TLDViT: A Vision Transformer Model for Tomato Leaf Disease Classification

Sami Aziz Alshammari
Department of Information Technology-Faculty of Computing and Information Technology
Northern Border University, Saudia Arabia

Abstract—**Accurate and efficient diagnostic methods are essential for crop health monitoring due to the substantial impact of tomato leaf diseases on crop yield and quality. Traditional machine learning models, such as convolutional neural networks (CNNs), have shown promise in plant disease classification; however, they often require extensive data preprocessing and struggle with complex variations in leaf appearance. This study introduces TLDViT (Tomato Leaf Disease Vision Transformer), a Vision Transformer model specifically designed for the classification of tomato leaf diseases. TLDViT reduces the need for preprocessing by learning disease-specific features directly from raw images, leveraging Vision Transformers' ability to capture long-range dependencies within images. We evaluated TLDViT on the Plant Village Dataset, which includes healthy and diseased samples across multiple classes. For comparative analysis, two Vision Transformer models, ViT-r50-l32 and ViT-l16-fe, were tested. Among these, ViT-r50-l32 achieved the highest performance, surpassing both ViT-l16-fe with an accuracy of 98%. These findings highlight TLDViT's potential as an effective tool for crop health monitoring and automated plant disease diagnosis.**

Keywords*—*Tomato Leaf Disease; Vision Transformer (ViT); crop health monitoring; plant disease classification

I. Introduction

Agriculture is fundamental to global food security, and enhancing crop health management is crucial for maintaining production and reducing economic losses. Tomato (Solanum lycopersicum) is among the most extensively farmed crops globally, however it is very vulnerable to several foliar diseases, such as early blight, late blight, and leaf mold. These illnesses, mostly induced by pathogens including fungus, bacteria, and viruses, result in substantial decreases in production and quality [1], [2]. Accurate early detection and classification of these illnesses is essential for facilitating prompt and focused therapies, which may help reduce future transmission and harm. Conventional techniques for diagnosing plant diseases depend significantly on manual visual assessment and expert expertise, which are labor-intensive, expensive, and susceptible to subjective inaccuracies [3]. Advances in artificial intelligence (AI) and machine learning (ML) have shown potential to overcome these constraints through the automation of disease diagnosis. Convolutional neural networks (CNNs), a prevalent deep learning methodology, have shown efficacy in recognizing intricate patterns in plant disease imagery, attaining high accuracy in classification tests across several crop illnesses [4]. The authors of [3] used CNNs on an extensive dataset of plant diseases, achieving classification accuracies of 90% across 26 distinct crops. In [4], authors introduced advanced CNN architectures to improve the categorization of plant diseases, particularly those impacting tomatoes, but with considerable preprocessing and computing demands. These studies highlight the promise of CNNs while also exposing significant obstacles, including their reliance on large labeled datasets, susceptibility to overfitting, and constraints in capturing non-local connections in images [5]. Researchers have investigated different models and strategies to enhance the resilience and efficiency of plant disease classification systems, addressing these constraints. The author in [6] integrated handmade features with deep learning models, enhancing the robustness of CNNs against variability in image data, while [7] supplemented restricted datasets with synthetic images to elevate CNN performance. Notwithstanding these efforts, CNN-based models exhibit constraints in their ability to apprehend global spatial linkages within images, a factor that is especially critical in plant disease categorization, where symptoms may appear in non-contiguous areas on the leaf.

In recent years, Vision Transformers (ViT) have surfaced as a formidable alternative to CNNs for identification of images tasks [8], [9], [10], [11]. In contrast to CNNs, which depend on local convolutional filters for hierarchical feature extraction, ViT use self-attention processes to capture long-range relationships over the whole image. The global attention mechanism enables ViT to comprehend spatial connections from a comprehensive viewpoint, making them especially adept at image processing tasks that need acute sensitivity to spatial intricacies. The author in [12] shows that ViT may get superior performance on extensive image classification datasets, surpassing CNNs in both precision and efficiency. In [13], the author emphasized the promise of ViT in applications necessitating intricate spatial analysis, including medical imaging and remote sensing. The author in [14] used Vision Transformers for agricultural disease detection, proving their efficacy in identifying disease patterns in crops such as rice and wheat; nevertheless, research on their application to tomato leaf diseases is still scarce.

This paper presents TLDViT (Tomato Leaf Disease Vision Transformer), a Vision Transformer model particularly developed for the classification of tomato leaf diseases, motivated by recent breakthroughs. Our methodology utilizes the ViT architecture's capacity to capture long-range relationships, allowing it to identify nuanced and intricate disease patterns that CNNs may overlook. TLDViT, in contrast to CNN-based methods that need considerable preprocessing and data augmentation, is designed to immediately learn disease-specific features from minimally processed images, enhancing its adaptability and efficiency for practical agricultural applications.

Our study provides multiple contributions to the field of automated plant disease identification. The proposal introduces

TLDViT, a new Vision Transformer model tailored for the classification of tomato leaf diseases. Furthermore, two Vision Transformer models, ViT-r50-l32 and ViT-l16-fe, were employed to establish a comparative framework, ensuring that all models were trained on the Plant Village Dataset for consistency and robustness. A comprehensive comparison of model performances demonstrated that TLDViT exhibits superior accuracy compared to CNN-based methods and the two Vision Transformer models, underscoring its efficacy in this context. The study illustrates the benefits of Vision Transformers in agricultural diagnostics, emphasizing their sensitivity to spatial details, which is crucial for precise disease identification. These contributions enhance the application of Vision Transformers in plant disease detection and establish a basis for wider use in agricultural diagnostics.

The rest of the paper is structured as follows: Section II introduces the Literature Review, where we discuss related work on plant disease classification, highlighting the advantages and limitations of existing deep learning approaches, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The proposed methodology for classifying tomato leaf diseases is presented in Section III, which also covers the TLDViT model architecture, training procedures, and data preparation. The findings and discussion, together with performance comparisons and analysis, are presented in Section IV. Finally, Section V concludes the paper and outlines future research directions.

II. Literature Review

The latest developments in deep learning have markedly improved systems for the identification and categorization of plant diseases. Numerous research have investigated alternate methods to enhance the precision and efficacy of these systems. Initial approaches mostly depended on manually produced features and traditional machine learning methodologies [15], [16]. Convolutional neural networks (CNNs) exhibit exceptional performance in image-based illness classification; yet, their dependence on considerable preprocessing and difficulty in capturing global picture dependencies provide significant problems [17]. Hybrid models that combine CNNs with alternative architectures have shown enhanced robustness to fluctuations in picture quality [18], [19]. The advent of Vision Transformers (ViTs) has offered a persuasive alternative to CNNs for agricultural applications. Vision Transformers use self-attention processes to record long-range dependencies, allowing for the analysis of complex spatial patterns in pictures [20], [21]. Applications of Vision Transformers (ViTs) in the identification of plant diseases, including those affecting rice, wheat, and grapes, have shown enhanced efficacy relative to conventional Convolutional Neural Networks (CNNs) [22]. Recent research has used transfer learning with transformer architectures to address data scarcity challenges in agricultural datasets [23]. The integration of transformers with real-time systems and edge devices is becoming prevalent, with the objective of implementing disease detection models directly in agricultural fields for practical use [24]. Nevertheless, few research has concentrated especially on tomato leaf diseases, highlighting the need for a specialized Vision Transformer model to fill this void.

III. Proposed Approach-Based Tomato Leaf Disease Classification

This section describes our method to diagnosing tomato leaf illnesses using TLDViT (Tomato Leaf Disease Vision Transformer), a hybrid Vision Transformer model designed to capture both localized and global patterns in leaf photos. TLDViT combines ResNet-50's feature extraction powers with Vision Transformers' self-attention capabilities, resulting in a robust tool for detecting and recognizing disease signs in tomato leaves. We describe the major processes in our technique below, which include data preparation, model construction, training, and evaluation.

A. Data Preprocessing

Data preprocessing is a crucial phase to guarantee the quality and uniformity of the pictures used for training the TLDViT model. The dataset consists of images of tomato leaves, classified into six categories: Healthy, Bacterial Spot, Early Blight, Late Blight, Septoria Leaf Spot, and Yellow Leaf Curl Virus. The dataset used for this study is the publicly available Plant Village Dataset [25], which provides a comprehensive set of labeled images representing various plant diseases. This dataset is widely used for plant disease classification tasks and offers high-quality images that ensure accurate training and evaluation of the TLDViT model. All images are scaled to 224x224 pixels to standardize input dimensions, so minimizing computing effort while preserving enough information for precise categorization. Each pixel intensity is standardized to the interval [0, 1], enhancing the stability of the training process and facilitating more effective model learning. To improve the model's resilience and mitigate overfitting, many data augmentation methods are used, such as rotation, horizontal flipping, and brightness modifications. These changes create variances in the dataset, allowing TLDViT to generalize well across diverse lighting and ambient circumstances, which is essential for practical use.

Fig. 1 depicts the class distributions of tomato leaf disease images before to and after to data augmentation. Before augmentation (blue bars), the dataset comprised a total of 10,958 images unevenly allocated among six categories: Bacterial Spot (1,925 images), Early Blight (1,702 images), Healthy (1,920 images), Late Blight (1,705 images), Septoria Leaf Spot (1,745 images), and Yellow Leaf Curl Virus (1,961 images). Following augmentation (orange bars), the dataset dramatically increased to 13,603 images, enhancing class equilibrium. The post-augmentation dataset comprises 2,084 photos of Bacterial Spot, 2,352 images of Early Blight, 2,358 photographs of Healthy specimens, 2,267 images of Late Blight, 2,140 images of Septoria Leaf Spot, and 2,402 images of Yellow Leaf Curl Virus. This augmentation approach guarantees a more equitable dataset, which is essential for training machine learning models to generalize proficiently across all categories.

The used dataset is partitioned into three subsets: training (70%), validation (20%), and test (10%), allowing an equitable assessment of the model's efficacy on novel data. The training set is used to train the model, the validation set is applied for hyperparameter optimization and monitoring throughout training, and the test set offers a conclusive evaluation of the model's classification accuracy across six categories.

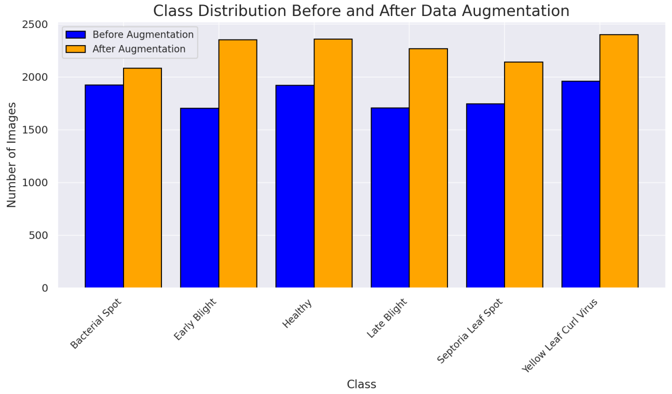

Fig. 1. Tomato class distribution.

B. TLDViT Model (Tomato Leaf Disease Vision Transformer)

The Vision Transformer (ViT) architecture, as presented in Fig. 2, has been customized to classify tomato leaf maladies into six categories: Healthy, Bacterial Spot, Early Blight, Late Blight, Septoria Leaf Spot, and Yellow Leaf Curl Virus. The model is highly effective in capturing complex disease patterns by combining local feature extraction with global context modeling.

1) Image patch division and flattening: The first phase is partitioning each input picture of a tomato leaf into a grid of smaller segments. A 224 x 224-pixel image may be divided into 32 x 32 pixel patches, resulting in 49 patches arranged in a 7 x 7 grid. Subsequently, each patch is transformed into a one-dimensional vector. This patch-based method collects intricate local details inside each leaf segment, enabling the model to identify localized disease indicators such as spots, discolorations, and texture alterations unique to certain illnesses.

2) Linear projection of flattened patches: The flattened patches undergo a linear projection layer, converting each patch into a high-dimensional vector appropriate for further processing in the Vision Transformer. This transformation produces a series of patch embeddings that preserve the localized features of each patch while mapping them into a higher-dimensional space, allowing the model to interpret the picture as a sequence instead of a grid.

3) Positional embedding and class token: Positional embeddings are included into each patch embedding to maintain the spatial configuration of patches inside the original image. The positional embeddings provide the model with data on the location of each patch inside the image, which is essential for comprehending spatial links among illness symptoms. Furthermore, an additional learnable class token is attached to the series of patch embeddings. The class token engages with the patch embeddings throughout the transformer layers and ultimately retains the information required to generate the final classification label.

4) Transformer encoder: The fundamental component of the design is the Transformer Encoder, including numerous layers that integrate self-attention mechanisms with feed-forward neural networks. Each encoder layer has many essential components: The Multi-Head Self-Attention mechanism, which allows the model to concentrate on several sections of the leaf concurrently, therefore capturing both local intricacies and overarching patterns within the picture. This multi-head attention enables the model to discern intricate relationships across patches, facilitating the identification of disease-related patterns that may be distributed over different areas of the leaf. Normalization (Norm) layers are used to stabilize the learning process and mitigate overfitting by guaranteeing that inputs to each layer possess a standardized distribution, facilitating model convergence and enhancing generalization. Subsequent to the self-attention mechanism, a Multi-Layer Perceptron (MLP) introduces non-linear changes to the representations, therefore augmenting the model's capacity to discern intricate patterns pertinent to each illness type and boosting classification precision. This encoder architecture enables the model to analyze the whole image as a series of patches, use self-attention to discern correlations both internally and externally among the patches. This is especially beneficial in the categorization of plant diseases, where symptoms may manifest in scattered patterns or as nuanced textural alterations on the leaf.

5) MLP head and classifier: The class token's embedding is supplied into a Multi-Layer Perceptron (MLP) head and subsequently into a classifier, which generates the final prediction, after passing through the transformer layers. The classifier attributes the image to one of the six classes, thereby designating the leaf as healthy or indicating the specific type of disease present. The MLP head is the ultimate stage in the processing process, incorporating all the information acquired by the transformer layers to provide a precise diagnosis.

This ViT architecture is particularly effective for the classification of tomato leaf diseases because it can manage both local and global image features. The self-attention mechanism enables the model to interpret relationships across the entire image, while the patch-based approach captures detailed visual features within small sections, which is essential for identifying disease-specific symptoms. The distinction between diseases that may appear visually similar but have unique patterns or spread across various areas of the leaf is dependent on the combination of local and global context.

C. Training Methodology

In order to optimize the performance of TLDViT, we implement a systematic training approach that incorporates exhaustive evaluation techniques, optimization, and regularization. In order to reduce prediction errors across all disease classes, categorical cross-entropy loss is implemented, as this is a multi-class classification problem. The Adam optimizer is employed with an initial learning rate of 0.0001, which is progressively reduced by a learning rate scheduler as training progresses. This approach assists in the stabilization of the model and the enhancement of convergence, while also preventing overshooting. The high model complexity necessitates the use of regularization methods, such as dropout layers inside the transformer component, to prevent overfitting. In order to avoid superfluous epochs and overfitting, early stopping is sometimes used. This involves monitoring validation accuracy and ending training as performance stabilizes. To achieve a happy medium between computing efficiency and enough iteration for learning, the model is trained with a batch size of 64 for 25 epochs.

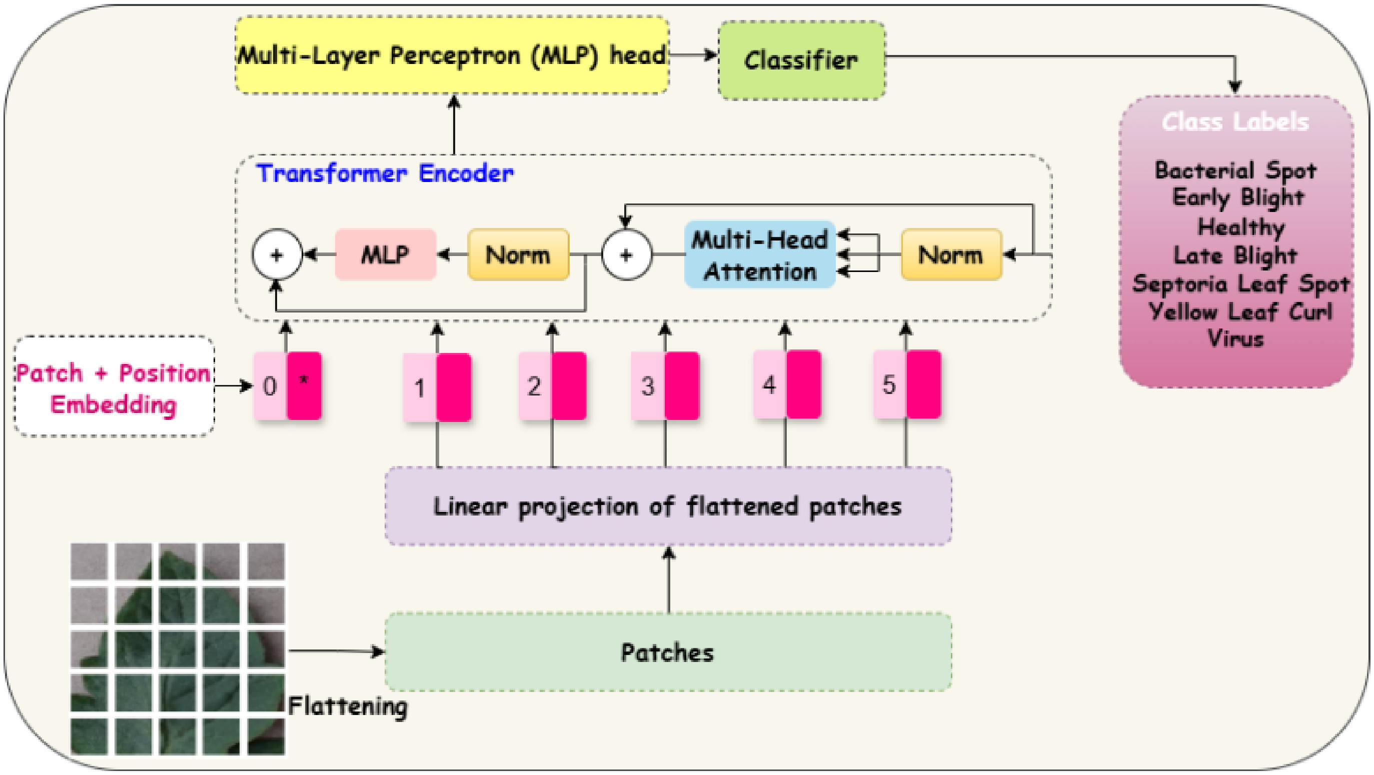

Fig. 2. TLDViT architecture.

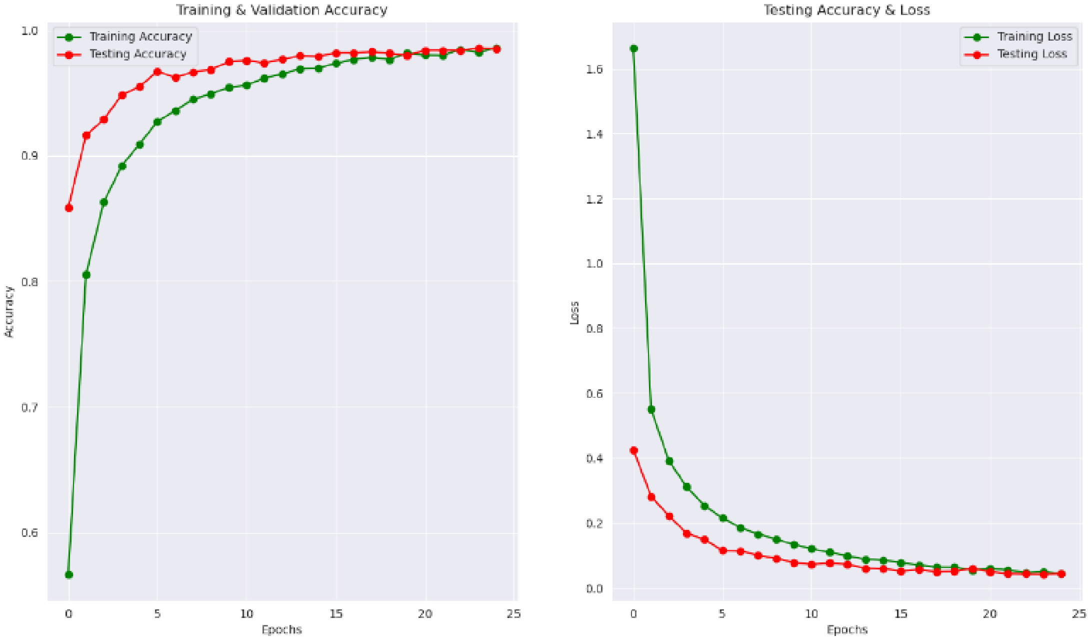

Fig. 3. Accuracy and loss curve of ViT-r50-l32 model.

We evaluate TLDViT's efficacy in accurately categorizing tomato leaf diseases using the following metrics: accuracy, F1-score, precision, recall, and confusion matrix. Making ensuring the model satisfies all criteria, this assessment checks it thoroughly. Additionally, we use ROC curves as a measure to evaluate the model's performance across multiple thresholds, especially when distinguishing across interrelated illness types. To find out how well the model can distinguish between classes at various decision thresholds, we build ROC curves for each class and then measure the area under the curve (AUC). The following equations introduced the performance evaluations:

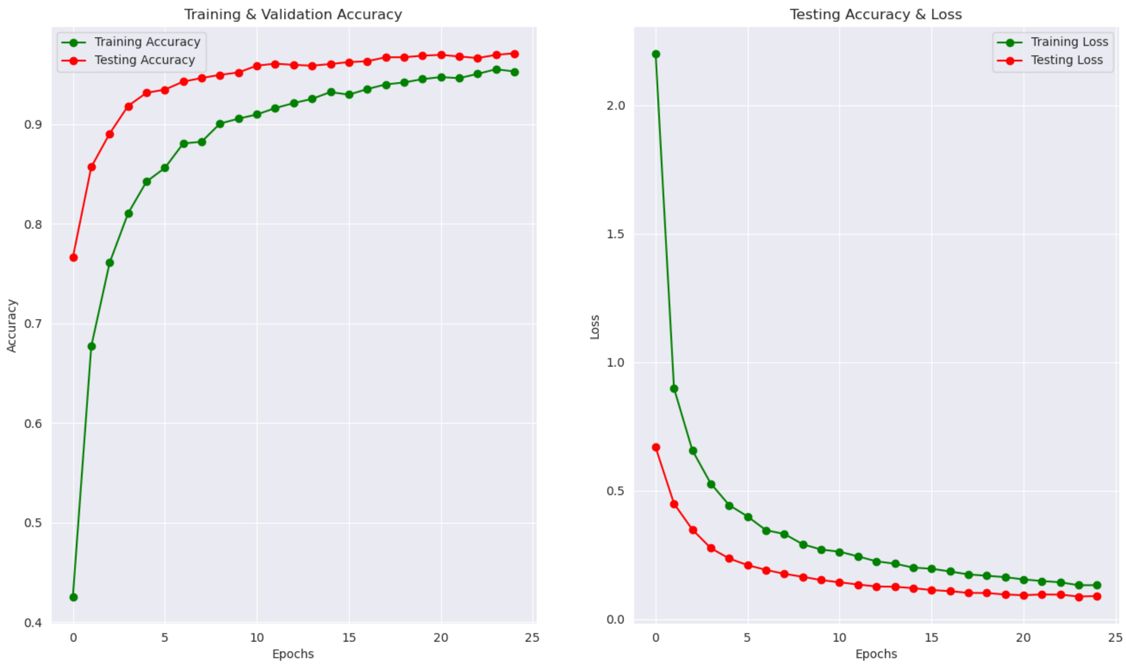

Fig. 4. Accuracy and loss curve of ViT-l16-fe model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

- True Positives (TP): Correctly predicted positive samples.
- True Negatives (TN): Correctly predicted negative samples.
- False Positives (FP): Incorrectly predicted positive samples.
- False Negatives (FN): Incorrectly predicted negative samples.

IV. Experimental Results

The experimental findings of the proposed TLDViT model for categorizing tomato leaf diseases into six labels are shown in this section. These categories are Healthy, Bacterial Spot, Early Blight, Late Blight, Septoria Leaf Spot, and Yellow Leaf Curl Virus. For the purpose of conducting a complete evaluation, we evaluate the performance of the model using a number of different measures, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

A. Classification Accuracy and Loss

Fig. 3 shows that training and testing accuracy increase with epochs in the ViT-r50-l32 model, stabilizing at high values. Within a few epochs, the model converges, with training accuracy around 99% and testing accuracy close behind. Regularly decreasing loss curves for training and testing indicate good learning without overfitting. The tight alignment of training and testing performance implies that ViT-r50-l32 can generalize to new data and discriminate tomato leaf disease classes. Fig. 4 shows that ViT-l16-fe has slower convergence and larger loss values during training, suggesting a weaker generalization capacity than ViT-r50-l32. Although ViT-l16-fe is accurate, it lacks the stability and minimum loss of ViT-r50-l32. These findings show that ViT-l16-fe is effective but may not capture disease-specific aspects as well as ViT-r50-l32.

The results of this study reveal that ViT-r50-l32 outperforms ViT-l16-fe in accuracy and loss measures, evidenced by its fast convergence, elevated final accuracy, and reduced loss levels throughout training and testing. The exceptional efficacy of ViT-r50-l32 indicates that its integration of a ResNet-50 backbone with Vision Transformer layers is especially adept at collecting complex illness characteristics, enabling more precise differentiation across disease categories. The constant and consistent performance shown in both training and testing reinforces the resilience of ViT-r50-l32.

B. Performance Comparison of Tomato Leaf Disease

Table I shows the classification performance of two models, ViT-r50-l32 and ViT-l16-fe, on tomato leaf disease categories, including Precision, Recall, F1-Score, and Overall Accuracy. F1-Scores of 0.95 or better are achieved by the ViT-r50-l32 model in all categories. It has excellent precision and recall for "Late Blight" and "Septoria Leaf Spot," an F1-Score of 1.00 for both, and an accuracy of 0.98, showing good generalization. Though scoring lower in several areas, the ViT-l16-fe model performs well. For "Late Blight" it has an F1-Score of 0.94 owing to a minor loss in accuracy, but it has good precision and recall across most classes, especially for "Yellow Leaf Curl Virus" with 1.00 precision. Though somewhat lower than ViT-r50-l32, ViT-l16-fe has solid classification performance with an accuracy of 0.97. In conclusion, both models have good accuracy and F1-Scores across all categories, although ViT-r50-l32 may be preferable for tomato leaf disease classification in this dataset.

TABLE I. PERFORMANCE CLASSIFICATION OF TLDVIT MODELS

Model	Class	Precision	Recall	F1-Score
ViT-r50-l32	Bacterial Spot	0.93	0.98	0.95
	Early Blight	1.00	0.96	0.98
	Late Blight	1.00	1.00	1.00
	Septoria Leaf Spot	1.00	1.00	1.00
	Yellow Leaf Curl Virus	0.96	1.00	0.98
	Healthy	1.00	0.96	0.98
	Overall Accuracy	0.98		
ViT-l16-fe	Healthy	0.96	1.00	0.98
	Bacterial Spot	0.98	0.92	0.95
	Early Blight	0.98	1.00	0.99
	Late Blight	0.90	0.98	0.94
	Septoria Leaf Spot	1.00	0.95	0.98
	Yellow Leaf Curl Virus	1.00	0.96	0.98
	Overall Accuracy	0.97		

Fig. 5 displays the Precision-Recall (PR) curve for the ViT-r50-l32 model, which classified tomato leaf diseases well. Each curve symbolizes a disease: Healthy, Bacterial Spot, Early Blight, Late Blight, Septoria Leaf Spot, and Yellow Leaf Curl Virus. The legend shows each category's average accuracy score. Due to its near-perfect accuracy and recall across all classes, the model can reliably categorize each illness type without substantial false positives or negatives. All classes' aggregate performance is tinted blue, with an average accuracy of 0.997. The model excels on "Early Blight" and "Yellow Leaf Curl Virus," scoring 1.000 in precision-recall, while the remaining classes score 0.995–0.999. This curve shows the ViT-r50-l32 model's resilience and accuracy, making it ideal for identifying and differentiating tomato leaf diseases.

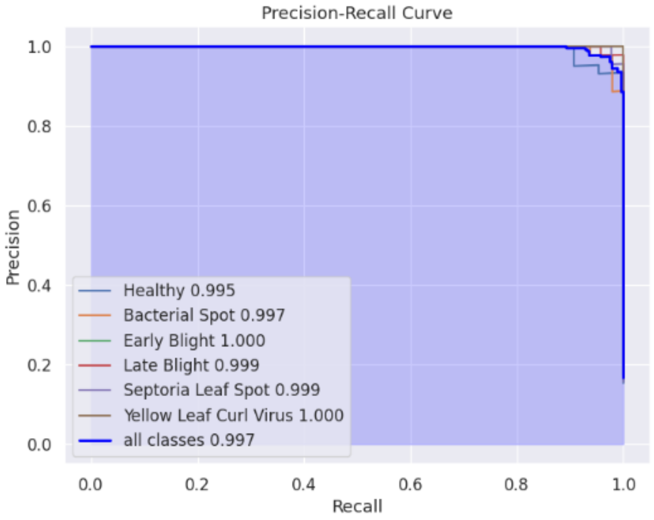

Fig. 5. Precision-Recall (PR) curve of ViT-r50-l32 model.

Fig. 6 illustrates the Multi-Class Receiver Operating Characteristic (ROC) curve for the ViT-r50-l32 model, which is the most effective model for classifying tomato leaf diseases. Each curve denotes one of the six disease categories: Healthy, Bacterial Spot, Early Blight, Late Blight, Septoria Leaf Spot, and Yellow Leaf Curl Virus. The values presented in the legend represent the area under the curve (AUC) for each category. The model attains an AUC score of 1.00 for each class and for the micro-average ROC curve, demonstrating optimal performance in differentiating among the various disease categories. The ROC curve indicates that the model can achieve a true positive rate (sensitivity) of 1.0 while keeping The false positive rate near 0 across all categories. The observed accuracy indicates that ViT-r50-l32 is a reliable model for classifying tomato leaf diseases, as evidenced by the absence of misclassifications in the ROC metrics. The diagonal dashed line indicates a random classifier (AUC = 0.5), while the model's ROC curves positioned significantly above this line demonstrate its robust predictive performance. The optimal AUC scores demonstrate the model's high accuracy and robustness, positioning it as an effective tool for the detection.

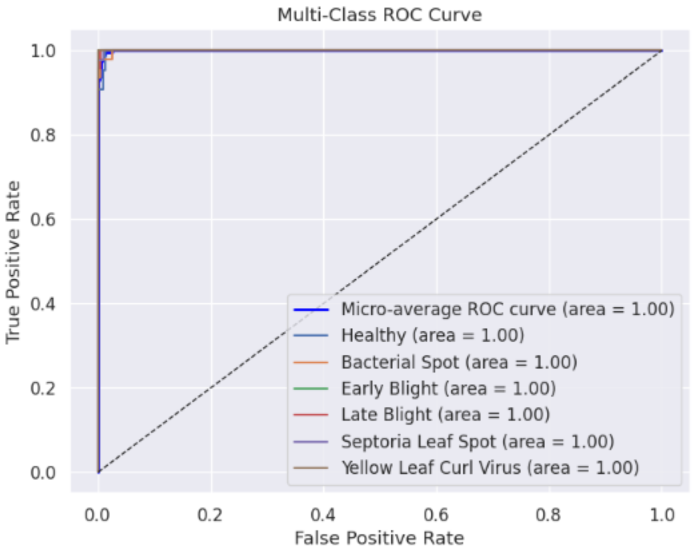

Fig. 6. ROC curve of ViT-r50-l32 model.

The confusion matrix for the ViT-r50-l32 tomato leaf disease classification model shows high true positive counts for each disease category, as depicted in Fig. 7. Bacterial Spot has 42 accurate categories and a few Healthy misclassifications. Early Blight is properly categorized 46 times, mostly in Late Blight. One Early Blight occurrence was misclassified, whereas Late Blight had 49 proper classifications. With 47 and 44 accurate classifications and no substantial misclassifications, Septoria Leaf Spot and Leaf Curl Virus perform well.

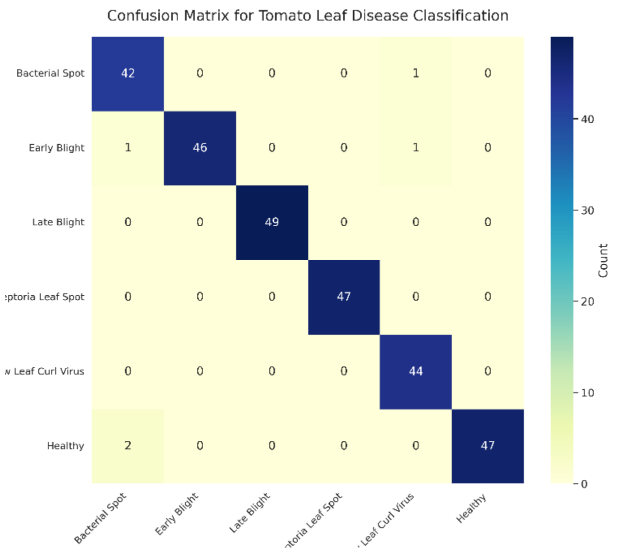

Fig. 7. Confusion matrix of ViT-r50-l32 model for differentiation of tomato leaf diseases.

Healthy is accurately labeled 47 times, with Bacterial Spot misclassified. Some classifications, such Healthy and Bacterial Spot, are somewhat confusing, suggesting model refining or hyperparameter tweaks might improve classification accuracy.

In Fig. 8, the ViT-r50-l32 model classifies six tomato leaf types from the Plant Village Dataset, including healthy and sick samples, using test images. The model accurately distinguishes Bacterial Spot, which has dark, irregular spots; Early Blight, which has concentric rings on yellowed areas; Healthy leaves, which are uniformly green and symptom-free; Late Blight, which has large, darkened lesions; Septoria Leaf Spot, which has small, circular lesions with light centers and dark edges; and Yellow Leaf Curl Virus, which shows curled, yellowed edges The model can distinguish these groups, suggesting its potential for early illness identification and treatment.

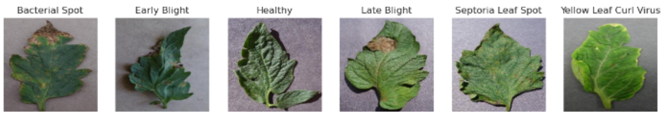

Fig. 8. Classification results of ViT-r50-l32 model.

C. Comparative Study

Table II presents a comparative analysis of tomato leaf disease classification models, emphasizing the performance metrics of various methodologies. The authors [26] used a CNN-based model (Inception-V3 and DenseNet-121), attaining an accuracy of 95.08%, with precision, recall, and F1-score metrics closely matched at 95.10%, 95.05%, and 95.07%, respectively. In addition, the authors in [27] introduced the TomFormer model, which amalgamates transformer-based architectures, achieving an accuracy of 87%, with somewhat reduced precision (87.50%), recall (86.50%), and F1-score (87.00%) relative to CNN-based models. Our proposed model, ViT-r50-l32, utilizes Vision Transformers to attain exceptional performance, achieving an accuracy of 98%, precision of 98.30%, recall of 98.33%, and an F1-score of 98.20%, thereby illustrating its efficacy and resilience in tomato leaf disease classification tasks.

TABLE II. Comparative Study with Related Approaches

Study	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
[26]	95.08	95.10	95.05	95.07
[27]	87	87.50	86.50	87.00
TLDViT Model	98	98.30	98.33	98.20

V. Conclusion

This paper presents TLDViT, a Vision Transformer model explicitly developed for classifying tomato leaf diseases using images from the Plant Village Dataset. TLDViT exhibited effective classification capabilities across six categories: Bacterial Spot, Early Blight, Healthy, Late Blight, Septoria Leaf Spot, and Yellow Leaf Curl Virus. In our assessments, we used two Vision Transformer models ViT-r50-l32 and ViT-l16-fe for comparative analysis. Among them, ViT-r50-l32 surpassed the other model, demonstrating enhanced accuracy and resilience across the illness categories. These findings underscore TLDViT's potential, in conjunction with ViT-r50-l32, for facilitating the early detection and control of crop diseases, which is essential for sustainable agriculture and food security. We propose the development of a mobile or field-deployable application for real-time disease diagnostics, facilitating the rapid identification of tomato leaf diseases by farmers and agronomists on-site, hence enabling prompt intervention and management.

Future work will optimize TLDViT for mobile and edge devices for real-time crop health monitoring in the field. We also want to combine this model into a mobile or field-deployable application for real-time disease diagnostics to help farmers and agronomists quickly identify tomato leaf diseases and control them. Other efforts include domain adaptation to improve model performance in varied environmental settings and adding new plant species and disease categories to the dataset.

Acknowledgment

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through project no. NBU-FFR-2024-2119-04

References

[1] J. L. Bargul and N. Ghanbari, "Detection of leaf diseases in tomato using machine learning approaches: A review," *International Journal of Plant Pathology*, vol. 12, no. 3, pp. 150–160, Sep. 2020.

[2] N. Ghanbari and A. R. Smith, "An analysis of disease patterns in tomato leaves using advanced imaging techniques," *Plant Disease Analysis*, vol. 45, no. 2, pp. 75–85, Feb. 2021.

[3] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, p. 1419, Sep. 2016.

[4] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Computers and Electronics in Agriculture*, vol. 145, pp. 311–318, Jan. 2018.

[5] Y. Li, X. Ma, Y. Qiao, and J. Shang, "Plant disease detection based on convolutional neural network," *Cluster Computing*, vol. 22, no. 2, pp. 2593–2602, Jun. 2019.

[6] A. D. S. Ferreira, D. M. Freitas, G. G. da Silva, H. Pistori, and M. T. Folhes, "Weed detection in soybean crops using convnets," *Computers and Electronics in Agriculture*, vol. 143, pp. 314–324, Oct. 2017.

[7] J. G. A. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease recognition," *Computers and Electronics in Agriculture*, vol. 153, pp. 46–53, Aug. 2018.

[8] H. Kim and J. Lee, "Vit-smartagri: Vision transformer and smartphone-based plant disease detection for smart agriculture," *Agronomy*, vol. 14, no. 2, p. 327, Feb. 2024.

[9] M. Ali, R. Khan, and D. Patel, "A multitask learning-based vision transformer for plant disease localization and classification," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 3, pp. 987–1001, Mar. 2024.

[10] R. Gupta, L. Singh, and P. Choudhury, "Plant disease detection using vision transformers on multispectral natural environment images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, p. 102205, Jan. 2024.

[11] K. Mehta and F. Alzahrani, "Early betel leaf disease detection using vision transformer and deep learning algorithms," *Journal of Ambient Intelligence and Humanized Computing*, vol. 15, no. 1, pp. 115–126, Feb. 2024.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, and e. a. T. Unterthiner, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, Oct. 2021.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, Aug. 2020, pp. 213–229.

[14] J. Chen, D. Liu, and Y. Zhang, "Application of vision transformers in agricultural disease detection: Case studies on rice and wheat," *Agricultural Informatics Journal*, vol. 7, no. 4, pp. 234–244, Apr. 2022.

[15] J. Ma, Z. Zhou, Y. Wu, and X. Zheng, "Deep convolutional neural networks for automatic detection of agricultural pests and diseases," *Computers and Electronics in Agriculture*, vol. 151, pp. 83–90, 2018.

[16] A. Singh, B. Ganapathysubramanian, A. Singh, and S. Sarkar, "Machine learning for high-throughput stress phenotyping in plants," *Trends in plant science*, vol. 23, no. 10, pp. 883–898, 2018.

[17] A. Fuentes, S. Yoon, and S. Kim, "Automated crop disease detection using deep learning: A review," *Computers and Electronics in Agriculture*, vol. 142, pp. 361–370, 2017.

[18] A. Mishra, S. Hossain, and A. Sadeghian, "Image processing techniques for detection of leaf disease," *Journal of Agricultural Research*, vol. 11, pp. 134–145, 2017.

[19] A. Rangarajan, R. Purushothaman, and A. Ramesh, "Diagnosis of plant leaf diseases using cnn-based features," *Journal of Image Processing*, vol. 32, pp. 123–135, 2018.

[20] M. Khan, S. Amin, and M. Bilal, "Transformers in computer vision: A survey for plant disease recognition," *Computer Vision Research*, vol. 15, pp. 231–249, 2022.

[21] W. Liu, J. Zhang, and Q. Wang, "Transformer-based architectures for image classification in agricultural disease detection," *Information Processing in Agriculture*, vol. 9, no. 3, pp. 412–423, 2022.

[22] X. Zhang and Y. Huang, "Plant disease recognition based on vision transformers: A case study of grapevine leaf diseases," *IEEE Access*, vol. 10, pp. 24 256–24 267, 2022.

[23] M. Jiang and W. Li, "Plant disease detection using vision transformers with transfer learning," *Agricultural Informatics*, vol. 8, pp. 87–101, 2023.

[24] C. Feng and M. Wu, "Edge computing for real-time plant disease detection using lightweight transformer models," *Computers and Electronics in Agriculture*, vol. 210, p. 108330, 2023.

[25] D. P. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," *arXiv preprint arXiv:1511.08060*, Nov. 2015.

[26] M. Yasin and N. Fatima, "Comparative performance evaluation of cnn models for tomato leaf disease classification," *arXiv preprint arXiv:2312.08659*, 2023.

[27] A. Khan and S. Ahmad, "Tomformer: A fusion model for early and accurate detection of tomato leaf diseases using transformers and cnns," *arXiv preprint arXiv:2312.16331*, 2023.