Received: October 24, 2023. Revised: November 30, 2023.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Tomato Disease Detection Using Vision Transformer with Residual L1-Norm Attention and Deep Neural Networks

Mamta Tiwari[1] Hemant Kumar[2]* Navin Prakash[3] Sunil Kumar[4] Rahul Neware[5] Shivneet Tripathi[1] Ritesh Agarwal[1]

[1]*Department of Computer Application, School of Engineering and Technology, Chhatrapati Shahu Ji Maharaj University, Kanpur, India*
[2]*Department of Information Technology, School of Engineering and Technology, Chhatrapati Shahu Ji Maharaj University, Kanpur, India*
[3]*Department of Computer Science & Engineering (AI), IIMT College of Engineering, Greater Noida, India*
[4]*Department of Computer Application, Allenhouse Business School, Kanpur, India*
[5]*National Environmental Engineering Research Institute (CSIR-NEERI), Nagpur, India*
* Corresponding author's Email: hemantime@gmail.com

Abstract: The tomato holds significant global significance as a crop, fulfilling crucial roles in agriculture and meeting essential human dietary requirements. The persistent presence of diseases poses a significant risk to tomato crops, compromising their productivity and economic viability. This study presents a novel approach that combines the vision transformer (ViT) with deep neural networks (DNN) to identify tomato leaf diseases. This research aims to develop a methodology for identifying tomato leaf diseases, specifically focusing on enhancing the model's performance and interpretability. The proposed technique integrates an improved multi-head self attention mechanism with L1-norm attention, enhancing its ability to capture essential image features. The capability of ViT to process images as sequences of patches is harnessed and integrated with the interpretative and classifying prowess of DNN. To train and test the model, we utilize a comprehensive dataset of 18,160 tomato leaf images classified into ten different disease types. According to the experimental findings, the proposed model achieved an exceptional accuracy rate of 99.74%. The proposed method surpasses other contemporary techniques in terms of both accuracy and precision (99.66%), recall (99.39%), and F1-score (99.53%) for disease classification. The model's exceptional performance highlights its potential for practical implementation in diagnosing tomato diseases. In conclusion, this study presents a novel and efficient method for detecting tomato leaf diseases, which fulfills the current agricultural need for scalable and precise techniques. The fusion of vision transformers and deep neural networks offers a robust solution with exceptional accuracy and interpretability, contributing to the sustainable cultivation of this crucial crop.

Keywords: Vision transformer, Tomato leaf disease, Multi-head self attention with L1-norm, Deep neural networks.

1. Introduction

The tomato (solanum lycopersicum) is a highly consumed and economically significant crop worldwide. It is crucial in global agriculture and human nutrition [1]. Tomatoes are widely consumed globally due to their versatility in cooking and high nutritional value. The sustainable cultivation of this crop is consistently hindered by the presence of diseases, which pose a significant risk to yield, quality, and economic sustainability. Tomatoes play a crucial role as a staple food and serve as a fundamental component of agricultural economies. They play a significant role in global agricultural production and trade, affecting the livelihoods of millions of people. Ensuring the health and productivity of tomato plants is of utmost importance, as they play a crucial role in ensuring food security and economic stability [2].

The importance of disease detection in tomato cultivation is of great significance. Tomato plants are vulnerable to diseases caused by various

pathogens, such as fungi, bacteria, and viruses [3]. If not properly controlled, these diseases can cause significant damage to crops, which in turn can jeopardize food availability and increase production costs. The prompt highlights the importance of promptly and accurately identifying tomato leaf diseases. This is crucial for effectively managing diseases, preserving crop yield, and promoting sustainable agricultural practices [4].

Numerous studies have been investigated by researchers for the identification of tomato leaf diseases. These methodologies encompass many approaches, starting from primary convolutional neural network (CNN) methods and progressing towards more sophisticated techniques. Sanida et al. [5] proposed a novel model that builds upon the VGGNet architecture, a well-known convolutional neural network (CNN) architecture. The authors showcased the application of proven approaches within the field to develop their model. In addition to the work above, Paul et al .[6] significantly improved the process's optimization. They proposed a convolutional neural network (CNN) model and integrated transfer learning techniques using VGG-16 and VGG-19 models. By employing data augmentation techniques, they were able to get enhanced outcomes. In recent studies, researchers, exemplified by LightMixer [7], have employed techniques to achieve notable advancements. These methodologies use concise models with limited parameters, prioritizing efficient feature fusion. As a result, these approaches have demonstrated remarkable accuracy rates in identifying tomato leaf diseases. Shoaib et al. [8] demonstrated the application of residual networks (ResNet) to improve bacterial detection and plant disease diagnosis. This research highlights using several convolutional neural network (CNN) designs for classification purposes. Bhandari et al. [9] utilized the EfficientNetB5 model innovatively to visually represent different illnesses caused by viruses in tomato leaves. Their study highlights the significance of model interpretability by employing approaches such as gradient-weighted class activation mapping (GradCAM) and model-agnostic explanations. Chen et al. [10] introduce LBFNet, a lightweight convolutional neural network, as the base model for tomato leaf disease recognition. The model utilizes a three-channel attention mechanism module to learn disease features and effectively minimize interference from redundant features. Introducing a cascade module aims to enhance model depth and address gradient descent issues.

Hybrid methods that combine vision transformers with CNNs are vital for enhancing performance. Thakur et al. [11] proposed a novel hybrid model combining vision transformers and CNNs. This model is specifically designed for lightweight and IoT-based agricultural systems, and it demonstrates impressive accuracy while prioritizing model explainability. Hossain et al. [12] investigated transformer-based models, including EANet, MaxViT, CCT, and PVT, to diagnose tomato leaf diseases. Their findings revealed that MaxViT exhibited superior performance to the other models, thus highlighting the promising capabilities of vision transformer-CNN hybrids within this field. Tabbakh and Barpanda [13] introduced TLMViT, a hybrid model that integrates transfer learning-based models with vision transformers. The TLMViT framework comprises four stages: data acquisition, image augmentation, leaf feature extraction using VGG19, and classification using the vision transformer (ViT) model. Gole et al.[14] introduced the trans-inception block as a modified encoder architecture in their study. This block replaces the MLP block found in the ViT model with a custom inception block. The purpose of this modification is to address the computational inefficiency of the MLP module in existing ViTs. Wang et al. [15] employ a hybrid approach, integrating global vision transformer (ViT) and local convolutional neural network (CNN) architectures to diagnose disease images in tomato plants. The ViT method is utilized for image recognition, and the CNN network is employed for feature extraction from plant images. Current crop disease classifiers encounter challenges in accurately identifying similar disease categories.

The aforementioned studies demonstrate the progression of methodologies for detecting tomato leaf diseases. This progression begins with primary convolutional neural network (CNN) approaches, progresses to more complex strategies, and vision transformer and culminates in innovative hybrid solutions that combine vision transformers with CNNs to improve accuracy and interpretability.

Tomato cultivation is crucial to worldwide agriculture, but illnesses may reduce productivity and quality. Traditional techniques are time-consuming and subjective; thus, automated and accurate illness-detection solutions are essential. This study intends to automate tomato disease detection and categorization using current deep learning advances, notably vision transformer models for image classification. This setting requires addressing illness symptom variability, real-time or near-real-time processing, imaging robustness, and scalability for substantial agricultural fields. This study intends to help farmers and agronomists identify tomato illnesses

early and correctly.

The following list summarizes the contributions of the proposed system:

- ***Improved multi-head self attention based on L1-norm*:** Introduce a multi-head self-attention technique that incorporates L1-norm attention and finds the most attentive patches for each attention head to better capture critical characteristics and interrelationships in image data, improving disease diagnosis.
- ***Hybrid improved vision transformer and deep neural networks frameworks:*** Proposes a novel hybrid approach that combines the enhanced vision transformer (ViT) with deep neural networks to identify tomato leaf diseases accurately. The approach utilizes feature extraction to provide a comprehensive and robust solution. This hybrid architecture integrates the advantages of ViT in processing images as patch sequences with the interpretability and classification capabilities of deep neural networks, leading to enhanced disease classification performance.

The subsequent sections of this paper are organized as follows: Section 2 proposes the improved vision transformer techniques incorporating L1-norm and deep neural networks for tomato leaf disease detection. Section 3 describes the experimental setups, technical details, and model evaluation metrics. Section 4 presents the results of the proposed approaches and compares them with state-of-art techniques—finally, section 5 concludes the study.

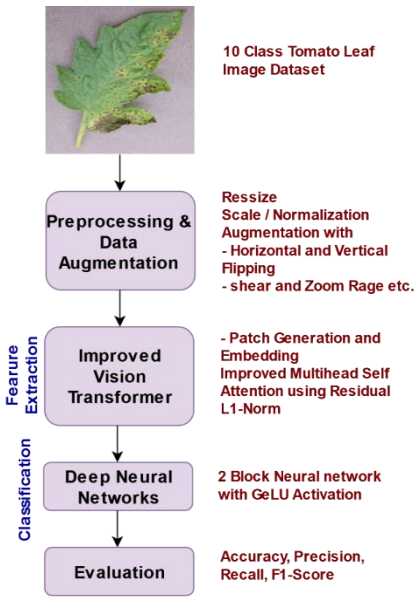


Figure. 1 Proposed model

Table 1. Tomato leaf dataset description

	Disease Name	Total	Train	Test	Valid
C0	Bacterial Spot	2127	1489	425	213
C1	Early Blight	1000	700	200	100
C2	Late Blight	1909	1336	382	191
C3	Leaf Mold	952	666	190	95
C4	Septoria Leaf Spot	1771	1240	354	177
C5	Spider Mite	1676	1173	335	168
C6	Target Spot	1404	983	281	140
C7	Yellow Leaf Curl Virus	5357	3750	1071	536
C8	Mosaic Virus	373	261	75	37
C9	Healthy	1591	1114	318	159
	Total	18160	12712	3632	1816

Figure. 2 Images after data augmentation (from left to right: Original, height shift range, horizontal flip, shear range, vertical flip, width shift range, zoom range)

2. Proposed methodology

Fig. 1 illustrates the proposed methodology for identifying tomato leaf diseases using a hybrid improved vision transformer and deep neural networks frameworks.

2.1 Dataset description and preparation

The study utilized tomato plant images from the PlantVillage dataset [16, 17], which consists of 18,160 tomato leaf images categorized into 10 classes. Table 1 presents images in the RGB color space, with a 256 × 256 pixels resolution.

2.2 Preprocessing

The preprocessing step enhances the image quality for leaf detection. The initial step involves resizing the images to dimensions of 96 × 96. Following this, the images are rescaled and subjected to data augmentation. The final training dataset for this research was generated using data augmentation approaches. This study employed various image augmentation techniques, including horizontal and vertical flip, shear and zoom range of 0.15, and width and height shift of 0.15. Fig. 2 depicts randomly chosen images and their corresponding enhanced versions.

2.3 Improved vision transformer

The vision transformer (ViTs) [18] architecture

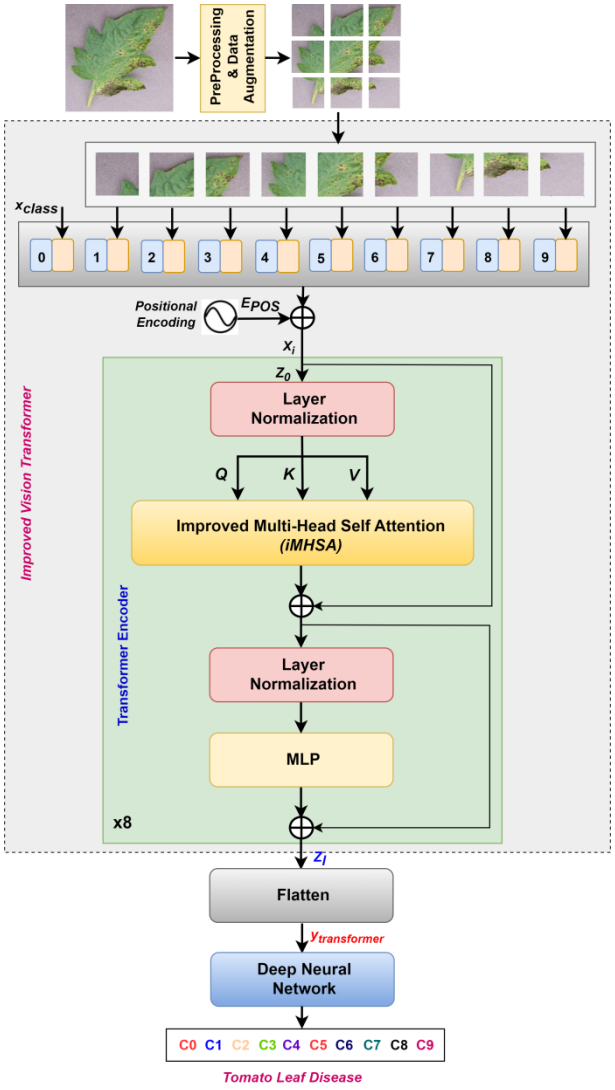


Figure. 3 Proposed model including improved vision transformer

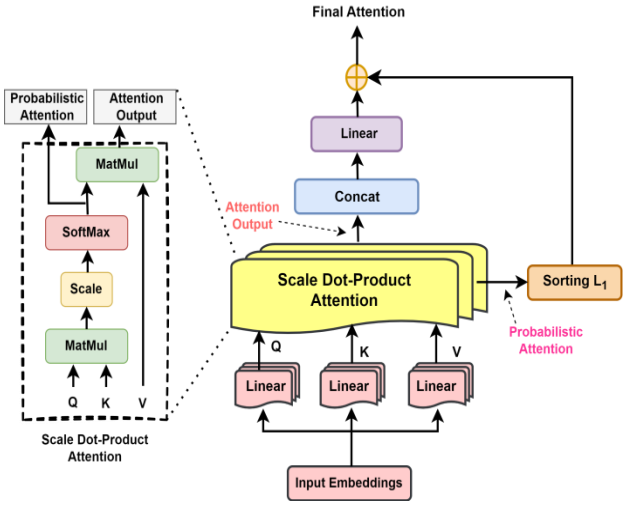


Figure. 4 Scale dot product with residual L1-norm

introduced the concept of processing images as sequences of patches, thereby leveraging the success of transformers[18] in natural language processing for computer vision tasks. While ViT has shown impressive results, there is room for improvement, particularly in enhancing the multi-head self-attention mechanism, show in Fig. 3.

A. *Patch generation & positional encoding*

ViT takes the 2D image $\boldsymbol{I} \in \mathbb{R}^{\boldsymbol{h\times w\times C}}$ and partition it into non-overlapping patches of size $\boldsymbol{P\times P}$. The image undergoes a transformation process, resulting in a patch sequence represented as $\boldsymbol{I_p} \in \mathbb{R}^{\boldsymbol{N\times(P^2\cdot C)}}$, where P is patch size, and N is the number of patches calculated as $\boldsymbol{N = (h\times w)/P^2}$. The process of embedding involves taking input patches, flattening them, and projecting them into a D-dimensional space using a linear projection $\boldsymbol{E} = \left[\boldsymbol{I_p^1}, \boldsymbol{I_p^2}, \ldots, \boldsymbol{I_p^N}\right]$. The patch embedding is integrated with the class embedding $\boldsymbol{x_{class}} \in \mathbb{R}^{\boldsymbol{1\times D}}$. Position embedding are incorporated into the combined embedding to preserve the positional details.

$$E_{POS} = \begin{cases} \sin\left(pos/1000^{2i/d}\right), & if\ i\ is\ even \\ \cos\left(pos/1000^{2i/d}\right), & if\ i\ is\ odd \end{cases} \quad (1)$$

The embedding procedure is represented by the equation.

$$X_i = \left[x_{class}; I_p^1, I_p^2, \cdots, I_p^N\right] + E_{POS} \quad (2)$$

where $\boldsymbol{X_i} \in \mathbb{R}^{\boldsymbol{(N+1)\times D}}$ represents final embedding for i-th patch and $\boldsymbol{E_{POS}} \in \mathbb{R}^{\boldsymbol{(N+1)\times D}}$, D represents the transformer's latent vector size.

B. *Improved multi-head self attention (iMHSA)*

In this segment, we introduce the improved multi-head self-attention (*iMHSA*) shows in the Fig. 4. iMHSA involves multiple attention heads, each of which learns different aspects of the relationships between patches. Attention is a critical component in capturing the relationships among sequential input patches.

Scale dot product with residual L1-norm

The scaled dot-product attention mechanism involves multiple queries, keys, and values, shows in Fig. 4. Queries and keys have a dimensionality of d_k, while values have a dimensionality of d_v. The dot product is calculated between the query and all keys. The values obtained are divided by the scale factor $\sqrt{d_k}$ and then subjected to a softmax algorithm. The process of attention calculation encompasses the utilization of three distinct embedding matrices, namely Key($\mathcal{K}$), Query($\mathcal{Q}$)

and Value($\mathcal{V}$). Let's assume we have ***H*** attention heads. For each head *h*,

$$\mathcal{Q}_{\mathrm{h}} = \mathrm{X_i} \cdot \mathcal{W}_{\mathrm{h}}^{\mathcal{Q}};\ \mathcal{K}_{\mathrm{h}} = \mathrm{X_i} \cdot \mathcal{W}_{\mathrm{h}}^{\mathcal{K}};\ \mathcal{V}_{\mathrm{h}} = \mathrm{X_i} \cdot \mathcal{W}_{\mathrm{h}}^{\mathcal{V}} \quad (3)$$

where $\mathcal{W}_{\mathrm{h}}^{\mathcal{Q}}, \mathcal{W}_{\mathrm{h}}^{\mathcal{K}}$ and $\mathcal{W}_{\mathrm{h}}^{\mathcal{V}}$ are weight matrices specific to each attention head.

Next, we compute the probabilistic attention scores ($\mathrm{PA_h}$) for each head by applying the softmax function to the scaled dot product of $\mathcal{Q}_{\mathrm{h}}$ and $\mathcal{K}_{\mathrm{h}}$

$$PA_h = SoftMax\left(\frac{\mathcal{Q}_h {\mathcal{K}_h}^T}{\sqrt{d_k}}\right) \quad (4)$$

The attention output is obtained by multiplying the probabilistic attention scores (PA_h) with the value ($\mathcal{V}$)

$$A_h = PA_h \cdot \mathcal{V} \quad (5)$$

After obtaining the attention output, we concatenate the outputs from all attention heads and apply a linear projection layer to reduce the dimensionality

$$Linear = Concatenate(A_1, A_2, \dots, A_H) \cdot \mathcal{W}_0 \quad (6)$$

Here, $\mathcal{W}_0$ is the weight matrix for the final linear projection.

The L1 norms of the probabilistic attention outputs from each attention head are computed and sorted. For each patch $\mathrm{P_{i,j}}$, this involves summing the absolute values of the attention scores

$$L1 - Norm_h = \sum_{k=1}^{N} |PA_{h,k}| \quad (7)$$

Identify the patch with the highest L1-norm, representing the most attentive patch (MA_{patch}) for this specific attention head

$$MA_{patch_h} = arg\max_h L1 - Norm_h \quad (8)$$

The residual term for the most attentive patch is calculated by directly using the probabilistic attention score

$$Residual = PA_{MA_{patch}} \quad (9)$$

where $PA_{MA_{patch}}$, attention corresponding to best L1-norm.

The improved multi-head self attention output ($iMHSA_{output}$) output is obtained by combining the attention from the linear projection layer and attention corresponding to best L1-norm.

$$iMHSA_{output} = Linear + Residual \quad (10)$$

C. ***Layer normalization and multi layer perceptron(MLP)***

The output of improved MHSA mechanism is normalized and then passed through a feed-forward network.

Layer normalization

Layer normalization is a technique used in deep learning to normalize the activations within a neural network layer. The proposed method is similar to batch normalization but normalizes the features within a single data point instead of across a batch of data. Batch normalization is a method that tackles the issue of internal covariate shift and improves the training and generalization capabilities of deep neural networks.

Multi layer perceptron (MLP)

Finally, the $iMHSA_{output}$ output is used as input to a *MLP* layer, which typically consists of two linear transformation. In the subsequent layers, non-linearity is introduced using a multilayer perceptron (MLP) with the Gaussian error linear unit (GeLU). The Z_0 signifies the initial input fed into a transformer encoder consisting of L layers. The encoder comprises an improved multi-head self attention $iMHSA_{output}$, layer normalization *LN* and multiplayer perceptron MLP.

$$Z_l' = iMHSA_{output}(LN(Z_{l-1}) + Z_{l-1} \quad (11)$$

$$Z_l = MLP\big(LN(Z_l')\big) + Z_l' \quad (12)$$

where $Z_l \in \mathbb{R}^{(N+1)\times D}$ and $l = 0,1,2,\dots,L$ is transformer block layer.

Now the final output of transformer's encoder blocks Z_l, is flattened for classification task.

$$y_{transformer} = flatten(Z_l) \quad (13)$$

The optimum parameters for feature extractor of proposed model are summarized in Table 2.

2.4 Deep neural networks classifier

The extracted features, represented as $\mathrm{y_{transformer}}$, are inputted into a deep neural network classifier for final classification. $\mathrm{y_{transformer}}$, undergoes a sequence of dense and dropout layers, yielding the output $\mathrm{y_1}$. The subsequent processing

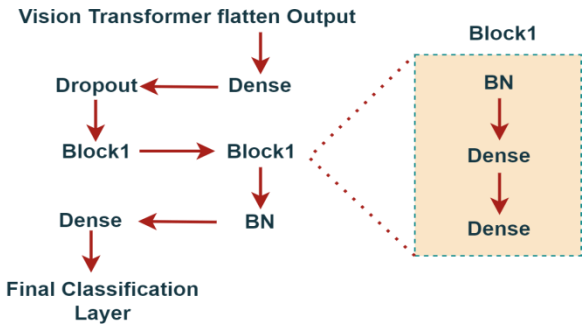


Figure. 5 DNN classifier

Table 2. Parameter for improved vision transformer

Parameter	Values
Image Size	96×96
Patch Size	6×6
Number of Patch (N)	256
Projection Dimension	64
Attention Head	3
Encoder layer	8
Transformer Unit	(128,64)
MLP Unit	(2048,1024)

involves two iterative blocks referred to as Block1. The block consists of three layers: a layer normalization (LN) layer and two dense layers, as depicted in Fig. 5.

$$y_1 = dropout\left(dense\left(y_{transformer}\right)\right) \quad (14)$$

$$Block1 = f(LN, dense, dense) \quad (15)$$

$$y_2 = Block1(y_1) \quad (16)$$

The concluding stage entails the utilization of a dense layer and a softmax layer to classify the accurate tomato leaf disease from the PlantVillage dataset.

$$Y_{prediction} = SoftMax(y_2) \quad (17)$$

3. Experimental setups

3.1 Implementation details

PyTorch was employed in our experimental setup on Kaggle. The model was trained on the PlantVillage dataset, explicitly focusing on tomato leaf images. The training method utilized categorical cross-entropy loss, and weight updates were executed using the adam optimizer, which enhances model performance through weight value adjustments. The training comprised 100 epochs with the implementing of a learning rate scheduler

Table 3. Hyperparameter for proposed model

Hyperparameter	Values
Learning rate	$1 - e4$
Epoch	100
Optimizer	$Adam$
Batch size	32
Scheduler	Learning Rate Scheduler
Loss function	Categorical Cross Entropy

Table 4. Comparing the performance of different decoder block and attention head configurations in the proposed frameworks

Number of Encoder	Number of Attention Head	Accuracy (%)
1	8	91.28
	12	92.57
2	8	94.91
	12	95.14
3	**8**	**97.24**
	12	96.59
4	8	96.27
	12	96.12

Table 5. Effect on modules on model performance

Models	Acc (%)	Loss (%)	Parameters
ViTs	97.24	15.92	2,535,889
ViTs + L1-norm	98.91	11.28	2,778,011
ViTs + L1-norm + Augmentation	99.74	8.10	3,159,348

for optimizing the learning rate. Table 3 presents a comprehensive summary of the hyperparameters.

3.2 Evaluation

Four generally used measures are utilized to objectively evaluate the effectiveness of the suggested methodology for detecting tomato diseases: precision (P), recall (R), F1 score, and accuracy (Acc).

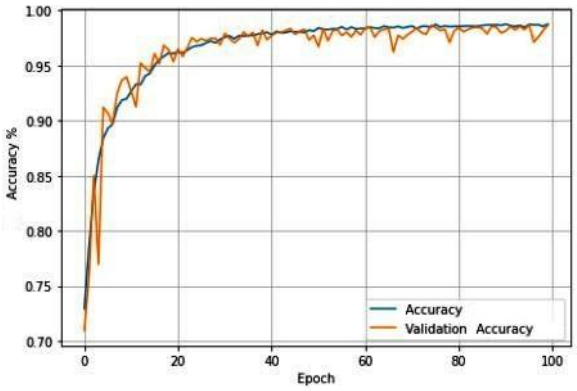
$$Accuracy\ (Acc) = \frac{P_T + N_T}{P_T + P_F + N_T + N_F} \quad (18)$$

$$Precision\ (P) = \frac{P_T}{P_T + P_F} \quad (19)$$

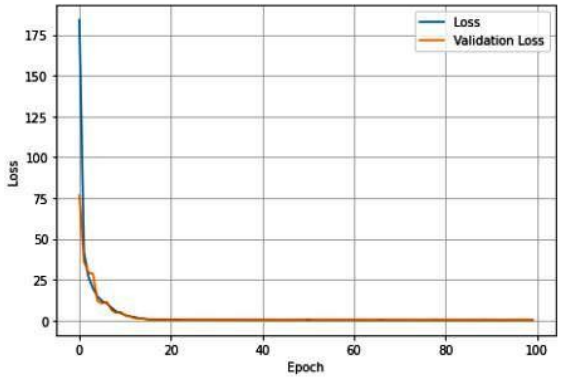
$$Recall\ (R) = \frac{P_T}{P_T + N_F} \quad (20)$$

$$F1 - score = \frac{2 \times (P \times R)}{(P + R)} \quad (21)$$

$Where\ P_T$, N_T, P_F, and N_F are true positives, true negatives, false positives and false negative respectively.



(a)



(b)

Figure. 6: (a) Accuracy curve and (b) Loss curve

4. Experimental result analysis

4.1 Ablation study

This study employed the vision transformer (ViTs) baseline model to examine the optimal combination of attention heads and decoder blocks for maximizing accuracy. Our study found that using 8 attention heads in both the encoder and decoder components produced positive results. Based on the positive results shown in Table 4, we intentionally adjusted the number of decoder blocks to three per iteration. The decision aimed to optimize the trade-off between model complexity and efficiency. The inclusion of three decoder blocks enhanced information processing and mitigated overfitting. Table 5 presents the impact of L1-norm regularization and data augmentation on the accuracy of the ViTs model. These techniques optimize model performance while maintaining the trainable parameter. The results suggest that these techniques have a positive effect on improving model performance.

4.2 Performance evaluation of model

The experiments were conducted on the Kaggle platform. Fig. 6 displays the epoch-wise graph of accuracy and loss for the tomato disease dataset during training and validation. Fig. 6(a) displays the accuracy curve for the training and validation datasets. The fluctuation observed in the curve can be attributed to the imbalanced distribution of tomato leaf image data within certain categories. The validation curve in Fig. 6(b) demonstrates perfect convergence.

The confusion matrix in Fig. 7 displays the results of the proposed model on the test set. It includes the number of correct predictions and errors for each class. The model demonstrates strong performance, with a majority of accurate predictions. The model exhibits minor errors, specifically false positives and false negatives in certain classes. As presented in Table 6, the proposed model demonstrates high effectiveness across various classes, particularly for early blight, late blight, and others, achieving near-perfect precision, recall, and F1 scores. This finding indicates that the proposed model demonstrates reliability and accuracy in its predictive capabilities. In certain classes, such as mosaic virus, there is potential for enhanced precision. The results suggest a robust classification model with high accuracy and reliability across various categories.

4.3 Comparison with other state-of-art methods

The present research comprehensively analyzes several cutting-edge algorithms used to identify tomato leaf diseases using the PlantVillage dataset. The findings of this investigation are succinctly presented in Table 7. The model using ViT + L1-norm regularization with augmentation performs well, with a remarkable accuracy rate of 99.79%. Furthermore, it has remarkable levels of accuracy (99.23%), recall (99.65%), and F1 score (99.44%). The two stage transfer learning approach has shown significant outcomes, achieving a noteworthy accuracy rate of 99.23% while preserving a well-balanced performance in terms of precision, recall, and F1 score. While the VGG19 with ViT model demonstrates satisfactory performance, it shows worse outcomes than the top two methods in terms of all analyzed criteria. ViT with an inception module and CNN with 3 channel attention show some parallels in their effects. However, it is essential to note that there may be compromises regarding accuracy, recall, and F1 score.

The combination of convolutional neural networks (CNN) and vision transformer (ViT) demonstrates significant levels of accuracy and recall, suggesting potential applications in specific disciplines. The findings provide valuable insights

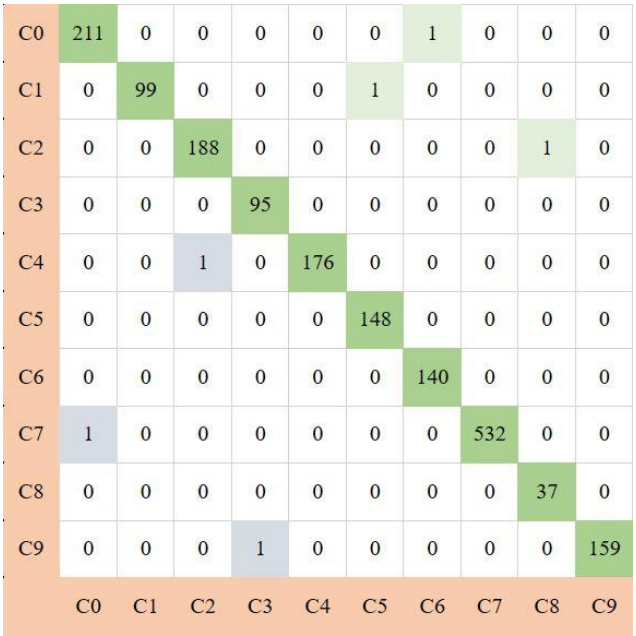


Figure. 7 Confusion matrix

Table 6. Experimental findings of our models(in %)

Classes	P	R	F1
Bacterial Spot	99.53	99.53	99.53
Early Blight	99.00	100	100
Late Blight	99.47	99.47	99.47
Leaf Mold	100	98.96	99.48
Septoria Leaf Spot	99.44	100	99.72
Spider Mite	100	99.33	99.66
Target Spot	100	99.29	99.64
Yellow Leaf Curl Virus	99.81	100	99.91
Mosaic Virus	100	97.37	98.67
Healthy	99.38	100	99.69

Table 7. Compare the model performance (in %) with other

Ref	Algorithms	Acc	P	R	F1
[5]	2 Stage Transfer learning	99.23	99.12	99.29	99.20
[13]	VGG19 + ViT	98.81	98.72	98.76	98.73
[14]	ViT with inception module in place of MLP	96.93	96.98	96.83	96.90
[10]	CNN + 3 Channel Attention	97.56	98.00	98.00	97.00
[15]	CNN + ViT	-	96.59	96.80	96.69
Our	ViT + L1-norm with Augmentation	**99.74**	**99.66**	**99.39**	**99.53**

for researchers and practitioners engaged in algorithm selection, particularly emphasizing the importance of accuracy and recall metrics tailored to individual tasks.

5. Conclusion

This study introduces a novel approach for the detection of tomato leaf diseases. The approach utilizes a hybrid improved vision transformer and deep neural networks framework. The results indicate that the model has the ability to accurately classify various tomato leaf diseases, achieving an impressive accuracy rate of 99.74%. The study showcases that incorporating L1-norm attention and data augmentation techniques significantly enhances the model's performance. The model's ability to enhance precision, recall, and F1 scores for each disease category highlights its versatility and effectiveness. This study introduces a new disease detection method in agricultural technology that is scalable and efficient. Deep learning, particularly vision transformers, has made significant progress in disease diagnosis. The emphasis of this approach lies in the importance of model interpretability for informed agricultural decision-making. This study makes a significant contribution to improving tomato crop health and productivity, addressing food security concerns, and enhancing economic sustainability in agricultural communities. Future refinements and integration of internet of things (IoT) technologies promise to enhance efficiency and sustainability in agricultural practices.

Notation list

Variable	Description
h, w, C	Height, width and channel of image
N, P	Number of patches and patch size
x_{class}, E_{POS}	Class and Positional Embedding
X_i	Final embedding
Q, K, V	Query, Key, and Value matrices
$iMHSA$	Improved Multi-Head Attention
$\mathcal{W}_h^Q, \mathcal{W}_h^{\mathcal{K}}, \mathcal{W}_h^V$	Weight matrices
PA_h	Probabilistic attention of each head
A_h	Attention output of each head
$Linear$	Linear Projection layer
$L1-Norm_h$	L1 normalization of probabilistic attention
MA_{patch_h}	Most attentive patch from each head
$iMHSA_{output}$	Output of improved multi-head self attention
LN	Layer Normalization
MLP	Multi Layer Perceptron
$y_{transformer}$	Flattened features from transformer
$Y_{prediction}$	Final classification output

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, Hemant Kumar and Mamta Tiwari; *Methodology*, Hemant Kumar and Navin Prakash; *Software*, Shivneet Tripathi and Sunil Kumar; *Validation*, Ritesh Agarwal and Rahul Neware; *Data curation*, Hemant Kumar; *Writing—original draft preparation*, Hemant Kumar and Rahul Neware and Mamta Tiwari; *Writing— review and editing*, Navin Prakash and Hemant Kumar and Sunil Kumar; *Visualization*, Shivneet Tripathi and Ritesh Agarwal.

References

[1] A. G. Ibañez and A. R. Muñoz, "Monitoring Tomato Leaf Disease through Convolutional Neural Networks", *Electronics*, Vol. 12, No. 1, 2023, doi: 10.3390/electronics12010229.

[2] A. P. A. S. Rani and N. S. Singh, "Protecting the environment from pollution through early detection of infections on crops using the deep belief network in paddy", *Total Environment Research Themes*, Vol. 3–4, No. November, p. 100020, 2022, doi: 10.1016/j.totert.2022.100020.

[3] K. Roy, S. S. Chaudhari, J. Frnda, S. Bandopadhyay, I. J. Roy, and S. Banerjee, "Detection of Tomato Leaf Diseases for Agro-Based Industries Using Novel PCA DeepNet", *IEEE Access*, Vol. 11, pp. 14983–15001, 2023, doi: 10.1109/ACCESS.2023.3244499.

[4] R. Reedha, E. Dericquebourg, R. Canals, and A. Hafiane, "Transformer Neural Network for Weed and Crop Classification of High Resolution UAV Images", *Remote Sensing*, Vol. 14, No. 3, 2022, doi: 10.3390/rs14030592.

[5] T. Sanida, A. Sideris, M. V. Sanida, and M. Dasygenis, "Tomato leaf disease identification via two–stage transfer learning approach", *Smart Agricultural Technology*, Vol. 5, 2023, doi: 10.1016/j.atech.2023.100275.

[6] S. G. Paul, A. A. Biswas, A. Saha, S. Zulfikar, N. A. Ritu, and I. Zahan, "A real-time application-based convolutional neural network approach for tomato leaf disease classification", *Array*, Vol. 19, 2023, doi: 10.1016/j.array.2023.100313.

[7] Y. Zhong, Z. Teng, and M. Tong, "LightMixer: A novel lightweight convolutional neural network for tomato disease detection", *Frontiers in Plant Science*, Vol. 14, 2023, doi: 10.3389/fpls.2023.1166296.

[8] M. Shoaib, T. Hussain, B. Shah, I. Ullah, S. M. Shah, and F. Ali, "Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease", *Frontiers in Plant Science*, Vol. 13, 2022, doi: 10.3389/fpls.2022.1031748.

[9] M. Bhandari, T. B. Shahi, A. Neupane, and K. B. Walsh, "BotanicX-AI: Identification of Tomato Leaf Diseases Using an Explanation-Driven Deep-Learning Model", *Journal of Imaging*, Vol. 9, No. 2, 2023, doi: 10.3390/jimaging9020053.

[10] H. Chen, Y. Wang, P. Jiang, R. Zhang, and J. Peng, "LBFNet: A Tomato Leaf Disease Identification Model Based on Three-Channel Attention Mechanism and Quantitative Pruning", *Applied Sciences*, Vol. 13, No. 9, 2023, doi: 10.3390/app13095589.

[11] P. S. Thakur, S. Chaturvedi, P. Khanna, T. Sheorey, and A. Ojha, "Vision transformer meets convolutional neural network for plant disease classification", *Ecological Informatics*, Vol. 77, 2023, doi: 10.1016/j.ecoinf.2023.102245.

[12] S. Hossain, M. T. Reza, A. Chakrabarty, and Y. J. Jung, "Aggregating Different Scales of Attention on Feature Variants for Tomato Leaf Disease Diagnosis from Image Data: A Transformer Driven Study", *Sensors*, Vol. 23, No. 7, 2023, doi: 10.3390/s23073751.

[13] A. Tabbakh and S. S. Barpanda, "A Deep Features Extraction Model Based on the Transfer Learning Model and Vision Transformer 'TLMViT' for Plant Disease Classification", *IEEE Access*, Vol. 11, pp. 45377–45392, 2023, doi: 10.1109/ACCESS.2023.3273317.

[14] P. Gole, P. Bedi, S. Marwaha, M. A. Haque, and C. K. Deb, "TrIncNet: a lightweight vision transformer network for identification of plant diseases", *Frontiers in Plant Science*, Vol. 14, No. July, pp. 1–19, 2023, doi: 10.3389/fpls.2023.1221557.

[15] Y. Wang, Y. Chen, and D. Wang, "Convolution Network Enlightened Transformer for Regional Crop Disease Classification", *Electronics*, Vol. 11, No. 19, 2022, doi: 10.3390/electronics11193174.

[16] G. Geetharamani and A. P. J., "PlantVillage Tomato leaf image Dataset from Mendeley", 2019, https://data.mendeley.com/datasets/tywbtsjrjv/1 (accessed Oct. 15, 2023).

[17] G. Geetharamani and A. P. J., "Identification of

plant leaf diseases using a nine-layer deep convolutional neural network", *Comput. Electr. Eng.*, Vol. 76, pp. 323–338, 2019, doi: 10.1016/j.compeleceng.2019.04.011.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", In: *Proc. of 9th International Conference on Learning Representations (ICLR)*, pp.1-22, 2021.