

AN INTELLIGENT ATTENTION-BASED DEEP LEARNING MODEL FOR TOMATO LEAF DISEASE RECOGNITION AND CLASSIFICATION

Swaira Zainab¹, Mehab Ishfaq², Abdul Ahad³, Sana Javed⁴, Ramsha Ahmad⁵,
Muhammad Ahsan Jamil^{6*}

^{1,2,4,5,6}Faculty of Computing, MNS-University of Agriculture, Multan, 60000, Pakistan.

³Department of Software Engineering, The Islamia University of Bahawalpur, 63100, Pakistan.

^{6*}ahsan.jamil@mnsuam.edu.pk

DOI: <https://doi.org/10.5281/zenodo.23387404>

Keywords

Tomato leaf classification; Vision Transformer; Self-attention; Plant disease recognition; Pest damage; Precision agriculture.

Article History

Received: 25 June 2026

Accepted: 16 September 2026

Published: 15 October 2026

Copyright @Author

Corresponding Author: *

Muhammad Ahsan Jamil

Abstract

Automated recognition of tomato leaf conditions can support crop monitoring by identifying visible symptoms from digital images. This paper presents a Vision Transformer (ViT) model and analyzes the eight-class evaluation documented in the supplied study. The evaluated categories comprise bacterial spot, early blight, late blight, leaf mold, Septoria leaf spot, two-spotted spider mite damage, tomato yellow leaf curl virus, and healthy leaves. The model combines image resizing and normalization, patch embeddings, positional encoding, transformer-based feature extraction, and a fully connected classification head. A balanced held-out test set contains 800 images, with 100 images per class. The source reports overall accuracy of 0.98 and macro-averaged precision, recall, and F1-score of 0.98. Class-level F1-scores range from 0.97 to 1.00. Early blight has the lowest recall (0.94), whereas Septoria leaf spot has the lowest precision (0.94). These findings indicate strong discrimination within the reported evaluation, while identifying the classes requiring further investigation. Average precision and feature-space visualization provide complementary evidence but do not establish calibration or field robustness. The analysis distinguishes documented results from proposed procedures and excludes unverified numerical comparisons with convolutional baselines. Missing dataset provenance, training configurations, prediction-level records, and external validation limit reproducibility and the strength of deployment claims. The model provides a promising basis for tomato leaf screening, subject to controlled replication and validation on independently collected field images.

INTRODUCTION

Tomato cultivation depends on maintaining healthy foliage throughout the growing season. Damage to leaf tissue can reduce the plant's capacity to support growth and fruit development, while delayed recognition of an adverse condition can complicate crop management. Visual symptoms provide an accessible source of information, but their

interpretation requires care: spotting, discoloration, necrosis, and deformation may occur across different diseases, pest injuries, and environmental stresses. An automated image classifier can help organize observations and identify cases requiring further assessment.

Traditional diagnosis draws on visual inspection, knowledge of crop history, and, when necessary,

laboratory confirmation. Image-based recognition addresses a different part of this process: it predicts a label from the visible appearance of a leaf. Such a prediction can support screening, but it does not by itself confirm a causal organism or determine the appropriate intervention. This distinction is particularly relevant when a dataset includes both infectious diseases and pest damage, as in the present eight-class evaluation.

Deep learning has enabled direct learning of discriminative features from labeled images. In plant disease recognition, convolutional neural networks (CNNs) provide a strong foundation Naseem et al. (2025) because they learn hierarchical representations of color, texture, and shape. Mohanty et al. (2016) demonstrated the potential of this approach using a large collection of labeled plant images. However, the same study also highlighted the difficulty of transferring performance from controlled images to images acquired under different conditions. High accuracy on a benchmark therefore requires a separate assessment of its relevance to field use.

The spatial distribution of symptoms motivates investigation of attention-based models. A leaf image may contain a localized lesion, widespread stippling, or broad morphological change. A Vision Transformer represents the image as a sequence of patch tokens and uses self-attention to exchange information between patches (Dosovitskiy et al., 2021). This architecture offers a mechanism for integrating spatially distributed visual evidence. Its suitability for a particular crop task nevertheless depends on data quality, model configuration, and evaluation design; the architecture alone does not guarantee robustness or superiority over CNNs.

The supplied study describes a ViT-based tomato leaf classifier and reports a balanced test evaluation across eight categories. Its introductory abstract mentions a broader ten-class scope, but its methodology and classification report specify eight evaluated classes. This paper follows the eight-class experimental record. The resulting task comprises six infectious disease categories, one pest-damage category, and one healthy reference category. References to disease classification are therefore

understood as a convenient description of the broader leaf-condition recognition task.

The objective is to present the documented model in a concise research article and critically analyze its reported performance. Three questions guide the analysis: how well does the classifier distinguish the eight evaluated categories; which class-level weaknesses remain despite high aggregate accuracy; and what evidence is still required for reproducibility and practical use? The contribution lies in consolidating the model description, class-wise results, and their implications while preserving the boundary between reported observations and future validation requirements.

LITERATURE REVIEW

Deep Learning for Plant Image Classification

Automated plant disease recognition has progressed from manually designed visual descriptors to learned image representations. Handcrafted pipelines typically separate feature extraction from classification, making their effectiveness dependent on the selected descriptors and image acquisition conditions. Deep networks instead learn a sequence of representations from the training data. This reduces the need to specify every useful lesion characteristic manually, although it introduces dependencies on annotation quality, sample diversity, and the integrity of the training and test partitions.

Mohanty et al. (2016) investigated CNN-based plant disease recognition and established a widely cited example of strong performance on a controlled image collection. Their work provides an important methodological lesson: success on familiar acquisition conditions does not establish comparable performance on images from independent environments Khan et al. (2024). The distinction between benchmark recognition and field generalization is directly relevant to interpreting the tomato leaf results reported here.

Too et al. (2019) compared fine-tuning approaches for plant disease identification. Their study supports treating transfer learning and architecture selection as empirical questions rather than assuming that a single network family is consistently best Qudsi et al. (2025). For the present task, established

convolutional architectures remain relevant candidates because lesion texture and color are useful signals. A valid architecture comparison requires common partitions, compatible preprocessing, and equivalent evaluation criteria. Scores copied from different experimental settings cannot isolate the effect of a backbone.

Vision Transformers and Spatial Context

Dosovitskiy et al. (2021) introduced an image classification architecture that processes a sequence of patch embeddings with a transformer. Positional embeddings preserve information about patch location, while self-attention allows tokens to combine information from other parts of the image. This offers a different representational structure from local convolution. In leaf analysis, it provides a plausible way to relate localized marks to their distribution across the lamina and to larger shape changes.

The potential value of global interaction should not be confused with a demonstrated causal explanation of performance. A high classification score does not establish that the network uses biologically meaningful regions, and a CNN can also obtain broader contextual information through its depth and receptive fields. The relevant research question is whether a particular transformer implementation improves discrimination under a controlled evaluation and whether any improvement persists outside the original dataset.

Touvron et al. (2021) examined data-efficient training of image transformers and attention-based distillation. Their work reinforces the importance of the training recipe when assessing transformers. For agricultural applications, pretrained representations and carefully specified augmentation may be useful, especially where annotated data are limited. In the supplied tomato study, however, the checkpoint, pretraining source, and fine-tuning configuration are not documented sufficiently to reconstruct the reported model. They remain unresolved details rather than confirmed experimental settings.

Interpretability, Visualization, and Confidence

Attention can be inspected to study interactions among tokens, but interpreting a visualization

requires additional evidence. Abnar and Zuidema (2020) proposed attention rollout and attention flow to account for information propagation across transformer layers. These methods provide tools for examining attention patterns; they do not make any individual attention image a complete explanation of the prediction. In a tomato classifier, explanations should be checked against lesion regions and tested for sensitivity to changes in irrelevant background content.

Feature visualization serves another purpose. The t-distributed stochastic neighbor embedding method projects high-dimensional representations into a lower-dimensional space for inspection (van der Maaten & Hinton, 2008). The supplied study describes separated clusters for several classes. Such a plot can help generate hypotheses about representation structure, but its appearance depends on projection settings Mudasar et al. (2025). It cannot replace a confusion matrix, a quantitative test, or evaluation on independent data.

Confidence also requires a distinct assessment. Guo et al. (2017) showed that neural network probability outputs can be poorly calibrated and examined temperature scaling as a calibration method. Consequently, a high SoftMax score is not automatically a reliable estimate of the chance that a tomato diagnosis is correct. A useful decision-support system should examine calibration on validation data and evaluate a review mechanism for uncertain cases.

Research Gap and Scope of the Present Study

The literature supports continued investigation of image-based crop screening, contextual representation learning, and interpretable outputs. It also indicates that accuracy, explanation quality, and deployment reliability are separate properties. The present study contributes an eight-class performance analysis, with emphasis on the difference between aggregate success and remaining class-level errors. It does not establish early-stage detection, treatment effectiveness, calibrated confidence, or generalization to farms not represented in the source evaluation.

METHODOLOGY

The study follows a quantitative supervised image-classification design. The inputs are tomato leaf images, and the output is one of eight mutually exclusive labels. This article consolidates the methodology and results documented in the supplied Word study; no new model training or independent experiment is claimed. Numerical results are transcribed from its classification report and average-precision table. Where the source describes a possible procedure without implementation details, the procedure is identified as proposed or incompletely specified.

Dataset and Class Taxonomy

The evaluated test set consists of 800 images, evenly distributed across eight classes. The source associates these labels with the tomato subset of PlantVillage-style repositories, but it does not provide an exact dataset release, download record, complete image inventory, or acquisition metadata. The broader PlantVillage literature supplies background for the task (Mohanty et al., 2016), rather than independent verification of the provenance of these particular test images.

The categories are bacterial spot, early blight, late blight, leaf mold, Septoria leaf spot, two-spotted spider mite damage, tomato yellow leaf curl virus, and healthy leaves. Spider mite damage is a pest condition rather than an infectious disease. Late blight is associated with an oomycete rather than a true fungus. These distinctions preserve the biological meaning of the labels without changing the classification task. Tomato mosaic virus and target spot are outside the documented eight-class evaluation.

Each test class contributes 100 images. Equal support makes class-wise comparisons straightforward, but it does not reproduce naturally occurring disease prevalence. The source describes separate training, validation, and test partitions, with the test set reserved for final evaluation. The training and validation counts, split ratios, grouping criteria, and random seed are not provided. Independence between partitions is therefore a stated design intention that cannot be audited from the available records.

Image Preparation and Data Integrity

Resizing and pixel normalization are the preprocessing operations explicitly described in the model pipeline. Resizing places images on a common input grid, while normalization controls their numeric scale. The source does not specify the final image dimensions or channel statistics. These values should be recovered from the implementation before attempting replication. Dataset-derived normalization statistics should be estimated from training images and then applied consistently to the other partitions.

The source discusses rotation, flipping, cropping, and controlled changes in brightness and contrast as augmentation strategies. Their exact probabilities, ranges, and implementation are absent. Accordingly, this paper treats augmentation as part of the described training approach without attributing the results to a particular policy. Augmented variants of a test image must not enter training, and transformation settings should preserve the visual evidence required for the label.

Although the original abstract refers to segmentation, the methodological discussion presents leaf-region cropping or segmentation conditionally. No implemented segmentation algorithm, mask set, or segmentation evaluation is supplied. The reported model is therefore described using confirmed resizing and normalization; background removal is a possible extension. Duplicate checking and label verification are also recommended in the source, but their completion cannot be established without image-level records.

Transformer Feature Extraction

Let an RGB image have height H , width W , and three channels. For square non-overlapping patches of side P , the number of patch tokens is: $N = (H/P)(W/P)$, provided H and W are divisible by P . (1)

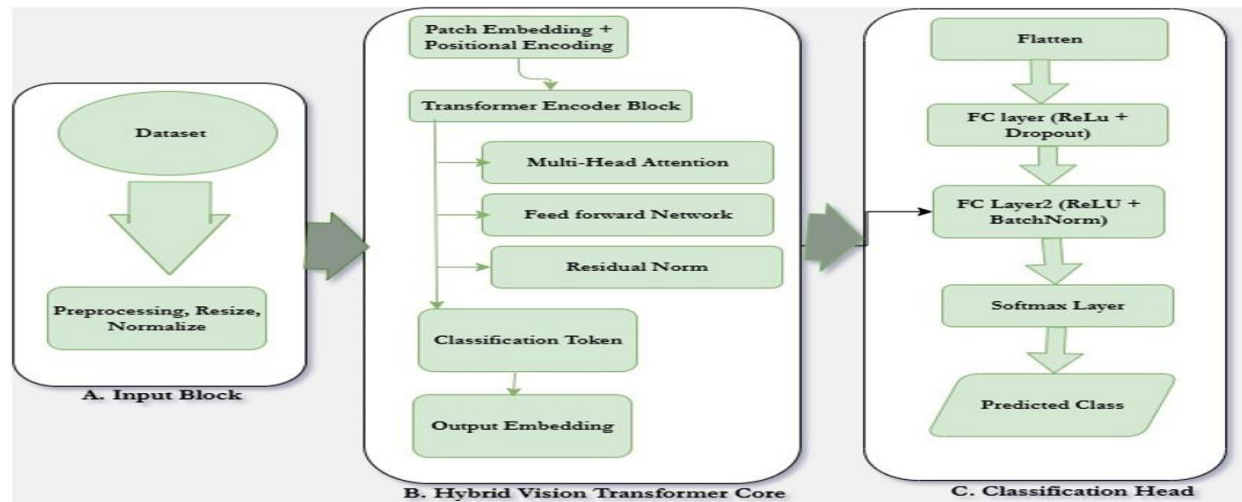
Each patch is flattened and projected into a D -dimensional embedding. A learnable classification token and positional embeddings are incorporated into the sequence. The representation is processed through transformer encoder blocks containing multi-head self-attention, feed-forward transformations, residual connections, and layer

normalization. This description follows the standard ViT formulation (Dosovitskiy et al., 2021) and the architecture reported in the source.

For query, key, and value matrices Q , K , and V , scaled dot-product attention is expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V. \quad (2)$$

Here, d_k is the key-vector dimension. Multiple attention heads compute complementary token interactions, and their outputs are combined. In the tomato task, this structure allows information from different image regions to contribute to a joint representation. Whether specific heads identify lesions, veins, or background cues remains an empirical question requiring attribution analysis.



The source describes extracting an output embedding and passing it through a fully connected classification head. Its stated components include ReLU activation, dropout, another fully connected transformation, batch normalization, and an eight-output SoftMax layer. Layer widths, dropout rates, the precise pooling or flattening operation, and the transformer depth are not specified. These missing values prevent an exact reconstruction of the network even though its main processing stages are identifiable.

Figure 1. Architecture of the Proposed Vision Transformer-Based Classification Model

Training Objective and Reproducibility
The source describes supervised optimization using a multiclass cross-entropy objective. For one-hot target components y_c and predicted probabilities p_c , the per-image objective is:

$$L = \sum_{c=1}^8 y_c \log(p_c). \quad (3)$$

The predicted label is the class with the largest output probability. This mathematical specification clarifies the classification objective, but does not identify the exact loss implementation or optimizer used in the reported run. Adam and AdamW, learning-rate scheduling, early stopping, and weight decay appear as options or recommendations in the source rather than a complete experimental configuration.

Table 1. Experimental information available for reproduction.

Item	Documented status
Target classes	Eight; listed in the test report
Test sample count	800; 100 per class
Input preparation	Resizing and normalization

Item	Documented status
Main architecture	ViT with fully connected head
Training/validation sizes	Not specified
Dataset release and split seed	Not specified
Image/patch size; depth; heads	Not specified
Pretraining checkpoint	Not specified
Optimizer; learning rate; epochs	Not specified
Hardware and software versions	Not specified

The absence of these settings is material to interpretation. Reproducing a score requires reproducing the dataset partition and training process, rather than selecting plausible values after observing the outcome. The present account therefore retains symbolic dimensions instead of inventing hyperparameters. Checkpoint files, configuration exports, and training logs would resolve these omissions more reliably than retrospective descriptions alone.

Evaluation Metrics

Overall accuracy is the fraction of correctly classified test images. Precision, recall, and F1-score are calculated per class using a one-versus-rest interpretation. Precision is $TP/(TP + FP)$, recall is $TP/(TP + FN)$, and F1 is $2PR/(P + R)$. Macro averages assign equal weight to every class; weighted averages use class support. Because all eight supports are equal, the macro and weighted averages are expected to coincide. Their equality is a consequence of the evaluation design and is not, by itself, evidence of uniform performance.

The source also reports average precision from precision-recall analysis and a qualitative t-SNE visualization. Average precision evaluates class-score ranking over changing thresholds, while the classification report evaluates the final predicted labels. The two measures therefore address related but different behavior. All published class metrics are rounded to two decimal places. Exact prediction arrays, score vectors, and a confusion matrix are unavailable, so this paper does not reconstruct exact error destinations or claim statistical significance.

ANALYSIS & DISCUSSION

The source reports overall accuracy of 0.98, together with macro and weighted precision, recall, and F1-score of 0.98. This indicates strong performance on the specified balanced test set. The result applies to those eight labels and their represented image conditions; it does not quantify performance on unrepresented conditions or independently collected farm images. Table 2 retains the original rounded class metrics, with shortened labels for readability.

Table 2. Reported eight-class test results. P = precision; R = recall; F1 = F1-score; n = support.

Class	P	R	F1	n
Bacterial spot	1.00	0.99	0.99	100
Early blight	1.00	0.94	0.97	100
Late blight	0.96	1.00	0.98	100

Class	P	R	F1	n
Leaf mold	1.00	0.96	0.98	100
Septoria leaf spot	0.94	0.99	0.97	100
Spider mite damage	0.99	0.97	0.98	100
Yellow leaf curl virus	1.00	1.00	1.00	100
Healthy	0.96	1.00	0.98	100
Macro average	0.98	0.98	0.98	800
Weighted average	0.98	0.98	0.98	800

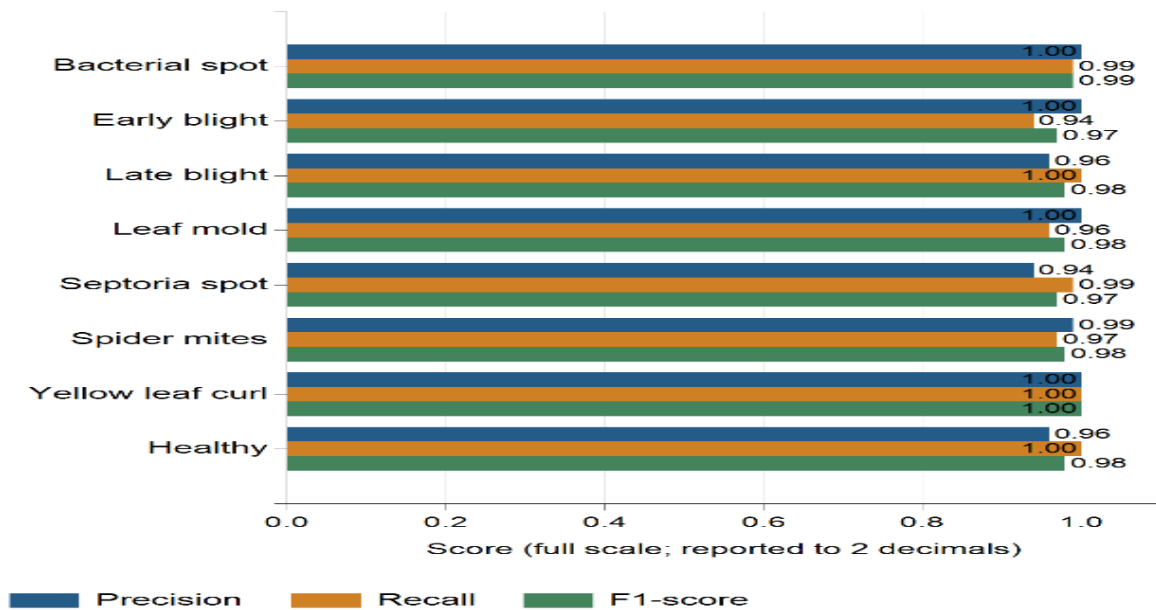


Figure 2. Class-wise precision, recall, and F1-score for the balanced eight-class test set (100 images per class). Bars use the full score scale; labels reproduce rounded source values. Early blight has the lowest recall and Septoria leaf spot the lowest precision.

Class-level F1-scores fall between 0.97 and 1.00, suggesting that no evaluated category exhibits a major collapse in predictive quality. Nevertheless, the precision–recall profiles differ. This matters because the costs of missing a symptomatic leaf and assigning the wrong disease label are not equivalent. The aggregate score compresses these distinctions

and should be read alongside the complete class report.

The displayed recalls sum to 7.85. With 100 images per class, their arithmetic suggests approximately 785 correct predictions among 800 images and approximately 15 errors. The resulting proportion, 0.98125, is consistent with an overall accuracy displayed as 0.98. This is a consistency calculation

from rounded metrics, not a verified count from prediction records. Exact counts should be obtained from the original outputs before performing confidence-interval or significance calculations.

Classes with the Strongest Results

Tomato yellow leaf curl virus has displayed precision, recall, and F1-score of 1.00. This means that the published report shows no class-level error at its stated precision. Its result is compatible with strong discrimination of the represented examples, potentially because broad color and shape characteristics distinguish them from some spot-based categories. That explanation remains a hypothesis: neither the metrics nor the architecture demonstrate which visual cues caused successful predictions.

Healthy leaves and late blight both show recall of 1.00, with precision of 0.96. Thus, high recall does not imply complete separation from other labels. The report indicates that some predictions assigned to these categories are incorrect, although their originating categories cannot be determined without the confusion matrix. In particular, the healthy class's lower precision prevents a claim that every leaf predicted healthy was truly healthy within the labeled test set.

Bacterial spot has precision of 1.00, recall of 0.99, and F1-score of 0.99. The displayed metrics indicate a small remaining sensitivity shortfall with highly reliable positive assignments in this evaluation. These results should still be distinguished from pathogen confirmation, as the experiment assesses agreement with image labels rather than an independent laboratory diagnostic standard.

Early Blight, Leaf Mold, and Pest Damage

Early blight has the lowest recall, 0.94, despite precision of 1.00. The model is therefore selective when assigning the early-blight label but misses a larger proportion of true early-blight examples than it does for other classes. On the reported support, the recall suggests about six missed examples. These cases account for the largest estimated contribution to the remaining errors and merit inspection before claiming a uniformly reliable screening system.

Leaf mold has recall of 0.96 and precision of 1.00. Its profile also suggests missed true examples despite highly precise positive predictions. Spider mite damage has precision of 0.99 and recall of 0.97, indicating both types of error at a small level. Since mite injury is included as a separate category, the task tests differentiation between pest damage and the other represented conditions, rather than exclusively identifying disease-causing infections.

Localized or visually overlapping patterns offer plausible explanations for the remaining errors, but the supplied records cannot establish specific confusions. Early blight may be misassigned to Septoria leaf spot, late blight, healthy, or another category; none of these possibilities is demonstrated by recall alone. A scientifically useful follow-up should review the actual misclassified images and their alternative scores, then distinguish label ambiguity, image degradation, and systematic model behavior.

Septoria Leaf Spot and False-Positive Assignments

Septoria leaf spot has the lowest precision, 0.94, and recall of 0.99. This differs from the early-blight profile: the classifier retrieves almost all labeled Septoria examples but also assigns the label to some non-Septoria images. Its F1-score of 0.97 reflects the combined effect. The pattern identifies a potential specificity problem at the class boundary, without revealing which competing categories supply the false positives.

The practical implication is that improvement should be class-specific. Increasing recall for early blight and improving precision for Septoria leaf spot may require different changes in data coverage or representation. A single global accuracy target can overlook this distinction. Additional training examples should be selected after inspecting the original errors, rather than assuming that every difficult class is affected by the same visual mechanism.

Precision–Recall Ranking and Feature Visualization

Table 3 summarizes the average-precision values reported in the source. Early blight has a displayed average precision of 0.99; all other classes have

displayed values of 1.00. These ranking results are compatible with high discriminative quality across thresholds. They do not imply error-free final

classification, and the rounded values do not establish that every underlying score ranking is perfect.

Table 3. Reported average precision (AP).

Class	AP
Bacterial spot	1.00
Early blight	0.99
Late blight	1.00
Leaf mold	1.00
Septoria leaf spot	1.00
Spider mite damage	1.00
Yellow leaf curl virus	1.00
Healthy	1.00

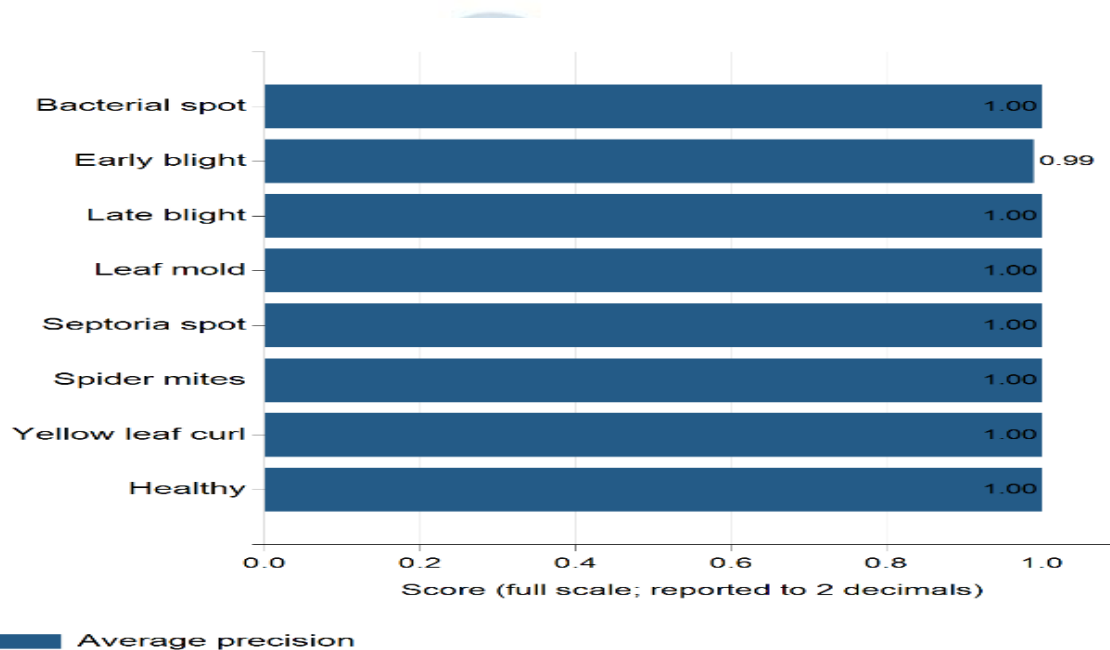


Figure 3. Average precision for the eight evaluated classes, using the rounded values reported in the source. A displayed score of 1.00 does not establish error-free multiclass predictions. These bars summarize AP values rather than reconstructing unavailable precision-recall curves.

A class may rank its positive examples highly yet lose some cases under multiclass argmax prediction because another class receives an even larger score. This helps explain how leaf mold can have displayed AP of 1.00 and recall of 0.96 in the classification report. The two results are not necessarily contradictory. Their consistency should nevertheless

be checked using the original class-score vectors and documented AP calculation settings.

The source describes compact t-SNE clusters for several categories and closer representations among some visually related conditions. This is qualitative support for learned class structure, rather than an independent measure of generalization. Without the underlying embeddings, projection seed, perplexity,

and preprocessing settings, the visual analysis cannot be reproduced. Nor can relative distances in a two-dimensional projection establish the exact direction of prediction errors.

Architecture Interpretation and Baseline Comparison

Self-attention provides a plausible mechanism for integrating distributed symptoms, but the available evidence does not isolate its contribution to the reported score. The classification head, data preparation, sample selection, and any pretraining could also influence performance. Ablation experiments would be required to determine the effect of specific components, including dropout, batch normalization, contextual attention, and leaf-region preparation.

The source includes numerical comparison entries for VGG16, MobileNetV2, ResNet50, and DenseNet121, but explicitly states that the added baseline values require validation against real training logs. Those entries cannot establish experimentally verified superiority. They are therefore excluded from the paper's numerical results. A defensible comparison should train or fine-tune each candidate on the same verified partitions and report performance variation across repeated runs, alongside parameter count, memory use, and inference time.

This distinction does not weaken the documented eight-class report; it limits the conclusion that can be drawn from it. The present findings support strong reported ViT performance on the selected test set. They do not establish that a transformer is more accurate, faster, or more suitable for deployment than all convolutional alternatives. Too et al. (2019) provide relevant background for comparative plant-classification experiments, but their results do not substitute for a matched comparison in this study.

Implications for Agricultural Screening

The classifier could form one component of a crop-monitoring workflow in which a user submits a leaf image and receives a likely condition label. In such a workflow, image quality, uncertainty, and applicability to the supported classes should guide whether a result is passed for review. A label

prediction alone does not provide lesion severity, treatment choice, or an assessment of mixed symptoms. These capabilities would require separate evidence and suitable data.

Field validation should include independent farms, seasons, cultivars, cameras, and capture sessions. It should also include common sources of uncertainty such as cluttered backgrounds, occlusion, blur, nutrient deficiency, and conditions outside the eight labels. Real prevalence should be preserved in at least one deployment-oriented test set. A balanced benchmark can reveal differences between classes, while a naturally distributed set examines performance under the workload that users may encounter.

Confidence calibration should be evaluated separately from accuracy. Temperature scaling is a possible validation-set procedure (Guo et al., 2017), but no calibration result is supplied here. Reliability diagrams and proper scoring measures would help assess whether high probabilities are justified. An uncertainty review rule should be selected using validation data and evaluated on held-out images rather than set after inspecting test errors.

Limitations and Priorities for Verification

The principal limitation is incomplete experimental documentation. Dataset provenance, training and validation sizes, model dimensions, optimization settings, and implementation versions are unavailable. These omissions prevent independent reconstruction and leave data leakage checks unresolved. The reported result should consequently be treated as a documented source finding pending reproducible verification, rather than a newly replicated experiment.

A second limitation is the absence of prediction-level evidence. Rounded metrics permit useful description, but not a full error audit. Original labels, predictions, probability vectors, and the confusion matrix are needed to verify counts, identify disease-pair errors, and assess ranking behavior. Repeated runs are needed to assess training variability; no confidence intervals or hypothesis tests are claimed in the present paper.

Third, the evaluation does not establish early-stage sensitivity. The source motivates early detection, but

supplies no disease-stage annotations or stage-specific results. Similarly, robustness to lighting or field backgrounds is not demonstrated by a single aggregate report. Stratified and external evaluations must precede those stronger claims. Attention visualization and feature clustering also require validation before being interpreted as biologically faithful explanations.

The immediate research priorities are to recover the dataset manifest and model configuration, audit partition independence, regenerate the classification report from saved predictions, and complete verified baseline runs. Subsequent work should emphasize error-informed data collection, external testing, calibration, and explanation checks. These steps would convert a promising reported benchmark result into a stronger body of evidence for practical crop monitoring.

CONCLUSION

This paper presents a Vision Transformer Model for eight-class tomato leaf condition recognition using the results documented in the supplied study. The evaluation covers six infectious disease classes, spider mite damage, and healthy leaves on a balanced test set of 800 images. The source reports accuracy and macro F1-score of 0.98, with class F1-scores from 0.97 to 1.00. These results suggest strong discrimination within the evaluated image collection.

The class-wise analysis identifies early-blight sensitivity and Septoria leaf spot precision as the clearest remaining weaknesses. It also shows why high average precision and aggregate accuracy do not eliminate prediction errors or establish calibrated confidence. Unverified CNN comparison values are excluded, and no superiority claim is made without controlled baseline experiments.

The model is a promising basis for further tomato screening research. Its practical reliability depends on recovering reproducible training details, verifying the reported predictions, and testing independently acquired field images. Early-stage detection, biologically meaningful explanations, and deployment suitability remain objectives for subsequent validation rather than demonstrated outcomes of the present evaluation.

REFERENCES

- Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4190–4197. <https://doi.org/10.18653/v1/2020.acl-main.385>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. <https://arxiv.org/abs/2010.11929>
- Naseem, M., Ahmad, B., Razzaq, A., Qadri, S., Ullah, S., Saeed, S., & Jamil, M. A. (2025). Advance assessment and counting of ripe cherry tomato's via YOLO model. *Journal of Computing & Biomedical Informatics*, 8(2).
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 70, 1321–1330. <https://proceedings.mlr.press/v70/guo17a.html>
- Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 1419. <https://doi.org/10.3389/fpls.2016.01419>
- Khan, H. Y., Kajla, N. I., Chohan, S. R., Saifullah, M., Jamil, M. A., & Umar, M. (2024). Machine learning-based forecasting of mango production under diverse agricultural conditions. *Journal of Nanoscope (JN)*, 5(1), 71–80. <https://doi.org/10.52700/jn.v5i1.102>
- Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, 161, 272–279. <https://doi.org/10.1016/j.compag.2018.03.032>

- Qudsi, R., Hussain, A., Qadri, S., Iqbal, N., Ullah, S., Jamil, M. A., & Hussain, A. (2025). Onion price forecasting using machine learning models market of South Punjab. *Statistics, Computing and Interdisciplinary Research*, 7(2), 177-191. <https://doi.org/10.52700/scir.v7i2.200>
- Mudasar, A. Q., Bukhari, F., Talib, R., Khurshid, I., Jamil, M. A., & Naqvi, K. (2025). Predicting stock exchange closing prices using machine learning techniques. *Statistics, Computing and Interdisciplinary Research*, 7(1), 1-25. <https://doi.org/10.52700/scir.v7i1.177>
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *Proceedings of the 38th International Conference on Machine Learning*, 139, 10347-10357. <https://proceedings.mlr.press/v139/touvron21a.html>
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579-2605. <https://jmlr.org/papers/v9/vandermaaten08a.html>
- Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1
- Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1492-1500. <https://doi.org/10.1109/CVPR.2017.634>