Advances in Engineering Research : Possibilities and Challenges
ISSN Print: 3079-5192
ISSN Online: 3079-5206
https://doi.org/10.63313/AERpc.2013

Integrating ResNet-50 and Vision Transformer Architectures for Robust and Efficient Tomato Fruit Ripeness Classification

KASONGO WABANGA ELOGE
Zhejiang University of Science and Technology, Hangzhou 310023, China
Email: elogeskasongo@outlook.com

How to cite this paper: Kasongo Wabanga Eloge (2025). Integrating ResNet-50 and Vision Transformer Architectures for Robust and Efficient Tomato Fruit Ripeness Classification. Advances in Engineering Research: Possibilities and Challenges, 1(2), 102–112. ISSN Print: 3079-5192; ISSN Online: 3079-5206.
https://doi.org/10.63313/AERpc.2013
Published: 2025-05-16

Abstract

This article introduces a new hybrid deep learning framework that combines ResNet-50 and Vision Transformer (ViT) models to classify tomato fruits by ripeness and quality. The hybrid model takes advantage of the ResNet-50's capability to extract features in local spatial regions. Additionally, it combines ResNet-50 with ViT's ability to establish a global contextual relationship between elements with self-attention. The framework overcomes the limitations of utilizing a single model. The framework was trained and evaluated, using a balanced and diverse dataset. It consists of four tomato classes, ripe, unripe, unevenly ripened, and damaged, collected under controlled conditions in the field. The experiments conducted aggressively in the study demonstrated a high classification accuracy of 98% with a hybrid framework compared to individual ResNet-50 and ViT models. The results in addition to the general performance were validated further using precision, recall, the area under the curve (AUC) and the F1-score. The study also included a computational efficiency test to look for accuracy versus multiple time and spatial resource costs. The research successfully addresses the challenges with dataset diversity, computational costs, and real-time feature deployments through the conceptualization of strategies (data augmentation, transfer learning) to improve generalization. The hybrid framework is expected to provide similar design principles and measures for use with other agricultural products. It can help in real-time sorting applications in variable real-world scenarios in agricultural operations. The research provides a pathway for developing smarter, scalable, and sustainable solutions in food quality and safety systems within precision agriculture. Future work will incorporate deployment and optimize the framework for edge devices in a real-world variable farm environment.

Keywords

Deep Learning; Tomato Classification; ResNet-50; Vision Transformer; Computer Vision; Smart Agriculture; Crop Quality Control; Automated Sorting

1. Introduction

Tomato (Solanum lycopersicum L.) is among the ten most produced horticulture crops worldwide, with 186 million tonnes harvested in 2022. Tomatoes serve as a dietary staple that is rich in vitamins and antioxidants and has a large economic impact in both fresh and processed commodities and export markets. Accurate maturity and quality classification of tomato fruit is critical for pricing, post-harvest handling, predicting shelf-life and reducing food waste through targeted sorting. Figure 1 shows the production quantity of tomatoes and the harvested area.

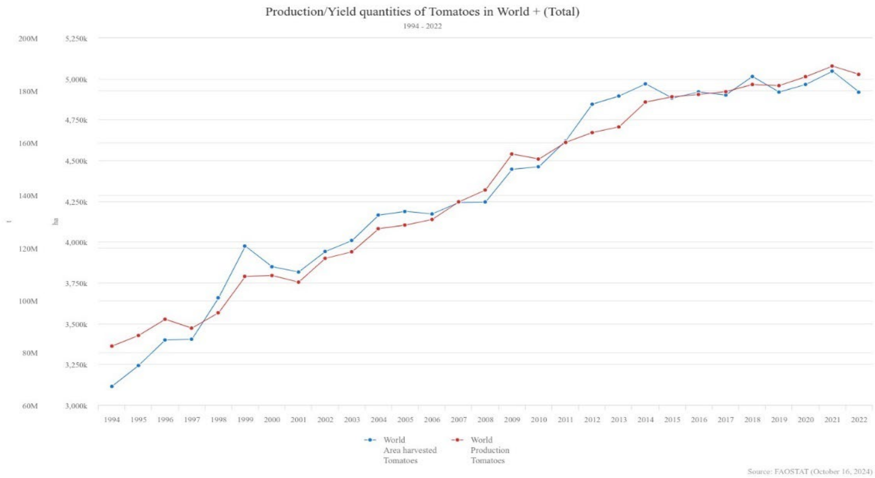

Figure 1. Production Market Statistics of Tomatoes

Traditional manual sorting is based on trained workers visualizing tomato fruit, but there is inherent variability in the classification due to human factors (e.g. fatigue, lighting, and human bias). This can create issues of economic loss, alter ripening distributions, and ultimately create situations whereby consumer satisfaction dwindles. Automated systems utilizing computer vision can achieve consistent and extensive classification of tomato fruit. However, they face challenges that seek adaptability and scalability within the environment of relentless change. The rise of deep learning and the use of imaging in agriculture have recently become prominent. ResNet-50 has exhibited superior hierarchical extracts of local features across edges, textures and color. It can struggle with recognizing long-range dependencies, such as spatial displacement across multiple fruit or background contexts. Vision Transformers (ViTs) use self-attention to model images as sequences of patches and facilitate a global context model, but typically require large data and significant computing power. In agriculture, hybrid models that incorporate CNNs' inductive biases and contextual reasoning of transformers have been proposed to address these limitations. They have shown successful applications in domains such as medical imaging, remote sensing, and plant disease detection.

This research explores how the ResNet-50 and ViT models can classify different types of tomato fruits while considering their ripeness levels. On the other hand, it is also necessary to implement significant changes in how farms operate to enhance the quality of food production for a better and sustainable future. Integrating the models allows the ResNet-50's localized feature extraction abilities to work alongside the contextual analysis provided by ViT. This enables the model to assess signals with greater and more beneficial accuracy for predictive maintenance approaches. It will be assessed using standard evaluation metrics. The method used in this study is unspecific to tomatoes; it can apply to other fruits and vegetables.

Here we present Integrating ResNet-50 and Vision Transformer Architectures for Robust and Efficient Tomato Fruit Ripeness Classification. A hybrid framework for classifying tomato fruit into four classes; ripe, unripe, uneven ripe, and damaged by the integration of feature representations ResNet-50 and ViT.

Our contributions are the following:

- A dataset of 2,741 high-resolution tomato images was expertly annotated.
- A hybrid fusion network utilizing concatenation of ResNet-50 and ViT embeddings before a classifier with a lightweight model.
- Extensive evaluation through ablation studies, class activation mapping, and latency benchmarks that demonstrate both state-of-the-art performance and feasibility.

The remainder of the article is organized as follows: Section 2 presents a review of related work; Section 3 describes the dataset, preprocessing, and model architecture; Section 4 presents results; Section 5 discusses limitations and future work; Section 6 provides conclusions.

2. Related Work

2.1. Traditional Methods

The classification of Tomatoes using traditional methods is mainly based on visual inspection, referring to a manual grading process paying attention to attributes like color, size and surface imperfections. Although these methods are inexpensive to implement, the entire process is intrusive, labor-intensive, subject to human error or bias, repetitive and generally unsustainable. These methods are greatly limited by the instability and lack of scale that these commercial agricultural processes must constantly contemplate. Over the last ten years, automated fruit classification has moved from classical image processing methods to deep learning models. Here, we review important developments of different classification methods for agriculture applications.

2.2. Machine learning Approaches

Conventional machine learning methods are still in use, and they continue to be applicable—albeit with algorithms such as support vector machines (SVM), decision trees and k-nearest neighbors (KNN), among others. These machine learning approaches require working with effective feature extraction; in other words, they require determining the essential features of the tomato images for classification. Although conventional methods can be impactful, the work required in feature extraction and algorithm tuning requires a fair amount of manual work.

2.3. Deep Learning Techniques

Deep learning, particularly convolutional neural networks (CNNs), has greatly improved the state of the art in classification accuracy. CNNs take raw image data as input and automatically learn hierarchical features, both improving classification accuracy and reducing preprocessing effort significantly. Similar to the tomato classification that I outlined above, data can now be analyzed in depth with limited human-engineered feature engineering or similar preprocessing functionality. Models such as VGG16 and ResNet-50 are highly favored for their outstanding feature extraction abilities from image data. This allows them to learn hierarchically and effectively to detect diseases, assess fruit ripeness, and perform quality grading. Studies have provided evidence of the increase in accuracy and reliability compared to using traditional models including studies applying transfer learning for faster training convergence using appropriately pre-trained weights.

Vision Transformer (ViT), a development designed for natural language processing, continues to be successfully used for image classification. They segment images into patches and use a modified self-attention schema to allow ViT to capture global contextual information. There is evidence that ViT models are superior to CNNs because they can model long-range (i.e., between spatially distant) dependencies with less reliance on pre-processing pipelines and pre-training periods. In some cases of plant disease recognition, ViT models have been shown to outperform CNNs. However, these models generally require large datasets and heavy computation.

2.4. Enhanced Classification Strategies

An increasing number of research studies are exploring the possibilities of hybrid models that combine CNNs for local feature extraction with ViT models to enhance global attention. This effectively overcomes the limitations that come with relying solely on one approach. Models that integrate ViT and CNN architectures have achieved better classification accuracy and robustness. In recognizing plant images under diverse field conditions. Allowing them to handle uncertainties arising from environmental variations, occlusions caused by plant structures, and subtle disease indicators. Other approaches have incorporated classic machine learning techniques, like SVMs, to classify based on detector features extracted using CNN architectures and improved scores from classification. Other complementary technologies, such as IoT devices and real-time image analysis, have been considered in some systems to allow for 24/7 crop monitoring and real-time disease identification. Data augmentation techniques and explainable AI methods (such as Grad-CAM++) are methods to improve the generalizability of the model and are helpful when farmers evaluate trust and ease of deployment.

Nonetheless, several challenges remain with obtaining high-quality and highly enriched labeled datasets, overfitting on small datasets and computational efficiency for real-time capabilities for edge devices. These issues highlight the need for additional research into optimized architectures, complementary multimodal data (e.g., hyperspectral imaging), and more lightweight model alternatives.

3. Methodology

3.1. Overview of the Hybrid Model

ResNet-50 is a 50-layer deep, convolutional neural network (CNN) that provides residual skip connections to alleviate the vanishing gradient problem and enables training very deep architectures. In the case of transfer learning, ResNet-50 will enable the model to train to tomato-based imagery using weights from large public datasets (like ImageNet). This would require less time and dramatically smaller computational memory than training a model from scratch.

The ViT was originally developed with tasks in natural language processing in mind. To employ an elegant approach to processing images by cutting them into fixed-size patches, and relying on self-attention to identify global dependencies. This ability provides more flexibility in thinking about the images that make up complex visual displays, with the understanding of frequent change (i.e., occlusions, variable light conditions) in agriculture. In the hybrid model we envision, we will take ResNet-50 with its high-level feature maps and cut the images into patches for the ViT layers. The hybrid model combines the advantages of the two networks for maximum accuracy of local features (i.e., tomato) places with global features (i.e., local conditions) to consider the visual domain. Figure 2 depicts the steps of the Hybrid Model.

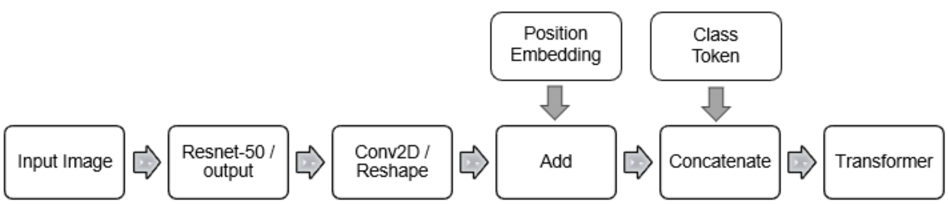

Figure 2. The phase of the combination of the ResNet-50 with ViT

3.2. Dataset and Experimental setup

A balanced dataset consisting of 2,741 tomato images corresponding to four classes: ripe, unripe, damaged, and uneven ripe, is prepared from publicly available repositories under different environmental conditions. The images are preprocessed by resizing to 224×224 (width x height) pixels, normalization, and extensive use of data augmentation through numerous features including flipping, rotating, brightness change, zooming, and cropping. This preprocessing is used to improve the robustness of the model and prevent overfitting. Figure 3 shows a Sample Representation of the Four classes of Tomatoes in the dataset. The experimental process involves using ResNet-50, which has been previously proven to produce local feature extraction, using pre-trained weights from ImageNet, and fine-tuned to perform tomato classifications. In this process, we replace its fully connected layers with a flattened and softmax layer corresponding to the four classes. In the first stages of model development, the lower layers are frozen and used to maintain general features learned in the lower layers of the model. Eventually, these layers will also be tuned for features specific to the domain. To achieve this, ViT takes input from the feature maps generated from the ResNet-50 and produces patches. Then feed these patches through transformer encoder layers, utilizing a multi-head self-attention mechanism to look for contextual relationships within the global space of the image; this feature is frequently missed with convolution networks alone. The fusion step will combine the higher-level feature maps generated by ResNet-50 with the global view produced from ViT's attention-driven layers. This will generate the patch embeddings for the feature maps and feed them into the layers of the transformer encoder which will train the entire model. Then, it will produce a final output of class labels through a fully connected layer taking the aggregated class token output provided from the final transformer encoder output.

(a) **(b)** **(c)** **(d)**

Figure 3. Illustrations of Tomatoes in our file (a-d) are enlarged views of Four group folders of the dataset

Hyperparameters were selected to balance convergence and generalization to train the model over fifty epochs for all labeled images in the dataset, utilizing categorical cross entropy for loss. The Adam optimizer was adopted with a learning rate set of 0.0001, 50 epochs and the batch size set for 32.

Implementation of the study is completed in Python utilizing TensorFlow and Keras for the utilization of a deep learning framework, using an AMD Ryzen CPU and AMD Radeon GPU. Other libraries include OpenCV for image processing and scikit-learn to evaluate models, metrics, and graphs to support the evaluation process. The development environment is VS Code, which allows a complete coding experience to produce code and debug and implement version control.

Evaluation of the model using the appropriate metrics has been defined to evaluate the overall performance of the model. Accuracy, precision, recall, F1-score, confusion matrix, ROC curves and area under the curve (AUC) have been provided as appropriate for each class of classification.

4. Results and Comparative Analysis

The hybrid model attained a notable classification accuracy rate of 98%, which is a marked improvement over the individual ResNet-50 (85.2%) and ViT (94.3%) models. Balanced precision and recall indicated the model's reliability of correctly classifying tomato classes under a range of conditions at approximately 0.99. The ROC-AUC score of 0.9646 provided additional evidence of the model's superior discriminating power. Table 1 is the summary of trained Models in this project.

Table 1. Experiment Results

Models	Accuracy	Precision	Recall	F1-Score	Training Time
ResNet-50	85.2%	0.85	0.83	0.84	~8 hours
ViT	94.3%	0.98	0.98	0.98	~20 hours
MobileNet	95%	0.94	0.92	0.93	~5 hours
EfficientNet	96%	0.95	0.93	0.94	~7 hours
Hybrid	98%	0.99	0.99	0.99	~15 hours

While ResNet-50 demonstrated efficiency in training and inference speed for classifying image input, it likely fails to capture broader supports or context cues, specifically multi-class tomato conditions; ViT was accurate, however computationally resource intensive. Table 2 shows the computational efficiency of the State-of-the-art models. The hybrid model succeeds in balancing the benefits of ResNet-50 and ViT. It has included all relevant feature analysis through multiple data analysis views while avoiding high computational resource expenditures, i.e., approximately 8.9 GFLOPs and 4 GB of memory. This makes the model deployable in various agricultural systems with mid-range capacity control computers.

Table 2. Computational Efficiency Metrics

Models	Memory Usage (GB)	FLOPs (GFLOPs)	Param. (Millions)	Inference time(ms)
ResNet-50	~2.1	~3.8	~25	200-250
ViT	~5.5	~17.6	~86	300–350
MobileNet	~1.2	~0.57	~4-6	100-150
EfficientNet	~1.8	~0.39	~5.3	150-200
Hybrid	~4	8.9	~112	250–300

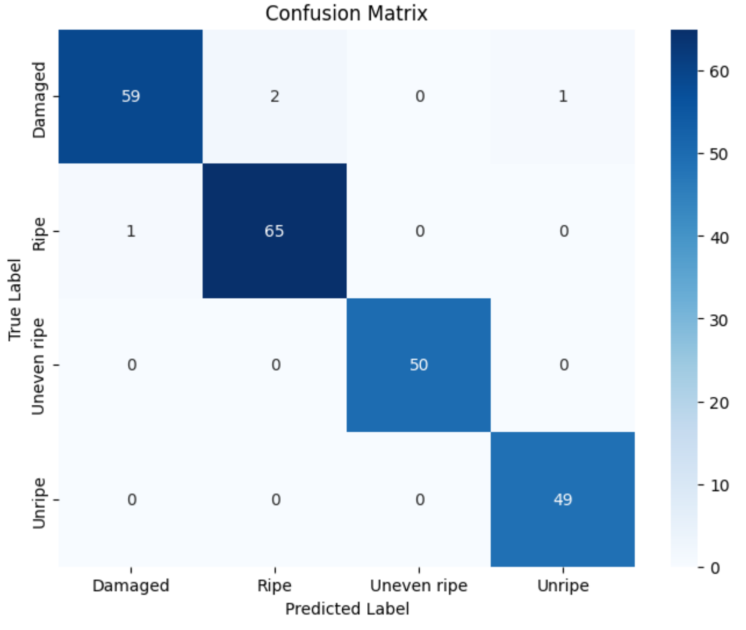

Figure 4. Confusion Matrix

Figure 4 provides a detailed breakdown of the model's performance for each class, allowing us to observe any misclassifications. The confusion matrix analyses illustrated very few misclassifications among tomato classes that are visually similar, such as ripe and uneven ripe. This is indicative of the complexity of tomato fruit maturity stages within a practical image agricultural use case, but still evidences model robustness.

5. Challenges and Future Directions

While the hybrid model demonstrates significant accuracy in the examination of disease resistance, it does carry its complexity when considering deployment in a resource-limited edge device. For instance, training the ViT components still requires extensive use of GPUs. In addition, the model cannot be novelly generalized without significant amounts of consistent labeled annotated data sets. Despite that, using both transfer learning and data augmentation approaches would reduce some of the limitations on both annotated and computational resource needs. Future work will aim to reduce our current hybrid model for model size and inference speed using model pruning or quantization. Refining the selected architecture will improve the speed. Finally, we think that utilizing multimodal imaging capabilities for things such as detecting internal defects using hyperspectral or thermal images would give value to future work. We'd hope to see either an integrated methodology for the hybrid. The development of an end-to-end station for real-time sorting that incorporates some of the foregoing methodologies with hardware (i.e., cameras on conveyor belts). The hybrid models' architectural methodologies add utility to other contexts beyond the scope of disease resistance projects, moving forward. For instance; generalizing the hybrid architecture to other crop projects would take the hybrid models to multiple possibilities as well.

6. Conclusions

The combination of ResNet-50 and Vision Transformer architectures creates a strong deep learning solution to both accurately and efficiently classify tomato fruit ripeness with robustness. This hybrid system demonstrates superior capability and performance than each model by integrating granular local feature extraction, along with holistic global context understanding. It is critical with the complexity and variability of agricultural images. Importantly, the promise of successful applications with this hybrid framework paves the way for more sophisticated automated quality control. This system reduces labor dependency and increases productivity for a wide range of precision agriculture settings. Therefore, this study, not only describes the functionalities and use cases of the hybrid design as relates to tomato fruit ripeness classification. It establishes a foundation for future precision agriculture technology represented by smart farming with artificial intelligence. It can also contribute to the methodological advancement of global food security and sustainable practices in agricultural production.

References

[1] FAO, "FAOSTAT," Food and Agriculture Organizatition of the United Nations, 2022. [Online]. Available: https://www.fao.org/faostat/en/#data/QCL/visualize. [Accessed September 2024].

[2] G. Aladekoyi, O. A. Ajala and O. A. Shakpo, "The Effect of Moringa Seeds Oil and Shea Butter Oil Waxing On the Post-Harvest Life of Tomato Fruit (Solanum lycopersicum)," European Journal of Food Science and Technology, vol. 11, no. 1, pp. 36-49, 2023.

[3] A. Kabas and I. Celik, "Effect of newly developed interspecific hybrid rootstocks on mineral nutrient composition and fruit quality in tomato (Solanum lycopersicum L.)," Acta Alimentaria, vol. 50, no. 3, p. 383–392, 2021.

[4] Z. Lang, Y. Wang, K. Tang, D. Tang, T. Datsenka, J. Cheng, Y. Zhang , A. Handa and J. Zhu, "Critical roles of DNA demethylation in the activation of ripening-induced genes and inhibition of ripening-repressed genes in tomato fruit," Proceedings of the National Academy of Sciences, vol. 114, no. 22, p. E4511–E451, 2017.

[5] L. Coman, "african american workers picking tomatoes from a conveyer belt sorting them by size," Shutterstock, January 2024. [Online]. Available: https://www.shutterstock.com/image-photo/african-american-workers-picking-tomatoes-conveyer-2412134441. [Accessed September 2024].

[6] V. Jaya, M. Sai, P. Rishitha and U. Prabu, "A Comprehensive Study on Fruit Classification and Grading Techniques," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 970-978, 2022.

[7] S. Kaur, A. Girdhar and J. Gill, Computer vision-based tomato grading and sorting, Springer, Singapore, 2018.

[8] Y. Chen, N. Zhang, A. Wang, Z. Liu and J. Yue, "Automatic classification of tomato leaf diseases based on MtConvNeXt model," 2024 8th International Conference on Electrical, Mechanical and Computer Engineering (ICEMCE), pp. 2021-2025, 2024.

[9] P. Das, J. K. P. S. Yadav and A. K. Yadav, "An Automated Tomato Maturity Grading System

Using Transfer Learning Based AlexNet," In Ingénierie des Systèmes d Inf., vol. 26, no. 2, pp. 191-200, 2021.

[10] C. Janiesch, P. Zschech and K. Heinrich, "Machine learning and deep learning," Electron Markets, no. 31, p. 685–695, 2021.

[11] M. O. Pallavi and A. Raj, "Tomato Disease Prediction Model using Machine Learning Algorithms and Image Processing Techniques," 2024 6th International Conference on Computational Intelligence and Networks (CINE), pp. 1-6, 2024.

[12] A. Hemalatha and J. Vijayakumar, "Automatic Tomato Leaf Diseases Classification and Recognition using Transfer Learning Model with Image Processing Techniques," 2021 Smart Technologies, Communication and Robotics (STCR), pp. 1-5, 2021.

[13] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[14] P. Shukla and M. Kumar, "Explicating ResNet for Facial Expression Recognition," International Journal of Computing, Intelligent and Communication Technology, vol. 11, no. 3, pp. 2319-748X, 2022.

[15] M. Sahaai, G. Jothilakshmi and D. Ravikumar, "ResNet-50 based deep neural network using transfer learning for brain tumor classification," AIP Conference Proceedings, vol. 2463, no. 1, p. 020014, 2022.

[16] S. Kumar, S. Pal, V. Singh and P. Jaiswal, "Performance evaluation of ResNet model for classification of tomato plant disease," Epidemiologic Methods, vol. 12, no. 1, p. 20210044, 2023.

[17] S. Alshammari, "TLDViT: A Vision Transformer Model for Tomato Leaf Disease Classification," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 12, pp. 841-848, 2024.

[18] E. Baek, "Attention Score-Based Multi-Vision Transformer Technique for Plant Disease Classification," Sensors (Basel, Switzerland)., vol. 25, no. 1, p. 270, 2025.

[19] Y. Xu, H. Wei, M. Lin, Y. Deng, K. Sheng, M. Zhang, F. Tang, W. Dong, F. Huang and C. Xu, "Transformers in computational visual media: A survey," Computational Visual Media, vol. 8, no. 1, pp. 33-62, 2022.

[20] A. Warohma, H. Hindersah and D. Lestari, "Speaker Recognition Using MobileNetV3 for Voice-Based Robot Navigation," 2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA), pp. 1 - 6, 2024.

[21] H. Mputu, A. Abdel-Mawgood, A. Shimada and M. Sayed, "Tomato Quality Classification Based on Transfer Learning Feature Extraction and Machine Learning Algorithm Classifiers," IEEE Access, vol. 12, pp. 8283-8295, 2024.

[22] M. Jagatheeswari and Y. Rao, "A Hybrid Approach based on Metaheuristics and Machine Learning for Tomato Plant Leaf Disease Classification," 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), pp. 1-6, 2022.

[23] A. Boukhris, J. Antari, H. Asri and A. Sadiq, "Detection and Identification of Tomato Disease Based on Deep Learning and Machine Learning algorithms," 2023 IEEE Afro-Mediterranean Conference on Artificial Intelligence (AMCAI), pp. 1-6, 2023.

[24] L. Quach, K. Quoc, A. Quynh, H. Ngoc and N. Thai-Nghe, "Tomato Health Monitoring System: Tomato Classification, Detection, and Counting System Based on YOLOv8 Model With Explainable MobileNet Models Using Grad-CAM++," IEEE Access, vol. 12, pp. 9719-9737, 2024.

[25] G. Ezhilarasan, S. Ranganathan, L. Mani and S. Kadry, "An intelligent deep residual learning framework for tomato plant leaf disease classification," International Journal of Electrical and Computer Engineering (IJECE), vol. 14, no. 3, pp. 3168-3178, 2024.

[26] R. Kasera, S. Nath, B. Das, A. Kumar and T. Acharjee, "IoT Enabled Smart Agriculture System for Detection and Classification of Tomato and Brinjal Plant Leaves Disease," Scalable Comput. Pract. Exp., vol. 26, pp. 96-113, 2025.

[27] M. Mahajan, D. Upadhyay, M. Aeri, V. Kukreja and R. Sharma, "Smart Agriculture: Lever-

aging CNN-RF Fusion for Enhanced Tomato Disease Classification," 2024 3rd International Conference for Innovation in Technology (INOCON), pp. 1-5, 2024.

[28] A. Ali and A. Abdulazeez, "Transfer Learning in Machine Learning: A Review of Methods and Applications," Indonesian Journal of Computer Science, vol. 13, no. 3, pp. 4227-4259, 2024.