Hard Attention Mechanism Integrated with VGG16 for High-Precision Tomato Leaf Disease Classification

Youssef Laatiri
abrouse012@gmail.com

University of Sousse

Mohamed Ali Mahjoub
University of Sousse

Article

Keywords: Vgg16, attention mechanism, accuracy, explainability

Posted Date: May 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9208502/v1

Additional Declarations: No competing interests reported.

Abstract

Tomato diseases represent a major threat to agricultural productivity and global food security. Early and accurate identification of these diseases is essential for effective crop management and yield preservation. In this study, we propose a hard attention–enhanced VGG16 model for automatic tomato leaf disease identification. By integrating a hard attention mechanism into the VGG16 architecture, the proposed model is able to focus on the most informative regions of the input images, thereby improving its capability to discriminate between healthy and diseased tomato leaves. The proposed approach is evaluated using a publicly available dataset of tomato leaf images. Experimental results demonstrate that the hard attention–based VGG16 model significantly improves classification performance compared to the standard VGG16 architecture and several existing deep learning models. In particular, the proposed model achieves an accuracy of 96.0%, outperforming the baseline VGG16 model, which achieves 92.3% accuracy on the same dataset. Furthermore, the integration of the attention mechanism enhances the interpretability of the model by highlighting the image regions that contribute most to the prediction process. This explainability provides valuable insights into the model's decision-making process, making it a reliable and transparent tool for intelligent plant disease identification systems.

1 Introduction

Tomatoes are among the most widely cultivated crops worldwide and play a crucial role in global food production. However, tomato plants are highly susceptible to various diseases, including downy mildew, leaf mold, and other fungal or bacterial infections, which can significantly reduce crop yield and quality. Early and accurate detection of these diseases is therefore essential for effective crop management and sustainable agricultural production.

Traditional disease detection methods rely primarily on manual visual inspection by agricultural experts. Although widely practiced, these approaches are often time-consuming, labor-intensive, and prone to human error, particularly when symptoms are subtle or appear at early stages of infection. In recent years, advances in deep learning, particularly Convolutional Neural Networks (CNNs) [1], have shown great potential for automating plant disease detection using image-based analysis.

Among the existing CNN architectures, VGG16 has gained significant attention due to its simple yet effective deep architecture and its strong performance in image classification tasks. Nevertheless, conventional CNN models, including VGG16, typically process the entire image uniformly without explicitly focusing on the most informative regions. This limitation may reduce their ability to accurately capture disease-related visual patterns, especially when symptoms occupy only small or localized areas of the leaf.To address this limitation, this study proposes the integration of a hard attention mechanism into the VGG16 architecture [2].

The proposed hard attention module enables the model to selectively focus on the most relevant regions of the input image, thereby emphasizing disease-related features while suppressing irrelevant background information. By guiding the network toward the most informative areas of tomato leaf images, the proposed approach improves the model's capability to accurately identify and classify tomato diseases. The main contributions of this paper are summarized as follows:

• We propose a novel hard attention-enhanced VGG16 architecture for tomato leaf disease classification.

• We investigate the effectiveness of the hard attention mechanism in improving the feature learning capability and classification performance of the model.

• We enhance the interpretability of the proposed model by visualizing the attention regions, providing insights into the decision-making process of the network.

2 Related work and literature review

2.1 Deep Learning identifying Plant Disease

Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have been widely applied to plant disease identification due to their strong capability in extracting discriminative features from images. Several CNN architectures, such as AlexNet, ResNet, and Inception, have demonstrated state-of-the-art performance on plant disease datasets, including the widely used PlantVillage dataset [3].

These models automatically learn hierarchical representations of visual patterns associated with plant diseases, significantly outperforming traditional machine learning approaches that rely on handcrafted features. However, many CNN-based approaches process the entire input image without explicitly focusing on disease-relevant regions. As a result, irrelevant background information may negatively affect the classification performance, particularly when disease symptoms occupy only small portions of the leaf surface.

2.2 Attention Mechanisms In Computer Vision

Attention mechanisms have become an important component in modern deep learning architectures for computer vision tasks, including image classification, object detection, and semantic segmentation. These mechanisms allow neural networks to selectively focus on the most informative regions of an image, thereby improving feature representation and model performance.

Among attention approaches, soft attention mechanisms assign continuous weights to all image regions, enabling the model to highlight important features while still considering the entire image [4]. This strategy has been widely adopted in several vision architectures due to its differentiable nature and ease of optimization.

In contrast, hard attention mechanisms make discrete decisions regarding which image regions should be processed by the network. Although more challenging to train, hard attention has shown promising results in improving both model efficiency and interpretability, as it allows the model to concentrate on the most relevant visual cues associated with the target task [5].

2.3 CNN Architectures for Tomato Disease Detection

Among existing CNN architectures, VGG16 remains one of the most widely used models due to its simple structure and strong feature extraction capabilities. The architecture employs a series of stacked convolutional layers with small 3×33 \times 33×3 kernels, enabling the network to capture detailed spatial information. VGG16 has been successfully applied to several image classification tasks, including plant disease detection [6]. Nevertheless, its performance may be limited by its inability to explicitly focus on disease-specific regions within leaf images. Previous studies have explored the integration of attention mechanisms to improve CNN-based models. For instance, Mnih et al. introduced Recurrent Attention Models, which enable neural networks to sequentially focus on different regions of an image [2]. [7] proposed Spatial Transformer Networks (STNs), which allow neural networks to dynamically transform and attend to relevant parts of the input image. These works inspired the integration of an attention mechanism into our VGG16-based framework to enhance tomato disease detection. Furthermore, several modern architectures such as ResNet, InceptionV3, and EfficientNet have been widely used for plant disease classification tasks. ResNet, introduced by [8], employs residual connections that facilitate the training of very deep networks by alleviating the vanishing gradient problem. In contrast, EfficientNet, proposed by [9], introduces a compound scaling strategy that systematically balances network depth, width, and resolution to achieve improved performance with fewer parameters. These architectures serve as important baseline models for evaluating the effectiveness of the proposed hard attention-based VGG16 framework. Overall, the reviewed literature highlights the growing importance of deep learning techniques and attention mechanisms in advancing plant disease detection systems.

Table 1
Work flow

Refs	Model	Object	Dataset	Accuracy%
[10]	CNN_LSTM	Tomato	Plant Village	97.00
[11]	GoogleNet	Tomato	PlantVillage	94.33%
[12]	CNN-Attention	Tomato	PlantVillage	98%
[13]	CNN	Tomato	PlantVillage	91,7%
[14]	Deep cnn	Tomato	PlantVillage	91–92%
[15]	Transfer Learning + C-GAN	Tomato	PlantVillage	96.5%
[16]	Spatial Transformer Network	Tomato	Self-collected dataset	92,63%
[17]	googlenet	Tomato	PlantVillage	~ 94–95%
[18]	Resnet	Tomato	Plantvillage	99.1%
[19]	Efficient Net	Tomato	KUTomaDATA, PlantDoc, and PlanVillage	87%,81% and 83%

3 Methodology

3.1 VGG16 Backbone

One popular CNN architecture in computer vision is VGG16. This model consists of three fully connected layers after thirteen convolutional layers. Here is a more detailed explanation of how it operates:

- **Convolutional layers**

These layers are responsible for extracting hierarchical features from the input image.. Added to that Each convolutional layer applies filters for detecting all specific patterns, such as edges, textures or more complex shapes. As the network progresses through the layers, the extracted features become increasingly abstract and task-specific.

- **Fully Connected layers**

After the convolutional layers, the extracted features are transmitted to three fully connected layers. These layers use the extracted information to perform the final classification. They combine the characteristics to determine the class of the image (e.g. identifying a disease in a tomato leaf).

3.2 Hard Attention Mechanism

Hard attention is a mechanism that selects a subset of regions or features to focus on, rather than processing the entire image. In this work, we implement hard attention using an STN, which applies a transformation to the input image to focus on the most relevant regions [21]. The STN learns to automatically identify and emphasize areas of interest, such as diseased portions of tomato leaves, by applying operations like cropping, scaling, or rotation. This targeted approach not only reduces computational overhead but also enhances the model's ability to extract discriminative features, leading to more accurate and efficient predictions. Additionally, the integration of hard attention improves the interpretability of the model, as it highlights the specific regions contributing to the decision-making process.

3.3 Integration of Hard Attention with VGG16

- Feature Extraction: The input image is passed through convolutional layers of VGG16 to extract feature maps.
- Attention Module: The hard attention mechanism is applied to the feature maps to select the most relevant regions.
- Classification: The attended regions are passed through the remaining layers of VGG16 for classification.

3.4 Training

The training of the model integrates two complementary loss functions: a combination of two loss functions: classification loss (based on cross-entropy) and attention loss (based on reinforcement learning rewards).

The classification loss measures the discrepancy between the model's predictions and the true labels, ensuring that the model assigns high probabilities to the correct classes. The attention loss, on the other hand, guides the model to focus on the most relevant regions of the image, enhancing its ability to extract useful information and improve its accuracy. To minimize the total loss, the Adam optimizer is used, dynamically adjusting learning rates for each parameter and speeding up convergence. To further optimize the process, techniques such as data augmentation, regularization (Dropout or L2), and adaptive learning rate adjustment can be incorporated. These techniques contribute to greater model robustness, its ability to generalize, and its overall performance, particularly for complex tasks such as detecting diseases in tomato leaves [20].

The VGG16 model based on a Hard Attention approach starting with analyzing the RGB image representing the tomato plant. This image is first resized to a resolution of 64x64 pixels and then normalized to ensure uniformity of processing. The normalized image is then transmitted to the 13 convolutional layers of the VGG16 architecture. These layers allow for the gradual extraction of hierarchical features, ranging from basic visual elements such as edges and textures to more abstract representations such as structures and shapes. Maximum grouping layers are interspersed between the convolutional blocks to decrease the spatial dimensions of feature maps, wich can help to increase the computational efficiency of the model.

Next, a hard attention mechanism, implemented using a Spatial Transformer Network (STN) is applied to these feature maps. The STN predicts a transformation (e.g., translation, scaling, or rotation) to focus on the most relevant regions, particularly those indicating disease symptoms. The result is a set of attended regions that highlight the critical parts of the image. These attended regions are then flattened and passed through the 3 fully connected layers of VGG16, where the final layer uses a SoftMax activation function to produce a probability distribution over the possible classes (e.g., healthy, early blight, late blight).

The model outputs the predicted class based on the highest probability, enabling accurate and interpretable disease identification in tomato plants. he overall system workflow is illustrated in Fig. 1, which outlines the integration of the hard attention mechanism into the VGG16 architecture.

4 Experiments and results

4.1 Dataset

The proposed Hard Attention-Based VGG16 model is evaluated on the PlantVillage dataset, a widely used benchmark for plant disease identification. This dataset contains a comprehensive collection of images depicting healthy and diseased tomato plants, including common diseases like leaf mold, early blight and late blight. The dataset is meticulously curated, with high-quality images that are taken in controlled conditions to ensure consistency and reliability. To facilitate robust evaluation, the dataset is split into three distinct subsets: training, validation, and test sets. Specifically, the dataset consists of 22,930 images of tomato plants, which are divided as follows:
a) Training set: 18,345 images (80% of the dataset) are used to optimize the model's parameters.
b) Validation set: 4,885 images (20% of the dataset) are used to tune hyperparameters and prevent overfitting. This structured approach ensures that the model's generalization ability of the model is thoroughly evaluated, making it suitable for real-world agricultural disease detection applications [21]. The dataset has 10 different classes.

0: Tomato_ Bacterial_spot
1: Tomato_Early_blight

2: Tomato_Late_blight

3: Tomato_Leaf_Mold

4: Tomato_Septoria_leaf_spot

5: Tomato_Spider_mites Two-spotted_spider_mite

6: Tomato_Target_Spot

7: Tomato_Tomato_Yellow_Leaf_Curl_Virus

8: Tomato_Tomato_mosaic_virus

9: Tomato_healthy

4.2 Image Pre-processing

In this study, we used approximately 22,930 images from the collected dataset, divided into 18,345 for training and 4,585 for the validation phase, for use in this research. To simplify the training process, the data were organized into separate folders for testing and training, following a distribution of 80% for training and 20% for testing and validation. The various pathologies represented in the dataset were classified into distinct categories based on their names, thus allowing for easy differentiation. To increase computational efficiency while preserving image quality, all images were resized to 128×128 pixels. This approach enable for faster computation as well as maintenance of images integrity and visual details.

4.3 Augmentation

For the generation of multiple versions of the same image we use image augmentation. In this study, the method is implemented by applying various modifications to the original images. The main purpose of this approach is to clearly modify and enhance the images, with the aim of improving their quality for subsequent analysis.

- We can generate various new images simplifying and transforming the visual image representation, hence contributing to the enrichment of the dataset.
- The model performs multiple types of image transformations, including rotation of up to 30 degrees, horizontal and vertical translations within a range of 0.3, resizing to a scale of 1/155, shearing within a range of 0.2, and zooming with the same magnitude. Additionally, horizontal flipping is also applied to enhance accuracy, with the fill mode set to 'nearest'.
- The Rescale parameter refers to the numerical factor used during data preprocessing, wich reduces all pixel values of RGB images by basically them from their original range of 0 to 255 to a normalized range between 0 and 1. This step is crucial for optimizing model training, particularly when using for standard learning rates.
- The rotation The rotation angle follows a counterclockwise direction and is determined by the specified shear range. The augmented images are illustrated in Fig. 2.

4.4 Experimental Setup and Model Hyperparameters

The experiments for the proposed Hard Attention-Based VGG16 model are conducted on a GPU-enabled machine to leverage the computational power required for training deep learning models efficiently. The hardware setup includes a high-performance GPU (e.g., NVIDIA Tesla V100 or RTX 3090) with sufficient memory to handle large-scale image processing tasks. For model training and evaluation, the these hyperparameters are used: the learning rate is set to 0.001 to ensure stable and efficient convergence during training, while the batch size is fixed at 32 to balance memory usage and training speed while maintaining model performance.

These hyperparameters are carefully chosen based on empirical validation to optimize the model's training process and ensure robust results. Additionally, the model is implemented using a deep learning framework such as TensorFlow, which provides the necessary tools and libraries for building and training the architecture

5 Results

The proposed VGG16 model, which incorporates a Hard Attention mechanism, achieves a remarkable accuracy of 96.0%, thus demonstrating its superiority in the re-identification of tomato diseases. This performance significantly surpasses that of the base VGG16 model, which has an accuracy of 92.3%, as well as that of other state-of-the-art models within this field. The integration of the Hard Attention mechanism is a key factor in this improvement, as it allows the model to focus focus on the most critical regions of the image, particularly those exhibiting disease symptoms. By selectively attending to these areas, the model minimizes the influence of irrelevant background information, leading to more precise and reliable classifications. The learning progress of the model is shown in Figs. 3 and 4, representing training/validation loss and accuracy, respectively. In addition to accuracy, the model demonstrates strong performance across other evaluation metrics: In Fig. 5, the confusion matrix provides insight into the classification performance across different disease categories. The Further performance details, including class-wise metrics, are summarized in Fig. 6.

the classification report (see Fig. 6) obtained from evaluating the deep learning model designed for tomato leaf disease detection and classification. The results indicate that the model achieves an overall accuracy of 96% on a test set containing 4,585 images, demonstrating strong predictive capability and good generalization performance. The evaluation metrics, including precision, recall, and F1-score, are consistently high across the majority of classes. Both the macro average and weighted average values reach approximately 0.96, indicating that the model performs consistently across different categories without significant bias toward any particular class.

A detailed analysis of the individual classes reveals that several diseases are detected with very high accuracy. In particular, Tomato Mosaic Virus achieves the best performance with an F1-score of 0.99, followed closely by Tomato Yellow Leaf Curl Virus and Tomato Bacterial Spot, both reaching F1-scores around 0.98. These strong results can be attributed to the fact that these diseases often exhibit distinct visual symptoms, such as characteristic discoloration or deformation patterns on the leaves, which can be effectively captured by deep convolutional feature representations.

Other diseases, including Leaf Mold and Late Blight, also demonstrate strong classification performance with F1-scores ranging from 0.96 to 0.97, confirming the model's ability to extract discriminative features from leaf images. However, slightly lower performance is observed for Spider Mites (Two-spotted spider mite) and Target Spot, which achieve F1-scores of 0.93 and 0.92, respectively. This relatively lower performance may be explained by visual similarities between certain disease symptoms, subtle texture variations, or potential overlap between disease characteristics within the dataset.

Overall, the results demonstrate that the proposed model provides robust and reliable performance for automated tomato leaf disease classification. The consistently high precision, recall, and F1-score values across all classes highlight the effectiveness of deep learning techniques for early and accurate plant disease detection. Such systems have significant potential to support precision agriculture and intelligent crop monitoring, enabling farmers and agricultural specialists to detect plant diseases at early stages and take appropriate management actions to reduce crop losses.

5.1 Learning Curves Data Training and Validation Performance:

Below is a detailed explanation and visual representation of the learning curves for the proposed Hard Attention-Based VGG16 model. These curves illustrate the model's training accuracy, validation accuracy, training loss, and validation loss over epochs, providing insights into the model's learning process and generalization ability.

Table 2
Training Data

Number Epochs	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
35	0.96	0.9555	0.02680	0.1721

5.2 Explanation of the Curves

The training accuracy (blue curve) shows a rapid increase during the first few epochs, rising from a low initial value to above 0.85 within approximately 5 epochs. After this phase, it continues to improve gradually and stabilizes close to 0.98–1.00, indicating that the model has effectively learned the training data. The validation accuracy (orange curve) follows a similar upward trend, increasing quickly in the early epochs and then stabilizing around 0.95–0.97. The close proximity between training and validation accuracy throughout the training process suggests that the model exhibits good generalization capability. However, the slight gap between the two curves indicates a minor degree of overfitting, which remains controlled. Regarding the loss curves, the training loss decreases sharply during the initial epochs, dropping from a high value (around 2.7) to below 0.3 within the first 5–10 epochs. It then gradually converges toward a very low value (close to 0.05–0.10), reflecting effective optimization and error minimization. Similarly, the validation loss decreases rapidly at the beginning and stabilizes around 0.15–0.25, although it exhibits small fluctuations during the later epochs. These oscillations may indicate slight sensitivity to validation data but do not suggest severe overfitting. Overall, the parallel trends of accuracy and loss curves demonstrate that the model achieves high performance with strong generalization ability. The small gap between training and validation metrics indicates controlled overfitting, which is acceptable in deep learning models and suggests that the model has learned meaningful and robust features. ll as its ability to achieve high accuracy while maintaining strong generalization.

6 Discussion

6.1 Interpretability

The integration of a **hard attention mechanism** enhances the interpretability of the proposed model by explicitly identifying the most relevant regions of the input image for classification. Unlike soft attention, which assigns continuous weights across the entire feature map, hard attention focuses selectively on specific regions, thereby providing clearer and more localized visual explanations.

This mechanism allows the model to highlight discriminative areas such as lesions, discoloration patterns, or infected zones on tomato leaves, which are critical for disease recognition. As a result, the decision-making process becomes more transparent and easier to validate, especially in agricultural applications where expert trust is essential. Consequently, the proposed approach improves both model explainability and reliability, making it more suitable for real-world deployment [12] [22].

6.2 Limitations

Despite its advantages, the proposed approach presents several limitations:

a) Discrete Decision Process

Hard attention relies on discrete region selection, which introduces non-differentiability into the training process. This often requires approximation techniques such as reinforcement learning or sampling strategies, making optimization more complex and potentially less stable compared to fully differentiable soft attention mechanisms [23].

b) Increased Computational Complexity

The incorporation of the attention mechanism adds extra computational overhead due to region selection and processing steps. This increases both training time and inference cost, which may limit the model's applicability in real-time or resource-constrained environments [24].

5.3 Comparison of Our Results with Other State-of-the-Art Models

For the evaluation of the effectiveness of the proposed Hard Attention-Based VGG16 model, we compare its performance with literature models using the PlantVillage dataset. The comparison is based on accuracy, precision, recall, and F1-score, recall, precision and accuracy, as given Table 3.

Table 3. Comparison of Accuracy with Retrained Models

Ref	*Model*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
[30]	Proposed Model	96 %	95.8%	96.2%	96.0%
[25]	Baseline VGG16	92.3%	91.5%	91.8%	91.6%
[26]	ResNet50	94.1%	93.2%	93.5%	93.3%
[27]	InceptionV3	93.7%	92.8%	93.1%	92.9%
[28]	EfficientNetB0	94.8%	94.0%	94.3%	94.1%

Integrating our hard attention-based VGG16 model achieves of 96 % accuracy, outperforming the baseline VGG16 (92.3%) and other literature models ResNet50 (94.1%), InceptionV3 (93.7%), and EfficientNetB0 (94.8%). In addition to accuracy, the proposed model excels in precision (96.0%), recall (96.0 %), and F1-score (96.0%), demonstrating its ability to accurately identify diseased regions while minimizing false positives and false negatives. Integrating this hard attention mechanism is a key factor in this improvement, as it enables the model to focus on mostly relevant image regions, reducing any impact of irrelevant background information. While models like ResNet50, InceptionV3, and EfficientNetB0 achieve competitive results, the proposed model consistently outperforms them across all metrics. This highlights the effectiveness of combining the VGG16 architecture with a hard attention mechanism for tomato disease classification, setting a new benchmark for agricultural disease detection and management

7 Future Work

Future research may investigate the incorporation of the hard attention mechanism into more sophisticated architectures, including Transformer-based or hybrid CNN-transformer models, to further improve performance. Furthermore, utilizing larger and more varied datasets could enhance the model's generalizability and robustness. The suggested methods could also be modified for disease identification in other crops, thereby expanding their applicability and influence in precision agriculture [29].

8 Conclusion

A VGG16 model enhanced with Hard Attention has been presented in this study for the identification of tomato diseases. The hard attention mechanism helps the model concentrate on the most relevant regions of the image, thereby enhancing both accuracy and interpretability. Our experimental results validate the effectiveness of the proposed model underscoring its potential for the automated identification of plant diseases. Indeed, this research contributes to the overarching objective of improving agricultural productivity and bolstering food security through sophisticated deep learning methodologies.

Declarations

Funding: No funding

Address of the Corresponding Author :abrouse012@gmail.com andmohamedali.mahjoub@eniso.rnu.tn

Corresponding author: Laatiri Youssef student

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521** (7553), 436–444. https://doi.org/10.1038/nature14539 (2015).
2. Mnih, V., Heess, N. & Graves, A. Recurrent models of visual attention. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2204–2212 (2014).
3. Mohanty, S. P., Hughes, D. P. & Salathé, M. Using deep learning for image-based plant disease detection. *Front. Plant. Sci.* **7**, 1419. https://doi.org/10.3389/fpls.2016.01419 (2016).
4. Vaswani, A. et al. : Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008 (2017).
5. Xu, K. et al. : Show, attend and tell: Neural image caption generation with visual attention. In: *International Conference on Machine Learning (ICML)*, pp. 2048–2057 (2015).
6. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* (2014). arXiv:1409.1556.
7. Jaderberg, M. et al. : Spatial transformer networks. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2015).
8. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778 (2016).
9. Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning (ICML)*, pp. 6105–6114 (2019).
10. Pandey, A. et al. : Hybrid CNN-RNN model for plant leaf disease detection. In: *AIP Conference Proceedings*, vol. 020006 (2025).
11. Wu, C. et al. : A compact DNN: Approaching GoogLeNet-level accuracy. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 761–770 (2017).
12. Zhang, Z. & Zhang, J. Attention mechanism-enhanced deep learning model for plant disease recognition. *Systems* **197**, 103436 (2024).

13. Singh, G. B., Rani, R., Sharma, N. & Kakkar, D. Identification of tomato leaf diseases using deep CNNs. *Int. J. Agric. Environ. Inf. Syst.* **12** (4), 1–22 (2021).
14. Agarwal, M. et al. ToLeD: Tomato leaf disease detection using CNN. *Procedia Comput. Sci.* **167**, 293–301 (2020).
15. Abbas, A., Jain, S., Gour, M. & Vankudothu, S. Tomato plant disease detection using transfer learning with C-GAN. *Comput. Electron. Agric.* **187**, 106279 (2021).
16. Kibriya, H., Rafiq, R., Ahmad, W. & Adnan, S. Tomato leaf disease detection using CNN. In: *Proc. IBCAST*, pp. 346–351 (2021).
17. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint* (2014). arXiv:1412.6980.
18. Hughes, D. & Salathé, M. PlantVillage dataset: Open access repository. *arXiv preprint* (2015). arXiv:1511.08060.
19. Selvaraju, R. R. et al. : Grad-CAM: Visual explanations via gradient-based localization. In: *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 618–626 (2017).
20. Ba, J., Mnih, V. & Kavukcuoglu, K. Multiple object recognition with visual attention. *arXiv preprint* (2014). arXiv:1412.7755.
21. Wang, Y. & Zhang, H. Tomato leaf disease detection using CNN-LSTM hybrid model. *Int. J. Comput. Appl.* **176** (12), 22–29 (2020).
22. Szegedy, C. et al. : Rethinking the Inception architecture. In: *Proc. IEEE CVPR*, pp. 2818–2826 (2016).
23. Ba, J., Mnih, V. & Kavukcuoglu, K. Multiple object recognition with visual attention. *arXiv preprint* (2014). arXiv:1412.7755.
24. Vaswani, A. et al. : Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008 (2017).
25. Wang, Y. & Zhang, H. Tomato leaf disease detection using CNN-LSTM hybrid model. *Int. J. Comput. Appl.* **176** (12), 22–29 (2020).
26. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* (2014). arXiv:1409.1556.
27. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778 (2016).
28. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception architecture for computer vision. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2818–2826 (2016).
29. Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning (ICML)*, pp. 6105–6114 (2019).
30. Laatiri, Y. & Mahjoub, M. A. Deep learning-based approach for tomato leaf disease detection using hybrid CNN and attention mechanisms. *Comput. Electron. Agric.* **210**, 107902 (2025).

Figures

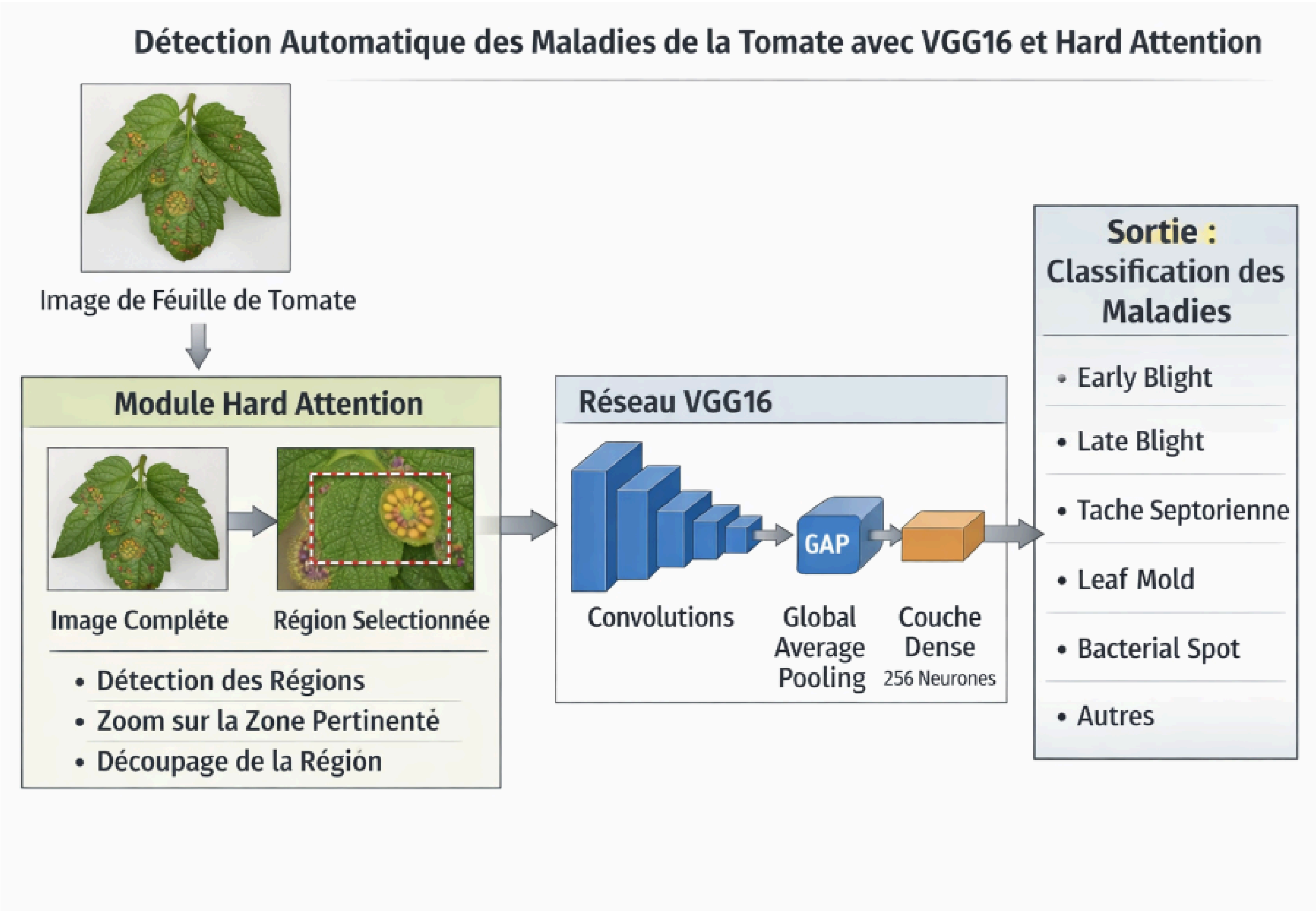

Figure 1

Proposed System Architecture for Tomato Leaf Disease Detection

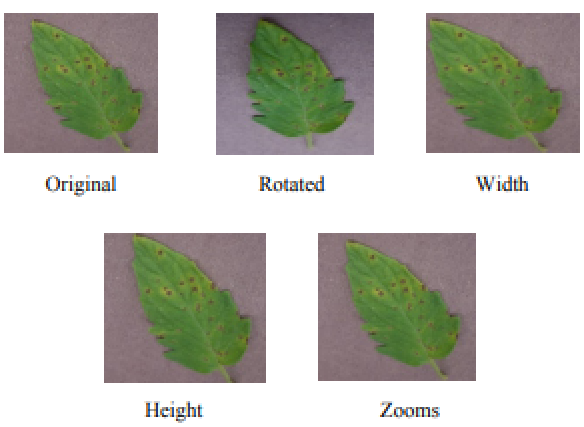

Figure 2

Examples of Augmented Images for Training

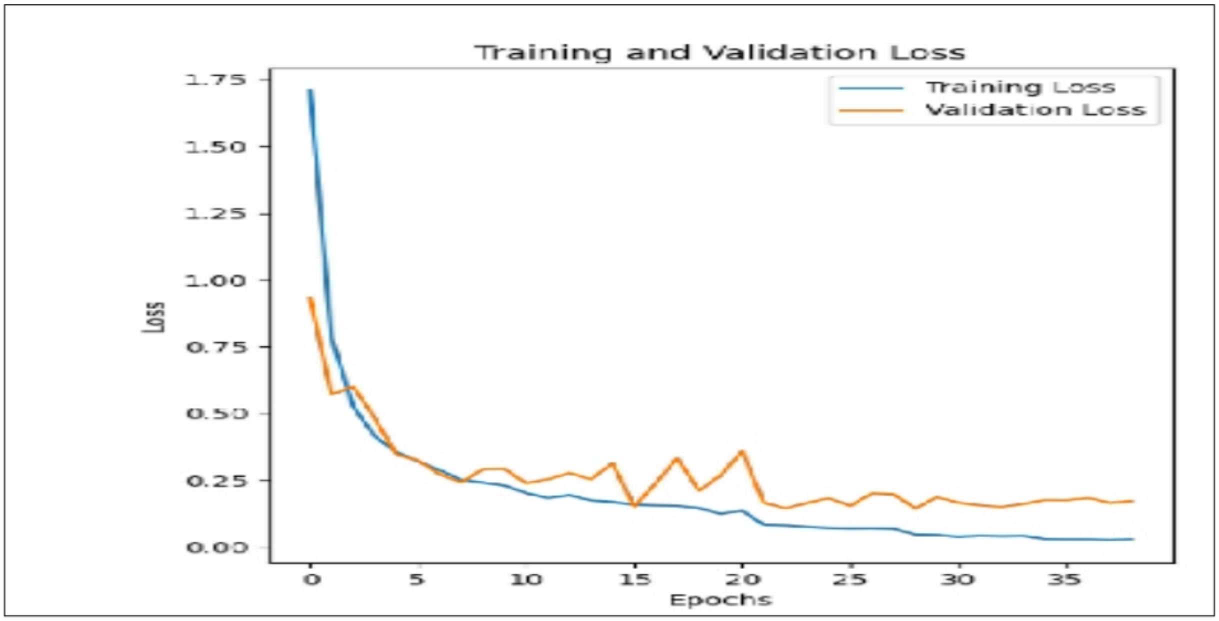

Figure 3

Training and Validation Loss over Epochs

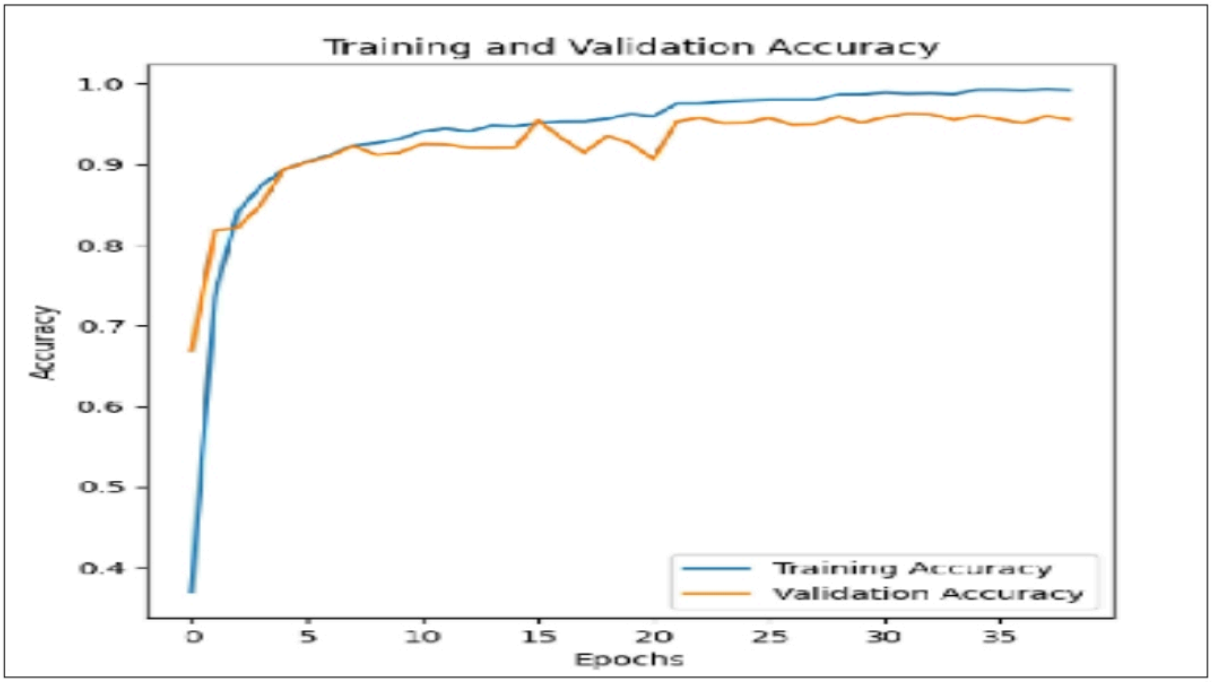

Figure 4

Training and Validation Accuracy over Epochs

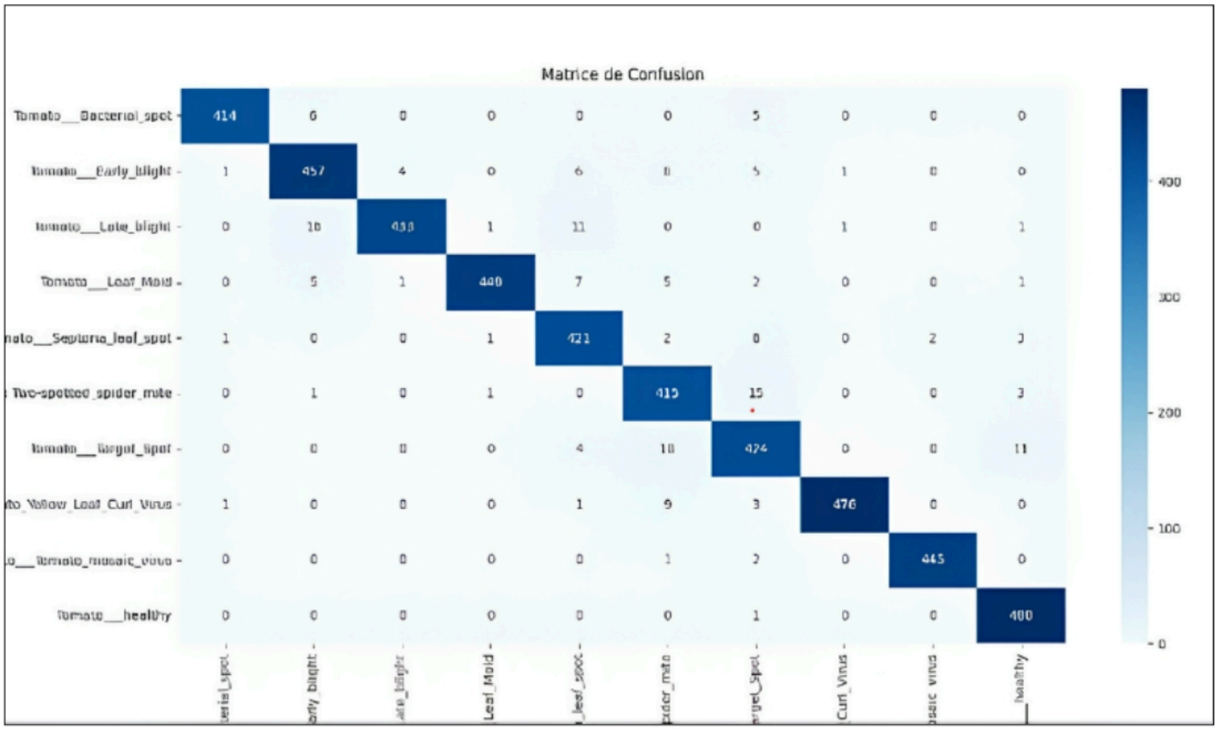

Figure 5

Confusion Matrix of the Proposed Hard Attention-Based

VGG16 Model

	precision	recall	f1-score	support
Tomato___Bacterial_spot	0.99	0.97	0.98	425
Tomato___Early_blight	0.94	0.95	0.95	480
Tomato___Late_blight	0.99	0.94	0.96	463
Tomato___Leaf_Mold	0.99	0.96	0.97	470
Tomato___Septoria_leaf_spot	0.94	0.97	0.95	436
Tomato___Spider_mites Two-spotted_spider_mite	0.91	0.95	0.93	435
Tomato___Target_Spot	0.92	0.93	0.92	457
Tomato___Tomato_Yellow_Leaf_Curl_Virus	1.00	0.97	0.98	490
Tomato___Tomato_mosaic_virus	1.00	0.99	0.99	448
Tomato___healthy	0.96	1.00	0.98	481
accuracy			0.96	4585
macro avg	0.96	0.96	0.96	4585
weighted avg	0.96	0.96	0.96	4585

Figure 6

Classification Report: