International Journal of Economics, Business and Innovation Research
https://e-journal.citakonsultindo.or.id/index.php/IJEBIR
E-ISSN : 2964-0865
IJEBIR

Implementation Of A Hybrid Cnn-Transformer Model For Tomato Leaf Disease Detection Under Background Noise

Mukti Adi Azhari[1], I Made Artha Agastya[2]

[1,2]Department of Informatics, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia

Email : m.adi.azhari@students.amikom.ac.id

Abstract

Tomato leaf diseases can significantly reduce crop productivity, while automated disease detection remains challenging when images are captured under complex field conditions with background noise. This study implements a Hybrid CNN-Transformer model for tomato leaf disease detection under background noise using the tomato subset of the PlantDoc dataset. The proposed architecture integrates MobileNetV3-Large for local feature extraction, a Feature Fusion Bridge for feature transformation, and a Vision Transformer encoder for global contextual representation. The dataset consists of 685 original images representing eight tomato leaf conditions and is divided into training, validation, and testing sets using an 80:10:10 stratified split. The proposed model is compared with pure MobileNetV3, ViT-B/16, and an EfficientNet-B0-based hybrid model. Experimental results show that the proposed Hybrid CNN-Transformer achieves 64.86% accuracy, 67.29% precision, 64.86% recall, and 64.49% F1-score on the independent test set. The model outperforms both single-model baselines by 5.40 percentage points in accuracy and the alternative hybrid model by 9.45 percentage points. Grad-CAM visualization indicates that the model primarily focuses on pathological leaf regions rather than irrelevant background elements. These findings demonstrate the potential of hybrid CNN-Transformer architectures for tomato disease classification under realistic background conditions.

Keywords: tomato leaf disease; Hybrid CNN-Transformer; MobileNetV3-Large; Vision Transformer; PlantDoc; background noise; Grad-CAM.

INTRODUCTION

Tomato (*Solanum lycopersicum*) is one of the economically important horticultural crops, but its productivity is highly vulnerable to various diseases affecting leaves and other plant organs. Common tomato leaf diseases include early blight, late blight,

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

bacterial spot, leaf mold, Septoria leaf spot, mosaic virus, and yellow leaf curl virus. These diseases can reduce photosynthetic activity, impair plant growth, decrease fruit quality, and ultimately cause substantial yield losses. Early and accurate disease identification is therefore essential for supporting timely disease management and maintaining agricultural productivity. Recent studies have demonstrated that deep learning and computer vision can provide effective approaches for automated plant disease identification, particularly through image-based classification methods (Singh et al., 2020; Tomato Leaf Disease Detection Using CNN, 2024).

Conventional disease identification in agricultural environments still relies substantially on visual inspection by farmers or agricultural experts. Although human visual diagnosis can be effective under favorable conditions, it is subjective, time-consuming, and strongly dependent on the availability of experienced personnel. The increasing availability of digital images and advances in artificial intelligence have consequently encouraged the development of automated computer vision systems for plant disease detection. Convolutional neural networks (CNNs), in particular, have demonstrated strong capabilities in learning discriminative visual features directly from plant images without requiring extensive manual feature engineering. However, the effectiveness of such systems depends strongly on the characteristics and diversity of the image datasets used for training and evaluation.

A major challenge in plant disease classification is the discrepancy between controlled laboratory images and images acquired under actual agricultural conditions. Many conventional plant disease datasets consist predominantly of images captured against relatively uniform backgrounds and under controlled illumination. In contrast, field or *in-the-wild* images contain substantial environmental variations, including soil, weeds, overlapping leaves, shadows, illumination changes, and partial occlusion. These elements may introduce visual features that are unrelated to the disease itself and can consequently interfere with the classification process. Singh et al. (2020) developed the PlantDoc dataset specifically to address the limited availability of non-laboratory plant disease images. The dataset contains images collected under more realistic visual conditions and was introduced as a resource for developing scalable computer vision approaches for plant disease detection.

The presence of background noise creates a particular challenge for CNN-based models. CNNs are highly effective at extracting local spatial patterns, including edges, textures, shapes, and localized lesion characteristics. These properties are particularly valuable for recognizing small pathological structures on plant leaves. Nevertheless, local feature extraction may also capture irrelevant background patterns when disease symptoms are visually similar to surrounding environmental elements. Consequently, although CNNs can provide strong local representations, their limited ability to model

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

long-range spatial relationships may restrict their capacity to comprehensively interpret the spatial context of complex *in-the-wild* images.

Vision Transformers (ViTs) provide an alternative approach by applying self-attention mechanisms to sequences of image patches. Rather than relying primarily on local convolutional operations, ViTs model relationships among image patches and can therefore capture broader contextual dependencies across an image (Dosovitskiy et al., 2021). This capability is potentially beneficial for plant disease classification because disease symptoms may be distributed across different regions of a leaf and may require contextual interpretation. However, pure Transformer architectures generally benefit from large-scale training data and may be vulnerable to irrelevant visual information when training data are limited or highly heterogeneous.

The complementary characteristics of CNNs and Transformers have motivated the development of hybrid CNN-Transformer architectures. In such architectures, convolutional layers can provide strong local feature extraction and inductive biases, while Transformer encoders can model global contextual relationships. This combination is particularly relevant for disease classification in *in-the-wild* environments, where the model must simultaneously recognize fine-grained pathological patterns and distinguish disease-related information from complex environmental backgrounds. The research problem addressed in this study is therefore not merely the classification of tomato diseases, but the development of an architecture capable of maintaining discriminative local representations while incorporating global contextual information under background noise conditions.

To implement the convolutional component efficiently, this study employs MobileNetV3-Large as the primary CNN backbone. MobileNetV3 was developed using hardware-aware neural architecture search and efficiency-oriented architectural improvements, making it suitable for computationally constrained computer vision applications (Howard et al., 2019). MobileNetV3-Large provides a relatively lightweight feature extraction framework while retaining strong image classification capability. Its use as the front-end feature extractor is intended to capture localized visual characteristics of tomato leaf lesions while reducing unnecessary computational complexity.

The extracted convolutional feature representation is subsequently transformed through a Feature Fusion Bridge before being processed by a Vision Transformer encoder. The Feature Fusion Bridge serves as an intermediate mechanism for converting the spatial feature representation generated by the CNN into a token-based representation compatible with Transformer processing. Through this configuration, local visual information obtained by MobileNetV3-Large can be integrated with the global self-attention mechanism of the ViT encoder. The proposed configuration is therefore

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

expected to overcome the limitations of using either a purely convolutional or purely Transformer-based architecture for complex field imagery.

Despite the increasing application of deep learning to plant disease recognition, an important research gap remains in the evaluation of lightweight hybrid CNN-Transformer architectures specifically under realistic background-noise conditions. Existing approaches have demonstrated the potential of CNNs, Transformers, and hybrid architectures for plant disease recognition, but the role of the feature-fusion mechanism in maintaining disease-relevant representations while suppressing irrelevant environmental information requires further investigation. In the context of tomato leaf disease classification, the present study addresses this gap by employing the tomato subset of the PlantDoc dataset, which represents realistic *in-the-wild* visual conditions. The study documented in the present research uses 685 original tomato images representing eight condition classes, followed by a stratified 80:10:10 partition and targeted training-data augmentation. □

Accordingly, this study proposes a Hybrid CNN-Transformer model for tomato leaf disease detection under background noise, integrating MobileNetV3-Large, a Feature Fusion Bridge, and a Vision Transformer encoder. The proposed architecture is evaluated against two single-paradigm baselines, namely pure MobileNetV3 and pure ViT-B/16, as well as an alternative hybrid architecture based on EfficientNet-B0. The evaluation considers Accuracy, Precision, Recall, and F1-Score to assess classification performance. In addition, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed as an Explainable Artificial Intelligence (XAI) technique to examine whether the proposed model focuses on pathological regions of tomato leaves rather than irrelevant background elements. The experimental results reported in the study indicate that the proposed hybrid model achieves an accuracy of 64.86% and an F1-score of 64.49%, outperforming both single-paradigm baselines and the alternative hybrid configuration.

The main contribution of this research is therefore the implementation and empirical evaluation of a lightweight Hybrid CNN-Transformer architecture designed for tomato leaf disease classification under realistic background interference. More specifically, the study contributes by integrating local feature extraction through MobileNetV3-Large with global contextual modeling through a Vision Transformer, evaluating the architecture using an *in-the-wild* tomato dataset, comparing it with single CNN and Transformer baselines, and examining model attention through Grad-CAM. This combination provides an experimental framework for investigating whether the integration of local convolutional representations and global self-attention can improve the robustness of tomato disease classification in visually complex agricultural environments.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

LITERATURE REVIEW

2.1 Tomato Leaf Disease Detection and PlantDoc Dataset

Tomato (*Solanum lycopersicum*) is susceptible to various fungal, bacterial, and viral diseases that can significantly affect plant growth, fruit quality, and productivity. Automated disease detection using computer vision and deep learning has become an important research direction because conventional visual diagnosis is subjective, time-consuming, and dependent on expert availability. Deep learning models, particularly Convolutional Neural Networks (CNNs), can automatically learn visual characteristics associated with disease symptoms such as spots, lesions, discoloration, and abnormal textures. However, many existing plant disease studies rely on controlled laboratory images, which may not adequately represent actual agricultural environments. Singh et al. (2020) introduced the PlantDoc dataset to address this limitation by providing plant disease images collected under non-laboratory and more realistic conditions. In this study, the tomato subset of PlantDoc is used to evaluate disease classification under *in-the-wild* conditions characterized by complex backgrounds and environmental variations.

2.2 CNN and Vision Transformer for Image Classification

CNNs are effective in extracting local visual features such as edges, textures, shapes, and localized pathological patterns. This characteristic makes CNNs suitable for identifying fine-grained disease symptoms on tomato leaves. Nevertheless, CNN-based architectures have limitations in explicitly modeling long-range relationships between distant image regions. Vision Transformer (ViT), in contrast, represents an image as a sequence of patches and uses self-attention to model global contextual relationships (Dosovitskiy et al., 2021). This capability enables ViT to capture relationships between spatially distant regions, but its performance generally benefits from large-scale training data and may be challenged by limited and heterogeneous datasets. MobileNetV3-Large is selected in this study as a lightweight CNN backbone because it provides an efficient balance between feature extraction capability and computational requirements (Howard et al., 2019). The combination of CNN-based local representation and Transformer-based global representation therefore provides a theoretical basis for developing a hybrid architecture.

2.3 Hybrid CNN-Transformer and Background Noise

Hybrid CNN-Transformer architectures combine the complementary strengths of convolutional feature extraction and Transformer-based global attention. CNNs can capture detailed local disease characteristics, while Transformers can model broader spatial relationships and contextual dependencies. This combination is particularly relevant to *in-the-wild* tomato leaf images, where soil, weeds, overlapping leaves,

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

shadows, illumination variations, and other environmental elements may interfere with disease recognition. Recent studies have demonstrated the potential of hybrid CNN-Transformer architectures for plant and tomato disease recognition under complex visual conditions. However, the specific integration of a lightweight MobileNetV3-Large backbone, Feature Fusion Bridge, and ViT Encoder for tomato disease classification under background noise remains an important research area. Therefore, the present study evaluates the proposed hybrid architecture against pure MobileNetV3, pure ViT-B/16, and an alternative EfficientNet-B0-based hybrid model. The Feature Fusion Bridge is used to transform CNN feature representations into token representations that can be processed by the Transformer, enabling the integration of local pathological features and global contextual information.

RESEARCH METHODOLOGY

This study employs an experimental quantitative approach to develop and evaluate a hybrid CNN-Transformer model for tomato leaf disease detection under background noise. The experiment uses the tomato subset of the PlantDoc Dataset, consisting of 685 original images representing eight tomato leaf conditions. The dataset is divided using a stratified 80:10:10 scheme into training, validation, and independent testing sets, with targeted physical data augmentation applied only to the training data, resulting in 1,614 augmented training images, 73 validation images, and 74 test images. The proposed architecture, referred to as Proposed Hybrid 1, integrates MobileNetV3-Large as a local feature extractor, a Feature Fusion Bridge for transforming CNN feature representations into Transformer-compatible tokens, and a ViT Encoder for global contextual modeling. Model performance is evaluated against two single-model baselines, namely pure MobileNetV3 and ViT-B/16, as well as Proposed Hybrid 2 based on EfficientNet-B0. Training is conducted for 25 epochs using a batch size of 16, Cross-Entropy Loss, and the AdamW optimizer, while model performance is assessed using Accuracy, Precision, Recall, and F1-Score. In addition, Grad-CAM is employed as an Explainable Artificial Intelligence technique to examine the spatial regions influencing model predictions and determine whether the model focuses on pathological leaf lesions rather than irrelevant background elements.

RESULTS AND DISCUSSION

4.1 Model Training Performance

The Proposed Hybrid 1 model, consisting of MobileNetV3-Large, Feature Fusion Bridge, and ViT Encoder, was trained for 25 epochs using a batch size of 16, Cross-Entropy Loss, and the AdamW optimizer. The training accuracy increased progressively and reached **98.70%** at epoch 25. Meanwhile, the highest validation accuracy of **68.49%**

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

was obtained at epoch 7. After this point, validation accuracy fluctuated between approximately 56% and 65%, while validation loss remained relatively high. This difference between training and validation performance indicates the presence of a generalization gap, suggesting that the model learned the training data effectively but faced greater challenges when processing unseen images containing complex natural backgrounds.

4.2 Comparative Classification Performance

The proposed model was evaluated using 74 independent test images and compared with two single-model baselines and an alternative hybrid architecture. Proposed Hybrid 1 achieved an **Accuracy of 64.86%, Precision of 67.29%, Recall of 64.86%, and F1-Score of 64.49%**. These results were higher than Baseline 1, which used pure MobileNetV3 and achieved 59.46% Accuracy and 57.67% F1-Score, and Baseline 2, which used pure ViT-B/16 and achieved 59.46% Accuracy and 59.60% F1-Score. Therefore, Proposed Hybrid 1 produced an absolute accuracy improvement of **5.40 percentage points** compared with both single-model approaches.

4.3 Comparison with the Alternative Hybrid Model

The proposed architecture was also compared with Proposed Hybrid 2, which employs EfficientNet-B0 as the CNN backbone. Proposed Hybrid 2 achieved an Accuracy of **55.41%** and an F1-Score of **52.56%**, whereas Proposed Hybrid 1 achieved 64.86% Accuracy and 64.49% F1-Score. Consequently, Proposed Hybrid 1 achieved an accuracy advantage of **9.45 percentage points** and an F1-Score advantage of **11.93 percentage points**. These findings indicate that backbone selection has an important influence on the effectiveness of the hybrid CNN-Transformer architecture.

4.4 Effectiveness of Hybrid CNN-Transformer Architecture

The superior performance of Proposed Hybrid 1 can be associated with the complementary roles of its architectural components. MobileNetV3-Large provides local feature extraction that is useful for identifying fine-grained pathological characteristics, while the Feature Fusion Bridge converts the convolutional representation into a token representation that can be processed by the ViT Encoder. The Transformer subsequently models global contextual relationships among different spatial regions. This integration enables the model to combine detailed local information with broader contextual information, which is particularly relevant for tomato leaf images captured under complex in-the-wild conditions. The results therefore support the research premise that combining CNN and Transformer representations can provide advantages over using either paradigm independently.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

4.5 Grad-CAM and Model Interpretability

Grad-CAM was employed to examine the spatial regions influencing the model's predictions. The visualization indicates that Proposed Hybrid 1 generally directs its attention toward pathological regions on tomato leaves rather than irrelevant environmental elements such as soil or weeds. This finding provides qualitative evidence that the model is learning disease-related visual characteristics instead of relying predominantly on background information. However, some misclassifications remain because different diseases may exhibit highly similar visual patterns. For example, coalescing bacterial spots can form dark necrotic regions that resemble early blight lesions, while image-resolution limitations may reduce the visibility of subtle diagnostic characteristics.

DISCUSSION

The experimental results demonstrate that **Proposed Hybrid 1 provides the best performance among the evaluated architectures**. Its combination of MobileNetV3-Large for local feature extraction and ViT for global contextual modeling produces better test performance than pure MobileNetV3, pure ViT-B/16, and the EfficientNet-B0-based hybrid model. The findings support the use of hybrid CNN-Transformer architectures for tomato leaf disease classification under background noise. Nevertheless, the test accuracy of 64.86% and the substantial difference between training and validation performance indicate that classification under realistic in-the-wild conditions remains challenging. Future improvements should therefore focus on increasing data diversity, strengthening generalization, and optimizing the hybrid architecture to better distinguish visually similar disease classes.

CONCLUSION

This study demonstrates that the proposed **Hybrid CNN-Transformer (Proposed Hybrid 1)**, integrating MobileNetV3-Large, a Feature Fusion Bridge, and a ViT Encoder, can effectively classify tomato leaf diseases under complex *in-the-wild* background conditions. Evaluated on 74 independent test images, the proposed model achieved an Accuracy of **64.86%**, Precision of **67.29%**, Recall of **64.86%**, and F1-Score of **64.49%**, outperforming both pure MobileNetV3 and ViT-B/16 baselines by **5.40 percentage points in accuracy** and outperforming the EfficientNet-B0-based hybrid model by **9.45 percentage points**. Grad-CAM analysis further indicated that the model predominantly focused on pathological regions rather than irrelevant background elements, supporting the effectiveness of combining local CNN features with global Transformer contextual modeling. Nevertheless, the observed generalization gap and remaining errors caused by visual similarity between disease classes indicate that tomato disease classification

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

under realistic field conditions remains challenging. Overall, the findings support the potential of lightweight Hybrid CNN-Transformer architectures asa practical approach for robust tomato leaf disease detection in complex natural environments.

REFERENCES

Barbedo, J. G. A. (2018). Factors influencing the use of deep learning for image-based plant disease recognition. *Biosystems Engineering, 172*, 84–91. https://doi.org/10.1016/j.biosystemseng.2018.05.013

Brahimi, M., Boukhalfa, K., & Moussaoui, A. (2017). Deep learning for tomato diseases: Classification and symptoms visualization. *Applied Artificial Intelligence, 31*(4), 299–315. https://doi.org/10.1080/08839514.2017.1315516

Chen, Z., Wang, G., Lv, T., & Zhang, X. (2024). Using a hybrid convolutional neural network with a Transformer model for tomato leaf disease detection. *Agronomy, 14*(4), 673. https://doi.org/10.3390/agronomy14040673

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

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture, 145*, 311–318. https://doi.org/10.1016/j.compag.2018.01.009

Fuentes, A., Yoon, S., Kim, S. C., & Park, D. S. (2017). A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors, 17*(9), 2022. https://doi.org/10.3390/s17092022

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv*. https://arxiv.org/abs/1704.04861

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1314–1324. https://doi.org/10.1109/ICCV.2019.00140

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708. https://doi.org/10.1109/CVPR.2017.243

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Hughes, D. P., & Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv*. https://arxiv.org/abs/1511.08060

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Karthik, R., Hariharan, M., Anand, S., Mathikshara, P., Johnson, A., & Menaka, R. (2020). Attention embedded residual CNN for disease detection in tomato leaves. *Applied Soft Computing, 86*, 105933. https://doi.org/10.1016/j.asoc.2019.105933

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision Transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

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

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., & Batra, N. (2020). PlantDoc: A dataset for visual plant disease detection. *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, 249–253. https://doi.org/10.1145/3371158.3371196

Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., & Stefanovic, D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience, 2016*, Article 3289801. https://doi.org/10.1155/2016/3289801

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114. https://proceedings.mlr.press/v97/tan19a.html

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

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

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*, 10347–10357. https://proceedings.mlr.press/v139/touvron21a.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. https://arxiv.org/abs/1706.03762

Zhang, X., Wang, G., & colleagues. (2024). Hybrid convolutional and Transformer-based approaches for tomato leaf disease recognition. *Agronomy, 14*, 673. https://doi.org/10.3390/agronomy14040673

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.