthermore, adoption of Vision Transformer-based models on inexpensive mobile or Internet of Things devices is difficult due to their high computational resource requirements. The efficiency of the system may be further diminished in remote farming areas by poor imaging quality, limited internet connectivity, and a lack of technical know-how. For dependable agricultural deployment, lightweight optimization and real-time field testing are crucial.

Figure 1 displays typical leaf infections across multiple plant species images from the PlantVillage dataset [Source: www.PlantVillage.org], including Apple Cedar Rust, Corn Common Rust, Pepper Bell Bacterial Spot, Potato Early Blight, Strawberry leaf disease, and Tomato Early Blight. These diseases can be accurately identified using DL-based classification techniques.

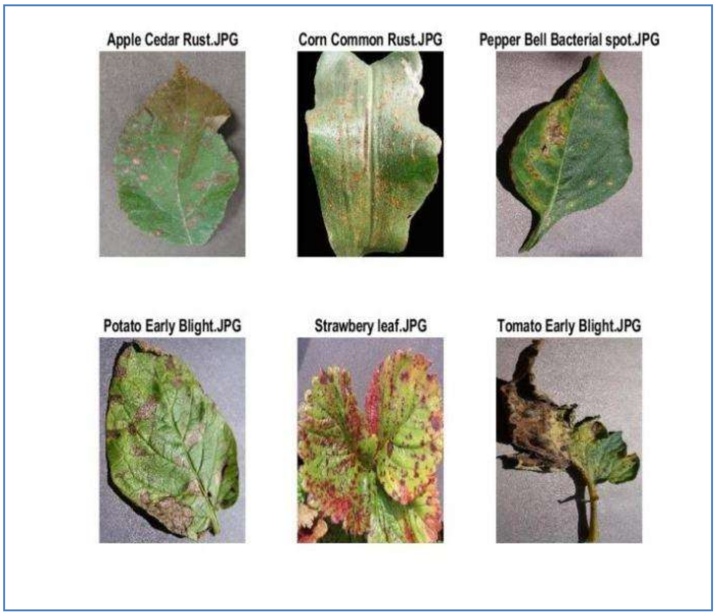

Figure 1. Typical leaf infections across multiple plant species.

The research addresses the issue of enhancing classifier performance when trained on source domain labeled data with diverse classes and tested on target domain data by introducing a unique framework named "Enhanced Hybrid Vision Transformer with Zero-shot learning classification (EHVZSC)". To investigate a several methods within this framework, such as data augmentation, segmentation and classification method. It also looks into the advantages of using robust discriminative losses to improve the performance of the classifier even more. The suggested EHVZSC framework introduces robust loss optimization, discriminative prototype-based zero-shot classification, and hybrid local–global feature modelling, in contrast to traditional ViT models improved through transfer learning. Together, these elements overcome the drawbacks of weak feature separability, poor unseen-class generalization, and high data dependency, allowing for noticeably better performance than conventional ViT + TL methods.

2. Related works

The detection of plant leaf diseases has been extensively researched to increase agricultural output and lower losses. The three primary categories of approaches found in the literature are classical ML, DL and advanced/transformer-based techniques.

2.1. Classical machine learning approaches

According to this study, contemporary organic farming is becoming more and more common in the agricultural sectors of many developing nations. Plant leaf disease is thought to be the most potent environmental element causing a deficiency in the quality of agricultural products, while there are many other environmental factors that also contribute to farming problems. The intention is to use machine learning and computer vision techniques to lessen this problem. The K-nearest neighbor (KNN) classifier was suggested by the authors as a method for the identification and classification of plant leaf diseases. For the classification, the texture elements are taken out of the photos of the leaf diseases [8].

2.2. Deep learning approaches

DL techniques for object detection, according to study, depend heavily on a large number of manually prelabeled training datasets as ground facts. Field item recognition, like bales, is particularly challenging due of three factors: (1) minimal prelabeled data available; (2) long-term image captures conducted in varying seasons and illumination conditions; and (3) a dearth of pretrained models and studies to utilize as a guide. This work builds an innovative pipeline of algorithms to boost the accuracy of bale recognition based on sparse data collecting and labeling. Furthermore, a type of transfer learning called domain adaptation (DA) is used to automatically label the training data and synthesize it under various environmental situations [9].

In this study investigated efficient supervised contrastive learning algorithms and tailored them to learn improved image demonstrations from imbalanced data. In order to realize the notion that better features lead to better classifiers; the authors specifically proposed a novel hybrid network structure that is made up of a cross-entropy loss for learning classifiers and a supervised contrastive loss for learning image representations. The learning process is gradually transitioned from feature learning to classifier learning. They investigated two variations of contrastive loss for feature learning, which differ in form but have the same basic principle of pushing samples from different classes apart and gathering samples in the normalized embedding space that are all from the same class [10].

In order to detect tomato diseases, presented a deep learning-based technique that creates artificial images of tomato plant leaves using the Conditional Generative Adversarial Network (C-GAN). After that, the tomato leaf photos are classified into ten disease categories using a DenseNet121 model that was trained using transfer learning on both artificial and real pictures [11].

According to this study discussed as the population has grown rapidly over time, which has raised the need for food production. Numerous bacterial, viral, and fungal plant diseases are among the main causes of reduced agricultural productivity. The majority of farmers apply fertilizers and insecticides to their entire field without explicitly considering the state of the plants [12]. Plant illnesses can be automatically detected by DL models from plant photos, which eliminate the need for human specialists. Suggested an EfficientNet DL architecture in the classification of plant leaf diseases. Transfer learning approach was used to train efficientNet architecture and other deep learning models. When transfer learning was being done, all the layers of the models were made trainable. EfficientNet is very computationally intensive, so it is not as practical to deploy it on low-power agriculture [13].

Introduced a powerful DL architecture of automatic plant disease detection and segmentation, DPD-DS, with an enhanced mask-region-based pixel-wise convolution neural network (CNN). In the proposed DPD-DS, a light head region convolution neural network (R-CNN) is suggested with the aim of saving the memory space and computational cost [14]. It improves detection accuracy and the speed of computations by changing the proportions of the Anchor in the RPN network and changing the backbone. DP-DS of instance segmentation is computationally expensive and is ineffective with small or overlapping areas of leaf diseases.

Achieving multi-classification of disease utilizing plant leaf image bearing in mind the chronological Flamingo search algorithm (CFSA) with transfer learning (TL). The pre-processing stage of leaf image is the Adaptive Anisotropic diffusion which removes noise [15]. In this case, the U-Net++ is used to segment plant leaf and trained using the Moving Gorilla Remora algorithm (MGRA). Flamingo search optimization framework presents excessive complexity in its algorithms and can easily overfit small plant disease data sets.

2.3. Transformer-based approaches

A study provided in the same label space; the goal of typical domain adaptation approaches is to transfer the information gained from a label-rich source domain to a target domain with fewer labels. Nevertheless, obtaining even the unlabeled target domain data for a task of interest is frequently challenging. Zero-shot domain adaptation (ZSDA) is the process by which the authors could transfer the domain shift between the source and target domains from an unseen task to the job of interest in such a scenario [16].

According to this study provided an example of pre-trained models are refined through the use of transfer learning. To combat overfitting, data augmentation techniques such as picture enhancement, rotation, scaling, and translation are used. The performance of several pre-trained neural networks was thoroughly categorized by the authors, who also included the results of a weighted ensemble of models pertinent to the identification of plant leaf diseases [17].

Transformer-based plant disease recognition system, in which the Vision Transformer architecture was improved to promote long-range interactions and the fine attributes of lesions in complex field images [18]. The model exhibited better robustness to illumination changes and background noises because of its self-attention mechanism, but fails in cases of training data being small or highly imbalanced.

The hybrid multispectral ViTCNN architecture of the plant disease detection by suggested a method that combined spectral data with deep spatial features to enhance the ability to discriminate similar diseases that appeared visually similar [19]. The fusion architecture was better at localizing features and decreasing misclassification in challenging outdoor scenarios, but multispectral acquisition makes the system more hardware-dependent and restricts its usage in low-resource agricultural settings.

PD-TR, an end-to-end transformer-based model that improves lesion localization as well as disease classification using a single self-attention encoder-decoder pipeline, was proposed by [20]. It enhanced the awareness of lesion boundaries and generated consistent predictions on different crop species; however, End-to-end transformer pipelines are still computationally expensive and cannot be deployed in real time on edge devices.

Zero-shot plant disease classification through the use of semantic attribute embeddings to classify unseen disease categories without the need to have labeled images [21]. This method greatly lowered dependency on annotation and generalized better to uncommon disease categories; however, Zero-shot performance is considerably reliant on the quality of semantic features, which may be noisy or not adequate to fine-grained diseases.

The transformer based few-shot learning model used in to detect barley disease was trained to acquire class prototypes using cross-attention mechanisms, only with a limited number of labeled data. The approach was successfully used to alleviate the lack of labeled agricultural data and achieve high recognition rates yet Few-shot transformer models are less stable and perform poorly when the intra-class variance is great [22].

Plant leaf disease detection techniques are summarized in **Table 1**, which demonstrates that while DL techniques like domain adaptation, contrastive learning, C-GAN-based augmentation, and lightweight CNNs improve accuracy but rely on large labelled datasets and frequently remain crop-specific, classical ML methods like KNN rely on handcrafted features and offer limited generalization. Although they lower data requirements, transfer learning and zero-shot domain adaptation have not been extensively investigated for multi-crop or multi-disease scenarios. Overall, the literature identifies important gaps: Transfer learning is underutilized, lightweight real-time models appropriate for field deployment are still underdeveloped, DL suffers from data imbalance and annotation demands, and classical ML lacks robustness.

Table 1. Comparison of plant leaf disease detection approaches.

Study	Approach	Key techniques	Strengths	Limitations
[8]	Classical ML	KNN, texture features	Simple, low computational cost	Handcrafted features, low robustness
[9]	DL	Domain Adaptation, sparse labeling	Handles sparse data, robust	Limited scalability
[10]	DL	Hybrid contrastive + cross-entropy loss	Improved feature learning	Computational complexity
[11]	DL	C-GAN + DenseNet121	Synthetic data generation improves classification	Requires GAN training
[12]	DL	Lightweight RCNN	Efficient, practical	Limited to specific crop
[13]	DL	EfficientNet	High accuracy, efficient scaling	Requires high-quality dataset
[14]	DL	Instance Segmentation (DPD-DS)	Precise localization + classification	High computational load
[15]	DL	Flamingo Search Optimization + Transfer Learning	Strong multi-class performance	Optimization complexity
[16]	Transfer Learning	Zero-Shot Domain Adaptation	Reduces labelled data requirement	Limited to certain domain shifts

Table 1. (*Continued*).

Study	Approach	Key techniques	Strengths	Limitations
[17]	Transfer Learning	Pretrained CNNs + data augmentation	Ensemble improves accuracy	Resource intensive
[18]	DL/Transformer	Vision Transformer	Robust to illumination and background variations	Requires large balanced datasets
[19]	DL/Hybrid ViT-CNN	Multispectral ViT–CNN	Improves local feature extraction, reduces misclassification	Requires multispectral hardware
[20]	DL/Transformer	PD-TR end-to-end transformer	Better lesion localization and cross-crop consistency	Computationally expensive, not real-time
[21]	Zero-Shot Learning	Semantic attribute embeddings	Recognizes unseen disease categories, reduces annotation	Performance depends on quality of semantic attributes
[22]	Few-Shot Learning/Transformer	Cross-attention prototype learning	Effective with few labelled samples, maintains high accuracy	Reduced stability with high intra-class variation

3. Proposed methodology

The PlantVillage dataset is used in the study methodology, which takes into account the sequence in which the experiments were conducted. Plant leaf diseases can be accurately predicted from plant photos using the proposed Enhanced Hybrid Vision Transformer with Zero-shot learning classification (EHVZSC) technique, yielding better results suitable for additional processing such as data augmentation [23]. **Figures 2** and **3** display the full suggested flow diagram for the EHVZSC training and testing process.

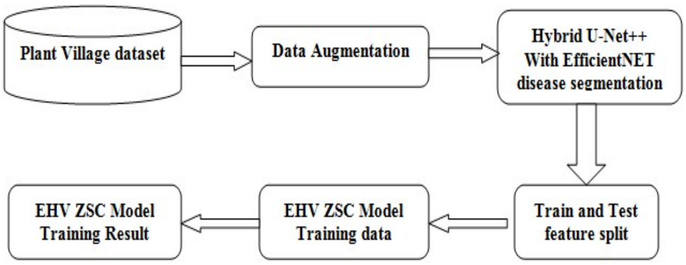

Figure 2. Proposed EHVZSC training process flow diagram.

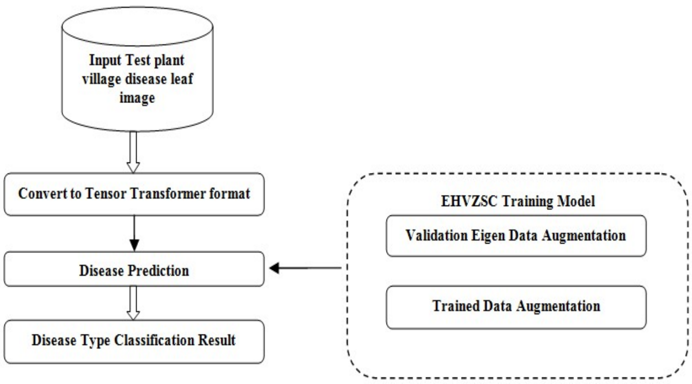

Figure 3. Proposed EHVZSC testing disease classification flow diagram.

3.1. Image acquisition

The PlantVillage dataset contains 54,305 images of healthy and diseased plant leaves across 38 classes. All images were preprocessed before training to ensure uniformity and improve feature extraction. First, each image was resized to 224 × 224 pixels and converted from RGB to HSV color space using Equation (1) to enhance leaf–disease contrast.

$$HSV = f(RGB) \tag{1}$$

Pixel intensities were then normalized to stabilize gradient updates using Equation (2):

$$I_{norm} = \frac{I - \mu}{\sigma} \tag{2}$$

where I is the input image, μ is the channel mean, and σ is the standard deviation. Basic augmentation (rotation, translation, contrast, hue, and saturation adjustments) was applied to strengthen generalization by Equation (3):

$$I' = T(I) \tag{3}$$

where $T(\cdot)$ represents the augmentation transformations. These preprocessing steps ensure consistent input quality and improve the robustness of the proposed EHVZSC model prior to segmentation and classification. **Figure 4** shows the sample plant disease images from the dataset [23].

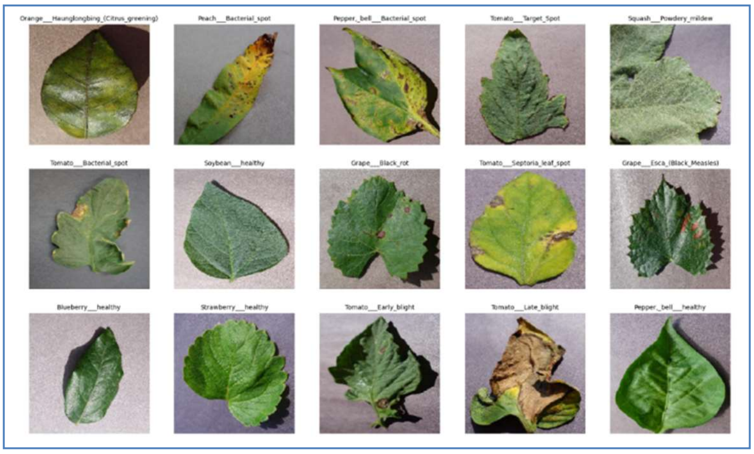

Figure 4. Plant disease sample from the plant Village dataset.

3.2. Data augmentation

The dataset was improved using seven different data-augmentation techniques, which are summarized in the **Table 2** along with their key parameters and effects on dataset variability and model robustness. While contrast, hue, and saturation adjustments enhance color-based variations, padding, rotation, translation, and affine transformations mainly increase geometric diversity. When combined, these augmentations increase the size of the effective dataset, lessen overfitting, and enhance the model's capacity to generalize under various lighting, orientation, and appearance conditions.

Table 2. Data augmentation techniques, parameters, and impact on dataset size.

Augmentation type	Key parameters	Impact on dataset size/model performance
Padding	Padding width p, padding mode (constant/reflect)	Slight increase in dataset variability; helps preserve aspect ratio.
Rotation	Angle range $\theta \in [-40\,°,\ +40\,°]$, rotation center	Significant dataset expansion; improves rotational invariance.
Translation	Shift in pixels (t_x, t_y)or % of image	Increases robustness to positional changes; moderate dataset growth.
Affine transformation	Scale, shear, rotation, translation matrix	High variability added; improves geometric distortion tolerance.
Contrast adjustment	Contrast factor α; brightness β	Enhances illumination diversity; minimal effect on dataset size but improves feature distinguishability.
Hue adjustment	Hue shift δH	Expands color domain; helps model generalize to different lighting conditions.
Saturation adjustment	Saturation scaling factor k	Increases color intensity variation; moderate improvement in color robustness.

3.2.1. Dataset split and hyperparameter settings

The dataset was split into three subsets—70 % for training, 15 % for validation, and 15 % for testing—to guarantee accurate training and assessment. This division preserves enough validation and test samples to adjust hyperparameters and evaluate generalization while offering a sizable training set for learning robust features. The model was trained with 50 epochs, a batch size of 32, and a learning rate of 0.0001 because these parameters provided stable convergence without overfitting. A dropout rate of 0.3 was used to prevent overfitting, and the Adam optimizer was chosen due to its capacity for adaptive learning. Patch sizes of 16 × 16 and embedding dimensions of 768 were utilized for the Vision Transformer component, in accordance with commonly used configurations for balanced computational complexity and performance. The final Data Augmentation results are described in **Figures 5–7**.

3.3. Disease segmentation

The segmentation operation is applied to separate the whole leaf area with the image background. To mined accurate features the model has to be able to capture finer details and address the differences in the sizes of the leaves, their shapes and features. In order to do so, a variant of U-Net++ architecture is modified to form the basis of plant-leaf segmentation. U-Net++ is an improved form of U-Net that is intended to be used in the accurate segmentation of images. It proposes better skip connections and multi-scale feature aggregation, and hence is applicable to isolating the plant leaf regions that may have structural differences [24].

The architecture has three main components: the encoder, decoder and dense cross layer connections, which assist in capturing leaf features in various spatial resolutions. U-Net++ will be used in the proposed hybrid framework with EfficientNet to enhance segmentation performance on the plant disease dataset, which consists of both healthy and diseased leaves. U-Net++ is the primary feature extractor, with the encoder doing the feature selection and the decoder doing the upsampling, which is also known as a transcoder. Convulsion and max-pooling are used repeatedly in the encoder path. Each two 3 × 3 convolutions are followed by a 2 × 2 max-pooling layer that reduces the spatial dimension and adds channel dimension.

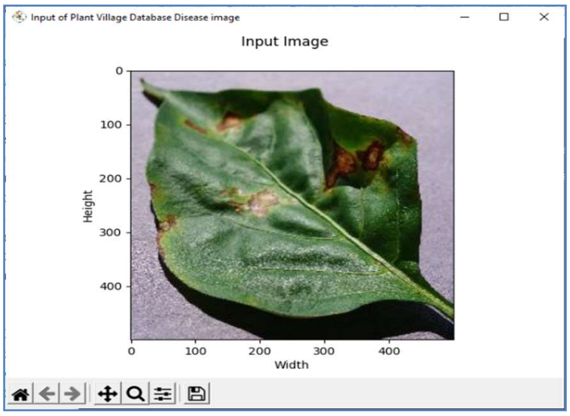

Figure 5. Input of pepper__bell___bacterial_spot image.

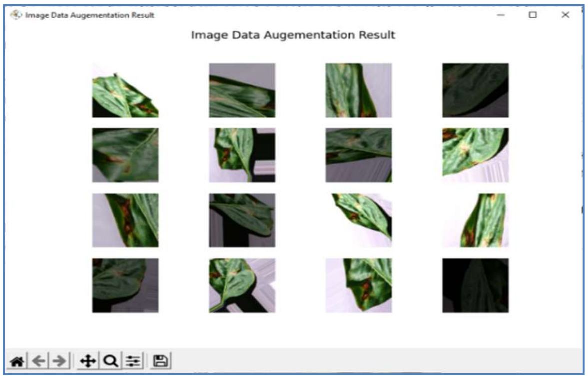

Figure 6. Data augmentation result-1.

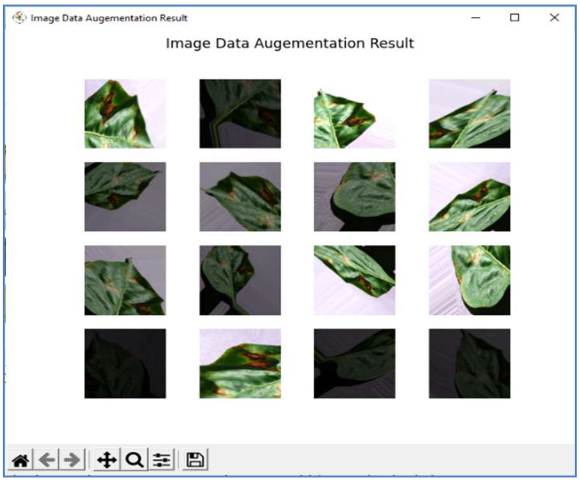

Figure 7. Data augmentation result-2.

The convolutions are all followed by a ReLU activation, and with such a network, it is possible to learn patterns in the down sampled feature maps. The network carries out upsampling in the decoder path through transposed convolution. The upsampled feature maps are refined after every transposed convolution with two 3 × 3 convolution layers and a ReLU activation. EfficientNet also improves the decoding step by providing powerful multi scale feature representations that increase the segmentation of diseased leaf areas. **Figure 8** shows the overall segmentation architecture.

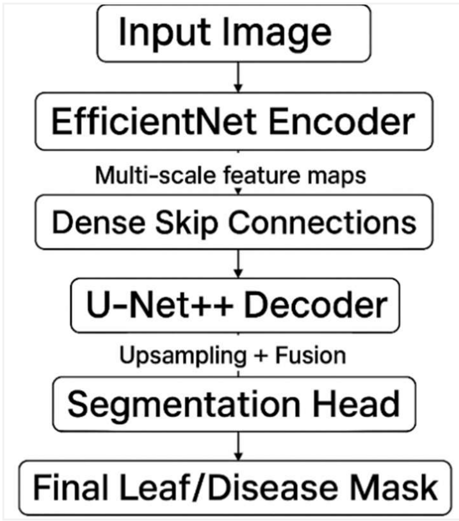

Figure 8. Segmentation architecture.

Mathematical implementations:

EfficientNetEncoder:

The encoder function in EfficientNet can be represented as Equation (4).

$$z = ENet\_encode(ip, L) \tag{4}$$

where, z is the output feature vector (encoding) of the input image; ip is the input image; L represents the learnable parameters of the encoder; *ENet_encodis* the encoder function, composed of the following components.

Convolutional Stem by Equation (5):

$$ip' = Conv(ip, Lconv) \tag{5}$$

where ip' is the stem output and $Lconv$ is the learnable parameter set of the convolutional stem.

EfficientNet Blocks (MBConv) by Equation (6):

$$ip'' = MBConv(ip', Lebconv) \tag{6}$$

where ip'' is the output of the *MBConv* block and $L_{(mbconv)}$ represents MBConvparameters.

Squeeze-and-Excitation (SE) Blocks by Equation (7):

$$ipSE = SE(ip'', Lse) \tag{7}$$

where $ipSE$ is the SE block output and $L_{(se)}$ is the SE module parameter set.

Global Average Pooling by Equation (8):

$$z = GAP(ipse) \tag{8}$$

Where, $GAP(\cdot)$ computes the global average pooling operation to generate final encoded features z.

Mathematically, the encoder function can be represented by Equation (9), which transforms an input image x into *a* feature vector z that captures its essential characteristics:

$$z = GAP(SE(MBConv(Conv(x, Lconv), Lmbconv), Lse), Lgap) \tag{9}$$

where *Lconv*, *Lmbconv*, *Lse*, and *Lgap* are learnable parameter sets for each component.

U-Net++ decoder:

The proposed model of U-Net++ decodingfeature maps are described as Equation (10),

$$feat^{a,b} = \begin{cases} L\left(feat^{a-1,b,}\right) & b = 0 \\ L\left(\left[\left[feat^{a,m}\right]_{m=0}^{b-1}, UP\left(feat^{a+1,b-1}\right)\right]\right) & b > 0 \end{cases} \tag{10}$$

where the concatenation layer is represented by [], the upsampling layer is indicated by *UP(.)*, and the convolution function is defined by *L(.)*. The input is received by the node at level $b = 0$ through the previous encoder layer. The node at level $b = 1$ receives the encoder and the sub-network input from two subsequent levels. The node at level $b > 1$ receives $b + 1$, where the ending input is the up-sampled result from the short skip pathway. The previous b nodes' outputs are referred to as inputs at level b.

Feature fusion and segmentation head:

Combine the features from the EfficientNet encoder with the upscaled features from the U-Net++ decoder using concatenation or attention mechanisms and finally add a segmentation head on top of the hybrid model to predict the final segmentation masks.

The proposed hybrid model of EfficientNET with U-Net++ feature fusion is described as Equation (11),

$$FF = Fusion\left(z, feat^{a,b}\right) \tag{11}$$

where *FF* is the fused feature map combining EfficientNet and U-Net++ outputs.

Add segmentation head on top of the hybrid model to predict the final segmentation masks by Equation (12).

$$seg = Seg\_head(FF) \tag{12}$$

where seg represents the predicted segmentation mask.

The hybrid model can be represented by Equation (13):

$$seg = Seg\_head(FF(z(x), Unet(z(x)))) \tag{13}$$

Where, *Unet(·)* denotes the U-Net++ decoder output.

The proposed Hybrid model segmentation results shown in **Figure 9**.

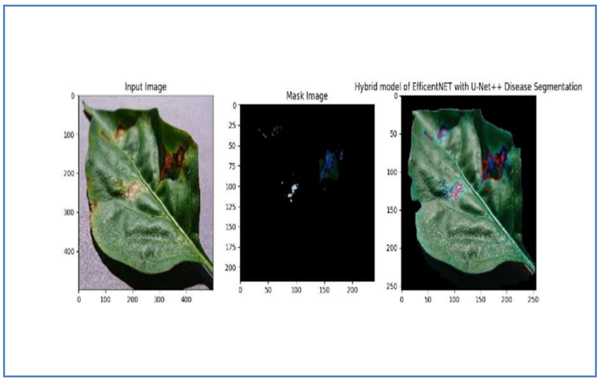

Figure 9. Proposed ECMDS segmentation result.

3.4. Enhanced hybrid vision transformer with zero-shot learning classification (EHVZSC)

A proposed hybrid vision transformer with zero-shot learning capabilities can be achieved by combining the Vision Transformer (ViT) architecture with zero-shot learning techniques for plant leaf disease prediction. In particular, the proposed vision transformer model only accepts input images with a resolution of 224 × 224. The chosen size of the suggested system was determined by the ease with which 224 × 224 could be split into 256 (16 × 16) image patches. Each picture patch is 14 by 14 pixels when there are 256 image patches. It displays the changes the image undergoes to become the 256 (14 × 14) pixel image patches.

The Vision Transformer (ViT) patch generation Equation (14):

$$x_p(i,j,k,l) = x(i \times P + k, j \times P + l, :) \tag{14}$$

Where, x_p represents the patch tensor - x is the input image tensor - i, j are the patch indices (0 to N_{h-1} and 0 to N_{w-1}, respectively) - k, l are the pixel indices within a patch (0 to P-1) - P is the patch size - Nh, Nw are the number of patches in the height and width dimensions, respectively.

The Vision Transformer (ViT) feature encoding process involves the following steps:

Patch embedding: Divide the input image into fixed-size patches and embed each patch into a vector using a linear layer.

Position encoding: Add positional encodings to the patch embeddings to retain spatial information.

Transformer encoder: Feed the patch embeddings and positional encodings into a transformer encoder, which consists of multiple identical layers. Each layer applies self-attention, feed-forward neural networks (FFNNs), and layer normalization.

Self-attention: Compute the attention weights between patches and perform weighted sum to capture patch relationships.

Feed-forward neural networks (FFNNs): Transform the patch representations using FFNNs.

Layer normalization: Normalize the patch representations to stabilize training.

Output: The final output is the encoded feature representation of the input image.

The Transformer encoder formation is described in Algorithm 1.

This mathematical model represents the Transformer encoder in ViT, which takes in a sequence of image patches and outputs a sequence of embeddings that capture the contextual relationships between the patches.

The classification method using zero-shot transform receives the output of the ViT Transform encoder as input. Zero-shot transform classification is a technique that enables a model to classify images into classes it has not seen before, without requiring any additional training data. In traditional classification, a model is trained on a dataset of labeled images, where each image is associated with a specific class label. However, in proposed zero-shot classification, the model is trained on a dataset of images from seen classes, but is then asked to classify images from unseen classes.

Algorithm 1 ViT Transformer encoder

```
Input:
  X – input image patch embeddings
  L – number of encoder layers
  h – number of attention heads
  d – feature dimension
d_head – dimension per head
Output:
  TC – transformer-encoded representation
Procedure:
1. Patch embedding:
  X ←Concat (Embed(patch1), …, Embed(patchN))
2. Encoder layers:
  for ℓ = 1 ... L do
    # Multi-Head Self-Attention
    Q ← X · W_Q
    K ← X · W_K
    V ← X · W_V
    for i = 1 ... h do
head_i ← Attention (Q·W_Q_i, K·W_K_i, V·W_V_i)
    end for;
    MSA ←Concat (head1 … head_h) · W_O
    # Add & LayerNorm
    X ← LayerNorm (X + MSA)
    # Feed-Forward Network
    FFN ← (ReLU (X·W1 + b1) ) ·W2 + b2
    # Add & LayerNorm
    X ←LayerNorm (X + FFN)
  end for;
3. Output:
  TC ← X
LayerNorm definition:
LayerNorm (x):
μ← Mean (x)
σ← Std (x)
return γ * ((x − μ)/σ) + β
```

The mathematical equation for zero-shot transform classification can be represented as by Equation (15).

$$Z = dot\_product(X, prototype) + bias \tag{15}$$

Where, Z represents output tensor representing the zero-shot classification scores; X is the input tensor representing the transformed image features; prototype is a learnable tensor representing the class prototypes; bias is a learnable tensor representing the bias terms; dot_product is the dot product operation.

The prototype tensor is made up of learnable class-specific vectors that are optimized during training, and semantic vectors that externally define the meaning of classes (e.g. attribute, word embeddings), do not change during the learning process. Prototypes are a feature space trainable anchor, which is trained to match the visual representation of the ViT encoder. Conversely, class embeddings are fixed and give semantic structure, which allows the model to make unseen classes related to each other on shared semantic or linguistic relationships.

The rewrite an Equation (15) with ViT Transform encoder like in Equation (16),

$$Y = softmax(dot\_product(TC, prototype) + bias \quad (16)$$

Where, TC is the Transformer encoder; SoftMax is the SoftMax activation function.

The above Equation (16) combines the patch embedding, transformer encoder, and zero-shot classification components into a single mathematical expression. This equation represents the entire hybrid vision transformer model with zero-shot transform classification, from input image to output classification. The proposed final EHVZSC Classification results are illustrated in **Figure 10**.

Figure 10. EHVZSC classification result.

4. Results and discussion

Hyperparameter tuning plays an important role in improving the performance and stability of the proposed EHVZSC model. Parameters such as learning rate, batch size, number of epochs, optimizer settings, and transformer depth were carefully adjusted to achieve optimal classification accuracy while avoiding overfitting. Different combinations of hyperparameters were tested experimentally, and the model performance was evaluated using validation accuracy and loss convergence. A learning rate of 0.0001, batch size of 32, and 50 training epochs were selected because they provided stable convergence, faster learning, and better generalization capability. Proper tuning also helped enhance sensitivity, specificity, F1-score, and overall robustness of the Vision Transformer with zero-shot learning framework

Using the PlantVillage dataset, the proposed EHVZSC model is compared with several recent state-of-the-art plant disease classification methods, including EfficientNet-based deep learning, DPD-DS, DbneAlexNet-MGRA& CFSA + TL with CNN + LeNet. The evaluation considers key performance parameters such as confusion matrix, training and validation accuracy, loss curves, and testing metrics including accuracy, specificity, sensitivity, F1-score and ROC curve. The experimental validation of truly unseen classes is still limited, despite the fact that the proposed EHVZSC system has zero-shot learning capability for recognizing unseen illness categories. The PlantVillage dataset, which has relatively limited variances in real-world unknown disease situations, is the main source of information used in this review. To fully confirm the robustness and generalization capability of the zero-shot classification process under realistic agricultural scenarios, more studies involving whole new disease categories, cross-species infections, and a variety of field datasets are required.

The confusion matrix in **Figure 11** illustrates how effectively the model differentiates diseased from healthy leaves: out of all truly diseased samples, the model correctly identifies 489 as diseased (true positives) while misclassifying 11 as healthy (false negatives). Similarly, among the truly healthy samples, the model correctly labels 475 as healthy (true negatives) but wrongly predicts 25 as diseased (false positives). Since the majority of samples lie along the main diagonal (489 and 475), the classifier demonstrates strong accuracy in correctly recognizing both diseased and healthy leaves.

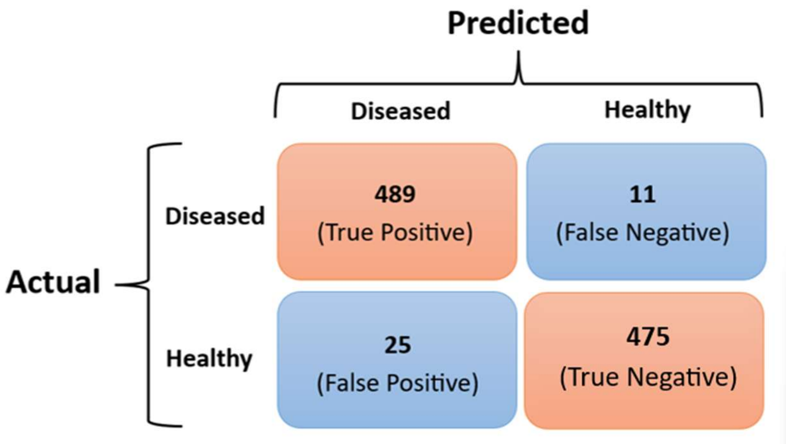

Figure 11. Confusion matrix.

Figure 12 shows the training and validation accuracy curve with 50 epochs. The training accuracy reaches above 95 percent in the early epochs and it persistently stays at the same high level, with no significant fluctuation in the convergence of the model. It is allowed to have a rapid improvement in the validation (test) accuracy in the initial epochs and then oscillates around 80 percent to 90 percent as it is observed to strongly generalize with slight fluctuations as a result of complexity of the classes and patterns that are not seen. All in all, the number shows that the EHVZSC model is highly accurate and it remains reliable in the course of training.

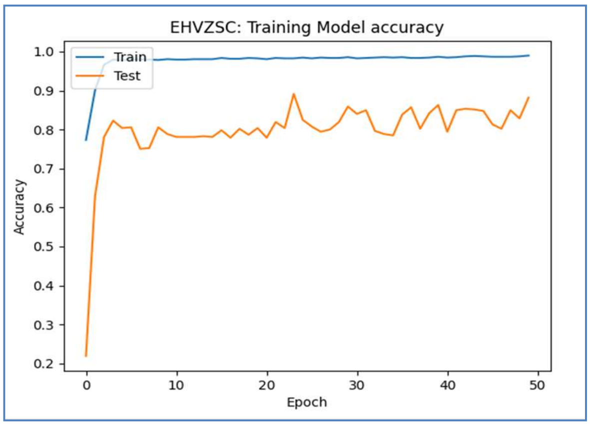

Figure 12. Training and validation accuracy.

Figure 13 shows the trend of training and validation loss with 50 epochs. The loss on training diminishes slowly and levels off at a low level and this signifies that the model is learning steadily and converging. Conversely, the loss in validation decreases rapidly during the initial epochs and then oscillates yet followed by an overall downward trend indicating that the model still has good generalization ability. The fact that there is no substantial variation between the curves indicates that overfitting is inhibited and the EHVZSC model acquires robust representations in training.

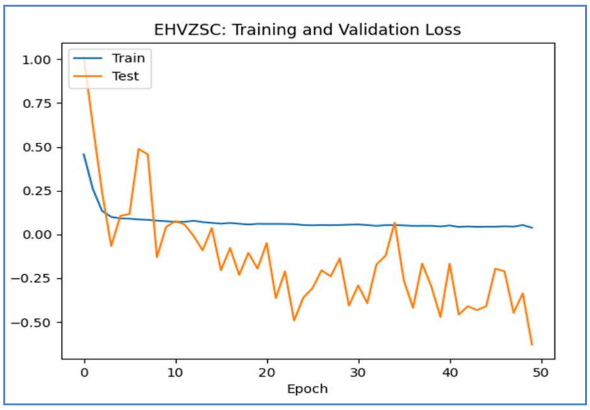

Figure 13. Training and validation loss.

The comparison of the accuracy in testing various plant-disease classification methods, including EfficientNet-based deep learning, DPD-DS, DbneAlexNet-MGRA, CFSA+TL with CNN + LeNet, and the proposed EHVZSC model, is shown in **Figure 14**, which was tested with 60 to 90 training epochs. There is a gradual increase in accuracy of all methods with increased number of epochs, but EHVZSC is always found to have the greatest accuracy at each stage with a sharp increase of 0.871 to 0.959 at 60 and 90 epochs respectively. In general, the accuracy of EHVZSC is 2–15% higher than the rest of the techniques and this is attributed to the fact that the model is better at capturing both global and fine-grained disease cues due to its architecture strength, such as attention-oriented feature refinement and zero-shot flexibility. In comparison to the traditional CNN-based or even hybrid models, where the localized features play an important role, EHVZSC hierarchical attention mechanisms enhance discriminative learning, whereas its zero-shot ability enhance the output to unseen patterns, leading to faster convergence, lower misclassification, and better robustness at each stage of training.

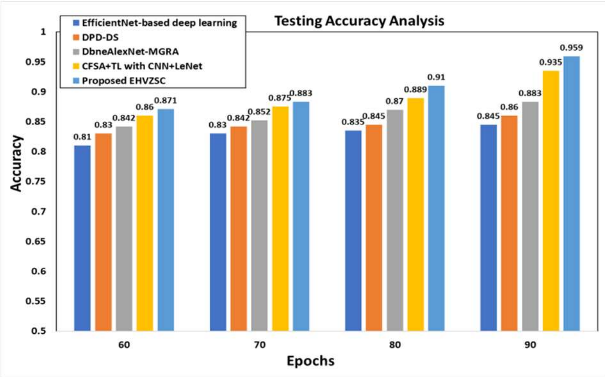

Figure 14. Testing accuracy analysis.

Figure 15 that displays the sensitivity analysis between 60 to 90 epochs indicates that all the compared approaches, such as EfficientNet-based deep learning, DPD-DS, DbneAlexNet-MGRA, and CFSA + TL with CNN + LeNet, are improving over time with training, but the proposed EHVZSC model is always most sensitive at any given time, which is 0.896 at epochs 60, and 0.978 at epochs 90. EHVZSC has a higher sensitivity of about 2–10 per cent than the other methods, proving that it has a greater capability to detect diseased samples with correct accuracy. This improvement in performance can be explained by the architectural benefits of the model, in particular, the ability to discriminate insidious aspects of disease and generalization to more complex or unobservable changes in performance, which has the characteristics of attention-enhanced processing and zero-shot adaptability. EHVZSC reduces the false negative rate and achieves high robustness per training epoch, compared to traditional CNN-based models that make use of few local feature images.

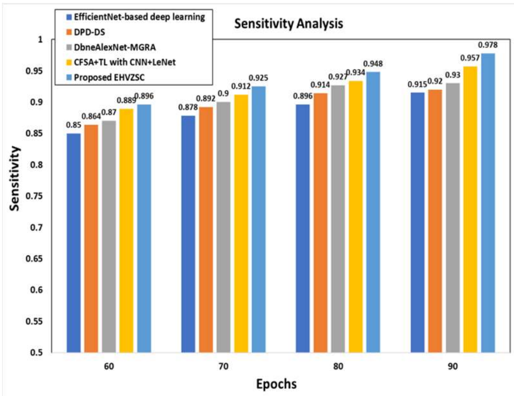

Figure 15. Sensitivity analysis.

Figure 16 shows the specificity comparison of the five plant-disease classification methods among 60 to 90 epochs where all the methods show a gradual improvement with epochs; the proposed EHVZSC model, however, shows the highest specificity rate of 0.88 at 60 epochs to 0.95 at 90 epochs. On the whole, EHVZSC works better than the rest of methods by a margin of 2–10 percent, showing its superiority in detecting healthy samples with the minimal number of false positives. This is due to its better architecture which incorporates attention-based feature refinement and zero-shot flexibility so that the model can capture finer differences between healthy and diseased regions. In contrast to classical CNN-based methods that use localized patterns as the primary ones, EHVZSC has an advantage of the global contextual knowledge and hierarchical acquisition of features, which leads to higher discrimination power and better generalization abilities of varying training epochs.

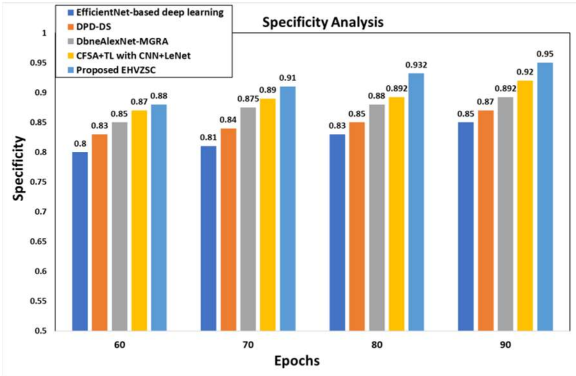

Figure 16. Specificity analysis.

The suggested EHVZSC model continuously gets the best performance at every step, increasing from 0.895 at 60 epochs to 0.968 at 90 epochs, according to **Figure 17**, which compares the F1-score of five plant-disease classification methods over 60 to 90 epochs. EHVZSC shows an overall improvement of 2–12% when compared to current approaches, demonstrating its improved recall and accuracy balance. The main cause of this steady improvement is EHVZSC's sophisticated attention-driven feature learning, which makes it possible for the model to identify both global and fine-grained disease patterns more successfully than conventional CNN-based methods. This lowers false positives and false negatives and produces a higher F1-score throughout all training epochs.

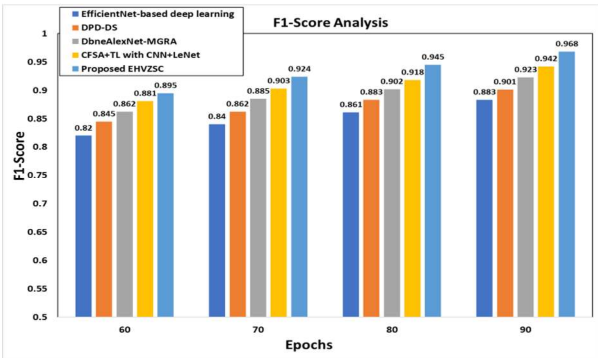

Figure 17. F1-Score analysis.

For five plant-disease classification models, **Figure 18** compares the ROC curve with the True Positive Rate (TPR) progression across increasing False Positive Rates (FPR). The suggested EHVZSC consistently gets the maximum TPR at every threshold. Even when false positives are kept to a minimum, EHVZSC beats the other approaches by around 5–25%, particularly at low to mid FPR values (0.02–0.40). EHVZSC's attention-driven and context-aware feature extraction, which allows the model to capture fine-grained disease patterns more successfully than traditional CNN-based techniques, is primarily responsible for this strong ROC performance. This leads to more confident and discriminative predictions across all decision thresholds.

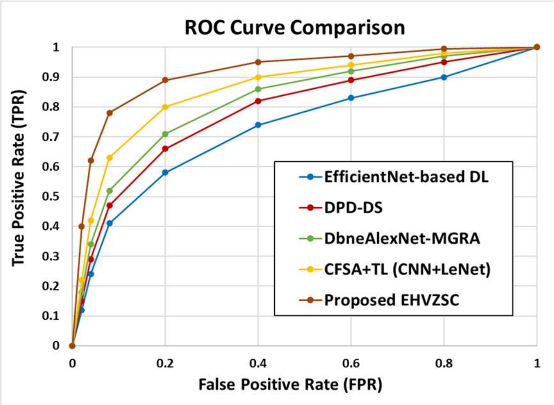

Figure 18. ROC curve comparison.

The Performance Comparison of EHVZSC with State-of-the-Art approaches in **Table 3** indicates that, at 90 training epochs, the suggested EHVZSC model consistently outperforms all baseline approaches in key parameters, including accuracy, sensitivity, specificity, and F1-score. With 95.9 % accuracy, 97.8 % sensitivity, 95.0 % specificity, and 96.8 % F1-score, EHVZSC reduces false positives in healthy leaves while simultaneously more accurately identifying sick leaves. Despite its great accuracy, the suggested EHVZSC model has many drawbacks. The PlantVillage dataset, which includes controlled photos and might not accurately represent real-field agricultural situations, is primarily used to evaluate the model. Real-time performance can be impacted by variations including dim illumination, complicated backgrounds, occlusions, and mixed infections. Furthermore, the Vision Transformer architecture is difficult to implement on inexpensive edge devices because to its high computing resource requirements and lengthy training period. Further validation using a variety of real-world datasets is also required for the zero-shot learning capabilities for undiscovered diseases.

Table 3. Performance comparison of EHVZSC with state-of-the-art methods.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
EfficientNet-based DL [13]	84.5	91.5	85.0	88.3
DPD-DS [14]	86.0	92.0	87.0	90.1
DbneAlexNet-MGRA [15]	88.3	93.0	89.2	92.3
CFSA + TL with CNN + LeNet [15]	93.5	95.7	92.0	94.2
Proposed EHVZSC	95.9	97.8	95.0	96.8

5. Conclusion

This work presents the development of EHVZSC, an Enhanced Hybrid Vision Transformer with Zero-Shot Learning, for accurate plant leaf disease classification, particularly addressing unseen disease classes. The model leverages Vision Transformers to learn robust attention-enhanced image representations, which are then used to generate prototypes for zero-shot classification. Evaluated on the PlantVillage dataset, EHVZSC achieves 95.7 % accuracy, 97.8 % sensitivity, 95.0 % specificity, and 96.45 % F1-score, with improved testing performance over existing methods in ROC, sensitivity, and specificity, demonstrating its capability to capture fine-grained disease features. Future work includes large-scale field validation to test robustness under real-world conditions such as variable lighting, occlusions, and mixed infections, and cross-domain generalization across different plant species, regions, and imaging conditions to enhance applicability. Integration with mobile and IoT-based agricultural systems could enable real-time disease monitoring and support timely crop management interventions, further improving efficiency and yield. By creating lightweight and energy-efficient EHVZSC models that can operate on smartphones, drones, and IoT-enabled edge devices, future research will concentrate on Edge AI deployment and mobile-based disease diagnosis for real-time, offline plant disease detection and precision agriculture support in remote farming environments.

Author contributions: Conceptualization, PE and AMJMDZR; methodology, PE; software, PE; validation, PE and AMJMDZR; formal analysis, PE; investigation, AMJMDZR; resources, AMJMDZR; data curation, AMJMZR; writing—original draft preparation, PE; writing—review and editing, PE; visualization, PE; supervision, AMJMDZR; project administration, AMJMDZR; funding acquisition, PE. All authors have read and agreed to the published version of the manuscript.

Funding: None.

Ethical approval: Not applicable.

Informed consent statement: Not applicable.

Acknowledgement: The authors appreciate the effort of the editors and reviewers.

Conflict of interest: The authors declare no conflict of interest.

References

1. J AP, Gopal G. Data for: Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network. 2019. doi: 10.17632/tywbtsjrjv.1
2. Patel S, Jaliya UK, Patel P. A survey on plant leaf disease detection. International Journal for Modern Trends in Science and Technology. 2020; 6: 129–134. doi: 10.46501/IJMTST
3. Liakos KG, Busato P, Moshou D, et al. Machine learning in agriculture: A Review. Sensors, 2018; 18(8): 2674. doi: 10.3390/s18082674
4. Shoaib M, Shah B, Ei-Sappagh S, et al. An advanced deep learning models-based plant disease detection: A review of recent research. Frontiers in Plant Science. 2023; 14: 1158933. doi: 10.3389/fpls.2023.1158933
5. Thakur PS, Chaturvedi S, Khanna P, et al. Vision transformer meets convolutional neural network for plant disease classification. Ecological Informatics. 2023; 77: 102245. doi: 10.1016/j.ecoinf.2023.102245
6. Tharani Pavithra P, Baranidharan B. A hybrid ViT-CNN model premeditated for rice leaf disease identification. International

Journal of Computational Methods and Experimental Measurements. 2024; 12: 35–43. doi: 10.18280/ijcmem.120104
7. Chen Z, Wang G, Lv T, et al. Using a hybrid convolutional neural network with a transformer model for tomato leaf disease detection. Agronomy. 2024; 14(4): 673. doi: 10.3390/agronomy14040673
8. Hossain E, Hossain MF, Rahaman MA. A color and texture based approach for the detection and classification of plant leaf disease using KNN classifier. 2019 international conference on electrical, computer and communication engineering (ECCE). IEEE; 2019. pp. 1–6. doi: 10.1109/ECACE.2019.8679247
9. Zhao W, Yamada W, Li T, et al. Augmenting crop detection for precision agriculture with deep visual transfer learning—a case study of bale detection. Remote Sensing. 2020; 13(1): 23. doi: 10.3390/rs13010023
10. Wang P, Han K, Wei XS, et al. Contrastive learning based hybrid networks for long-tailed image classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. pp. 943–952.
11. Abbas A, Jain S, Gour M, Vankudothu S. Tomato plant disease detection using transfer learning with C-GAN synthetic images. Computers and Electronics in Agriculture. 2021; 187: 106279. doi: 10.1016/j.compag.2021.106279
12. Jhoo WY, Heo JP. Collaborative learning with disentangled features for zero-shot domain adaptation. Proceedings of the IEEE/CVF international conference on computer vision. 2021. pp. 8896–8905.
13. Atila Ü, Uçar M, Akyol K, Uçar E. Plant leaf disease classification using EfficientNet deep learning model. Ecological Informatics. 2021; 61: 101182. doi: 10.1016/j.ecoinf.2020.101182
14. Kavitha Lakshmi, R, Savarimuthu N. DPD-DS for plant disease detection based on instance segmentation. Journal of Ambient Intelligence and Humanized Computing. 2023; 14(4): 3145–3155. doi: 10.1007/s12652-021-03440-1
15. Chilakalapudi M, Jayachandran S. Multi-classification of disease induced in plant leaf using chronological Flamingo search optimization with transfer learning. PeerJ Computer Science. 2024; 10: e1972. doi: 10.7717/peerj-cs.1972
16. Vallabhajosyula S, Sistla V, Kolli VKK. Transfer learning-based deep ensemble neural network for plant leaf disease detection. Journal of Plant Diseases and Protection. 2022; 129(3); 545–558. doi: 10.1007/s41348-021-00465-8
17. Rehana H, Ibrahim M, Ali MH. Plant disease detection using region-based convolutional neural network. ArXiv Preprint ArXiv:2303.09063. 2023. doi: 10.48550/arXiv.2303.09063
18. Parez S, Dilshad N, Alghamdi NS, et al. Visual intelligence in precision agriculture: Exploring plant disease detection via efficient vision transformers. Sensors. 2023; 23(15): 6949. doi: 10.3390/s23156949
19. De Silva M, Brown D. Multispectral plant disease detection with vision transformer–convolutional neural network hybrid approaches. Sensors. 2023; 23(20): 8531. doi: 10.3390/s23208531
20. Wang H, Nguyen TH, Nguyen TN, Dang M. PD-TR: End-to-end plant diseases detection using a transformer. Computers and Electronics in Agriculture. 2024; 224: 109123. doi: 10.1016/j.compag.2024.109123
21. Kumar P, Mathew J, Sanodiya RK, et al. Zero shot plant disease classification with semantic attributes. Artificial Intelligence Review. 2024; 57(11): 305. doi: 10.1007/s10462-024-10950-9
22. Rezaei M, Diepeveen D, Laga H, et al. A transformer-based few-shot learning pipeline for barley disease detection from field-collected imagery. Computers and Electronics in Agriculture. 2025; 229: 109751. doi: 10.1016/j.compag.2024.109751
23. PlantVillage. Available at: https://plantvillage.psu.edu/ (accessed on 29 September 2022).
24. Fan X, Zhou R, Tjahjadi T, et al. A segmentation-guided deep learning framework for leaf counting. Frontiers in Plant Science. 2022; 13: 844522. doi: 10.3389/fpls.2022.844522