OPEN ACCESS
International Journal of Advances in Artificial Intelligence and Machine Learning Vol 2 No 2 (2025) https://doi.org/10.58723/ijaaiml.v2i2.435
Check for updates
Research Article

A Comparative Study of Convolutional Neural Networks and Vision Transformers for Fruit Classification

Received: May 17, 2025	Revised: July 15, 2025	Accepted: July 21, 2025	Publish: July 23, 2025

Malik Jawarneh*, Arief Marwanto, Dedy Syamsuar, Maivi Kusnandar

Abstract:
Background of study: Accurate fruit classification is vital for agricultural automation, yet traditional methods are often subjective and inefficient. Convolutional Neural Networks (CNNs) are effective but struggle with global context in fine-grained tasks. Vision Transformers (ViTs), inspired by NLP models, offer global attention mechanisms that may improve classification in complex scenarios.
Aims and scope of paper: This study compares the performance of EfficientNet-B0 (a CNN model) and ViT-B/16 (a Transformer model) on a fruit classification task involving five fruit types. The goal is to evaluate their strengths and weaknesses under controlled experimental conditions using a moderately sized dataset.
Methods: A dataset of 10,000 fruit images was preprocessed with standard augmentation techniques and split into training and validation sets. Both models were fine-tuned using pretrained weights. Performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices.
Result: EfficientNet-B0 achieved higher overall accuracy (94%) than ViT-B/16 (92%). The CNN model performed consistently across all classes, particularly excelling in bananas and strawberries. ViT-B/16 showed superior results for strawberries but struggled with apples. Confusion matrices revealed class-specific strengths and weaknesses.
Conclusion: EfficientNet-B0 is better suited for general fruit classification due to its balanced performance, while ViT-B/16 excels in capturing fine-grained visual features. A hybrid approach may leverage both models' strengths for enhanced performance in real-world applications.

Keywords: Agricultural Automation, Convolutional Neural Networks, Fruit Classification, Image Classification, Vision Transformer.

1. INTRODUCTION

Accurate fruit classification using image data is an increasingly critical automation component in several industries, including agriculture, food processing, and retail supply chains (Pandey et al., 2023). Traditional fruit detection systems have relied on manual labour or basic machine vision techniques involving hand-crafted features (Shi et al., 2025). These approaches often suffer from subjectivity, inconsistency, and poor scalability. In the era of Industry4.0 and smart agriculture, there is a growing demand for intelligent systems that can perform rapid, reliable, and consistent classification of agricultural products, especially fruits, based on visual features such products, especially fruits, based on visual features such as colour, shape, texture, and size (Darwin et al., 2021).

Deep learning has revolutionised the field of computer vision in recent years, providing robust frameworks for feature extraction and image classification without extensive manual feature engineering (Elharrouss et al., 2024). Convolutional Neural Network (CNN) has become the de facto standard for image recognition tasks (Kiranyaz et al., 2021). The hierarchical structure of CNN enables it to learn low-level features such as edges and textures in early layers, and high-level semantic features in deeper layers (Shabir et al., 2025). As a result, CNN have achieved remarkable success across diverse computer vision tasks, including medical image analysis, object detection, and natural image classification (Dewi et al., 2024).

In agriculture, CNN have been widely adopted for applications such as plant disease detection, crop classification, and fruit recognition. (Altalak et al., 2022) conducted an extensive review highlighting the successful application of deep learning, particularly CNN, in smart farming. They demonstrated that CNN recognise different fruit species and quality grades when large annotated datasets are available.

However, while CNN perform well in general image classification tasks, challenges arise when dealing with

fine-grained classification scenarios. In fruit classification, distinguishing between varieties of fruits that share similar visual properties—like colour tone, surface glossiness, or curvature—requires more than local feature detection (Ghazal et al., 2021). This is a significant limitation of CNNs, which are designed with local receptive fields and may fail to capture global image dependencies unless architecturally deepened, leading to potential overfitting, increased memory usage, and slower training convergence (Alzubaidi et al., 2021).

Additionally, fruit images often exhibit intra-class variability due to changes in illumination, occlusion, background noise, or post-harvest conditions. Despite its robustness, CNN sometimes struggles to generalise well under such domain shifts unless explicitly trained with diverse and augmented data (Momeny et al., 2021).

The limitations of CNN in capturing global context have catalysed exploring alternative architectures. One of the most notable developments is the introduction of Vision Transformers (ViTs), a novel deep learning architecture inspired initially by the Transformer model used in natural language processing (Khan et al., 2022). (Kanadath et al., 2024) proposed ViTs as a pure attention-based model for image classification, removing convolutional operations altogether. ViTs treat an image as a sequence of fixed-size patches, embedding each patch and applying self-attention. This architecture excels at modelling global dependencies, which are often vital for tasks that require holistic reasoning about an image.

Unlike CNN, which relies on inductive biases such as translation equivariance and locality, ViTs learn to capture spatial dependencies solely through data. This enables them to represent contextual relationships between distant regions of the image more effectively. (Wang et al., 2022) further advanced this architecture through their work on Data-efficient Image Transformers (Deit), which showed that ViTs could match CNNS in performance even when trained on smaller datasets, provided that adequate regularisation and data augmentation strategies are applied.

Despite their theoretical advantages, Vision Transformers have not yet seen widespread adoption in domain-specific applications such as fruit classification. One reason is that ViTs typically require larger datasets and longer training times to outperform CNN. This raises concerns about overfitting and poor generalisation for moderate-sized datasets, such as those often encountered in agricultural settings (Zhang et al., 2025). Consequently, the comparative effectiveness of CNN and ViTs on datasets that are neither extremely small nor massive remains underexplored.

Recent work in hybrid models suggests a promising path forward. Hybrid architectures aim to combine the locality modelling strengths of CNN with the global attention mechanisms of ViTs. For example, the Conformer model integrates convolutional layers within a Transformer block to preserve local feature hierarchies while gaining the ability to model long-range dependencies (Peng et al., 2021). Similarly, Swin Transformers introduce hierarchical self-attention with shifted windows, mimicking the hierarchical representation learning of CNN while leveraging attention-based modelling (Liu et al., 2021).

In fruit classification tasks, particularly those involving subtle inter-class differences and environmental variations, such hybrid models could provide the best of both worlds (Oliullah et al., 2025). CNN locality is beneficial for capturing micro-textural patterns on fruit surfaces. At the same time, ViTs' global context modelling can help identify broader shape or arrangement patterns that are otherwise hard to detect through convolutional operations alone.

Moreover, data augmentation and regularisation become even more critical in these settings. ViTs benefit significantly from advanced augmentation strategies such as Mixup, CutMix, and RandAugment (Trigka & Dritsas, 2025). Applying these techniques effectively to moderate-sized datasets like fruit classification sets can help close the gap between CNN and ViTs.

Key performance metrics such as accuracy, precision, recall, and F1-score are computed to evaluate the models holistically (Miller et al., 2024). Confusion matrices inspect misclassification patterns, which can offer insight into whether certain classes (e.g., grapes vs. strawberries) pose consistent challenges across architectures.

While CNN is expected to perform strongly out-of-the-box due to their maturity and prevalence in image classification tasks, we hypothesise that ViTs may offer better generalisation in subtle inter-class differences, provided that data augmentation and regularisation are carefully tuned. We also explore the potential of ensemble or hybrid strategies that blend CNN and Transformer outputs, leveraging late fusion or shared feature encodings.

This comparative study aims to bridge a gap in the literature by systematically evaluating CNN and ViTs under consistent experimental conditions on a realistic fruit classification task (Espinoza et al., 2024). Few studies have directly compared these architectures in domain-specific applications, focusing on hybridisation potential.

Furthermore, this work offers practical recommendations on architecture selection and optimisation for practitioners working with moderate-scale image datasets in agriculture or retail. As the field progresses toward more intelligent supply chains and automated food quality systems, the insights from this study could inform model choices for deployment in edge devices, greenhouses, or production lines.

While CNN remains a powerful image classification tool, its architectural constraints can limit fine-grained and globally dependent tasks. Vision Transformers present an exciting alternative, especially when global reasoning is required, but their application is still

maturing in domain-specific contexts (Maurício et al., 2023). This paper seeks to illuminate the trade-offs and synergies between these two paradigms through rigorous experimentation.

Ultimately, the goal is not to crown a definitive winner between CNN and ViTs but to understand each's nuanced strengths and explore how they may complement one another in creating more robust, generalisable, and efficient fruit classification systems.

2. MATERIAL AND METHOD

Data Collection

The dataset used in this study consists of images of five different types of fruit: Apples, Bananas, Grapes, Mangoes, and Strawberries, which were obtained from the URL, https://www.kaggle.com/datasets/utkarshsaxenadn/fruits-classification (DeepNets, 2023). Each class contains 2,000 images, resulting in a total of 10,000 images. These images were sourced from publicly available online datasets, and each image is labelled according to its corresponding fruit type. The images in the dataset vary in terms of resolution, ranging from 128x128 pixels to 512x512 pixels, and include different lighting conditions, angles, and backgrounds, representing real-world variability. The dataset was divided into two main subsets to train and evaluate the models: training (80%) and validation (20%). The training set comprises 8,000 images (1,600 per class), while the validation set contains 2,000 images (400 per class). This dataset setup allows for practical training of the models and unbiased evaluation of their performance.

Data Preprocessing

All images were resized to 224×224 pixels to meet the input size requirements of the CNN and ViT models. Data augmentation techniques such as random horizontal flipping and rotation (15 degrees) were applied for the training set to improve model generalisation. After augmentation, images were converted to tensors and normalised using the standard ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. Only resizing, tensor conversion, and normalisation were performed for the validation and test sets to maintain evaluation consistency. This preprocessing pipeline ensured the dataset was standardised and enhanced for optimal model training.

Model Selection

For this study, two deep learning models were selected: Vision Transformer (ViT) and EfficientNet-B0, implemented using the timm library with pretrained weights. The first model is ViT Base Patch16/224, which treats an input image as a sequence of 16×16 patches and uses a self-attention mechanism to model long-range dependencies between patches. A pretrained ViT model was loaded, and its classification head was replaced with a new fully connected layer matching the number of fruit classes. Figure 1 shows the Vision Transformer (ViT) architecture model.

The second model is EfficientNet-B0, a Convolutional Neural Network (CNN) known for balancing model size and performance. EfficientNet scales network depth, width, and resolution systematically for better efficiency. Similarly, a pretrained EfficientNet-B0 model was loaded, and its classifier layer was replaced with a fully connected layer matching the number of output classes. By fine-tuning these pretrained models on the fruit dataset, the study compares the performance of a traditional CNN approach against a transformer-based method for image classification tasks. Figure 2 shows the Basic CNN architecture model.

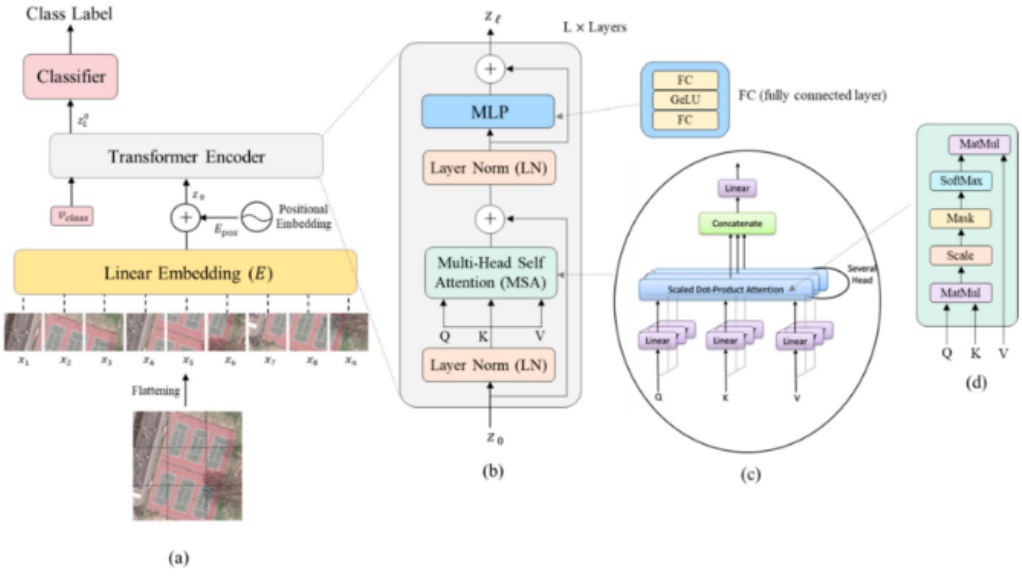

Figure 1. The Vision Transformer architecture model (Source: Wu et al., 2020)

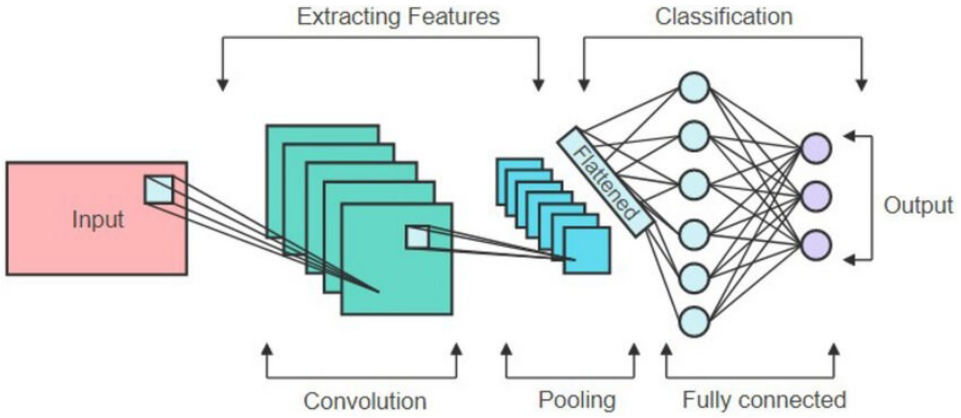

Figure 2. The Basic CNN architecture model (Ismail et al., 2023)

Training and Evaluation

The training process was conducted using the Adam optimiser with a learning rate of 1e-4 and a Cross-Entropy Loss function, which is standard for multi-class classification tasks. Models were trained for 30 epochs, and performance metrics, including training loss and training accuracy, were monitored after each epoch. The models were set to training mode during each epoch, and weights were updated via backpropagation.

After training, evaluation was performed on the validation dataset. The models were switched to evaluation mode, and validation loss and accuracy were computed without updating the gradients. Furthermore, detailed evaluation metrics were generated, including the confusion matrix, per-class accuracy, and a classification report covering precision, recall, and F1-score (Hinojosa Lee et al., 2024). Visualisation of the confusion matrix provided insights into class-wise model performance, highlighting any class imbalance or misclassification trends. Training loss and accuracy curves were plotted to assess model convergence behaviour and identify potential overfitting or underfitting.

3. RESULT AND DISCUSSION

3.1 Results

This section presents the experimental training outcomes and evaluates the CNN and Vision Transformer (ViT) models on the fruit classification dataset. The evaluation metrics included training loss, training accuracy, confusion matrix, per-class accuracy, and classification measures such as precision, recall, and F1-score.

First, we present the training accuracy and loss for both the CNN and ViT models throughout the training process. These graphs provide a clear picture of the models' learning dynamics. Figure 3 and Figure 4 show the training accuracy and loss for CNN and ViT models.

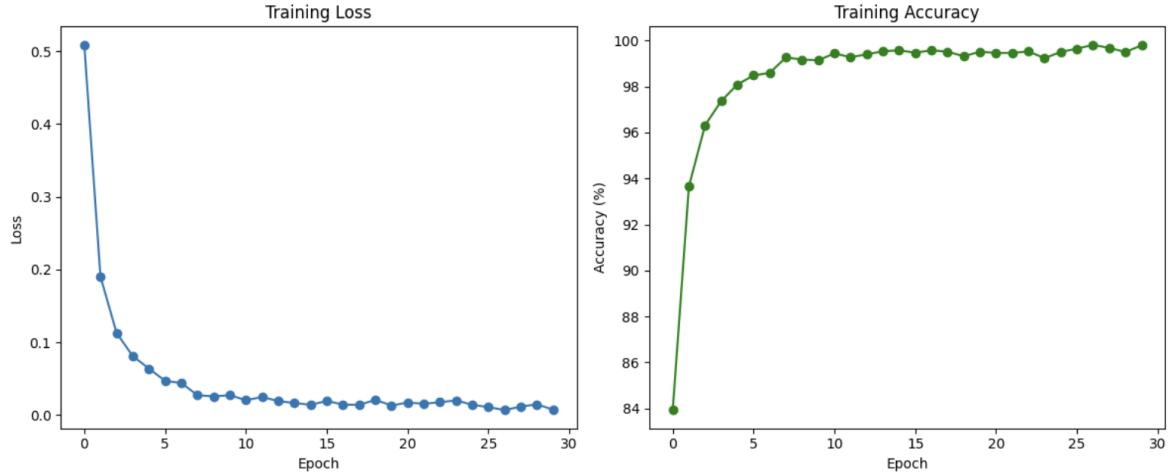

Figure 3. Training accuracy and loss graph for the CNN model

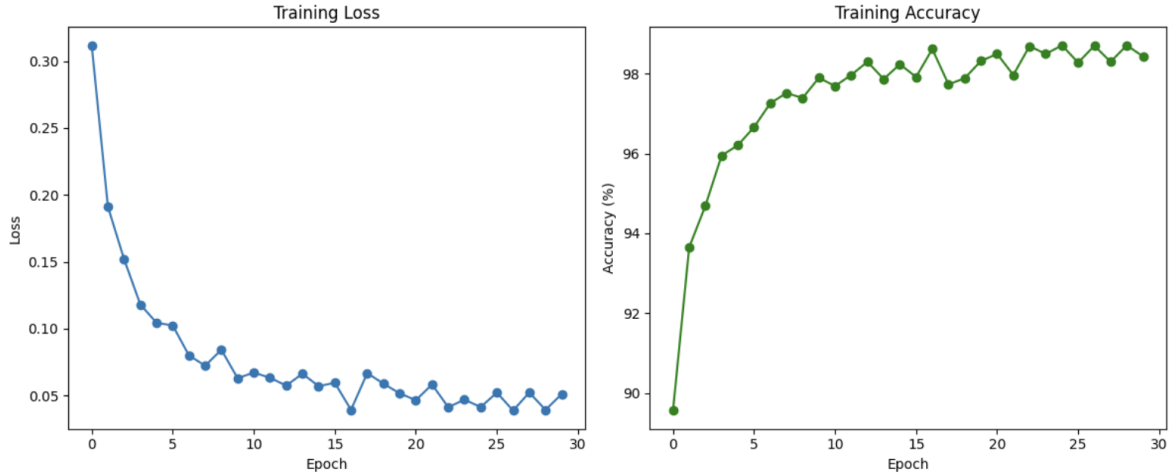

Figure 4. Training accuracy and loss graph for the ViT model

From these graphs, we can observe that both models converge relatively quickly, with the CNN model showing slightly more stability in terms of loss reduction and accuracy increase across epochs than the ViT model. Both models achieve high training accuracy, but the CNN model reaches a slightly higher accuracy earlier in the training. Table 1 and Figure 5 show the comparison results.

Table 1. The comparison results

Approaches	Accuracy	Class	Precision	Recall	F1-score
EfficientNet-B0	0.94	Apple	0.89	0.85	0.87
		Banana	1.00	0.95	0.97
		Grape	0.91	0.97	0.94
		Mango	0.95	0.95	0.95
		Strawberry	0.98	1.00	0.99
ViT-B/16	0.92	Apple	0.79	0.95	0.86
		Banana	0.93	0.93	0.93
		Grape	0.94	0.82	0.88
		Mango	0.95	0.90	0.92
		Strawberry	1.00	0.97	0.99

Figure 5. The comparison results are shown in graphs

Table 1 and Figure 5 compare the EfficientNet-B0 (CNN) and ViT-B/16 (Vision Transformer) models, focusing on accuracy, precision, recall, and F1-score for each class in the fruit classification task. The EfficientNet-B0 (CNN) model achieved an overall accuracy of 0.94, slightly outperforming the ViT-B/16 (Vision Transformer) model, which achieved an accuracy of 0.92.

The EfficientNet-B0 (CNN) model demonstrated consistent performance across all fruit classes, with high precision, recall, and F1-scores. Notably, strawberries and bananas performed exceptionally well, achieving perfect precision (1.00) and a recall of 0.95. Similarly, strawberries showed impressive precision (0.98) and a perfect recall of 1.00, resulting in an F1-score of 0.99. These results indicate that CNN models achieve high

precision and recall across the board, especially for more visually distinct fruit classes.

On the other hand, the ViT-B/16 (Vision Transformer) model, despite having a slightly lower overall accuracy of 0.92, demonstrated promising results in specific classes. The model performed particularly well with strawberries, achieving perfect precision (1.00) and a high recall (0.97), resulting in an F1-score of 0.99. However, the ViT-B/16 model struggled with apples, achieving a lower precision of 0.79 but maintaining a high recall of 0.95. This suggests that Vision Transformers, while slightly less accurate overall, may excel in fine-grained classification tasks, where understanding long-range dependencies in images becomes crucial.

Both models show strong performance in fruit classification, with EfficientNet-B0 (CNN) achieving higher accuracy overall, particularly excelling in precision and recall for apples, bananas, and mangoes. Meanwhile, ViT-B/16 shows competitive performance, especially in classifying strawberries, but struggles more with certain fruit types like apples. These results suggest that a hybrid approach, combining the strengths of both models, could potentially lead to improved performance in fruit classification tasks.

Comparison Metric

The following graph compares the performance of four models used in sentiment analysis of hotel reviews. Each model is evaluated using several metrics, including precision. Figure 6 shows the confusion matrices for both models.

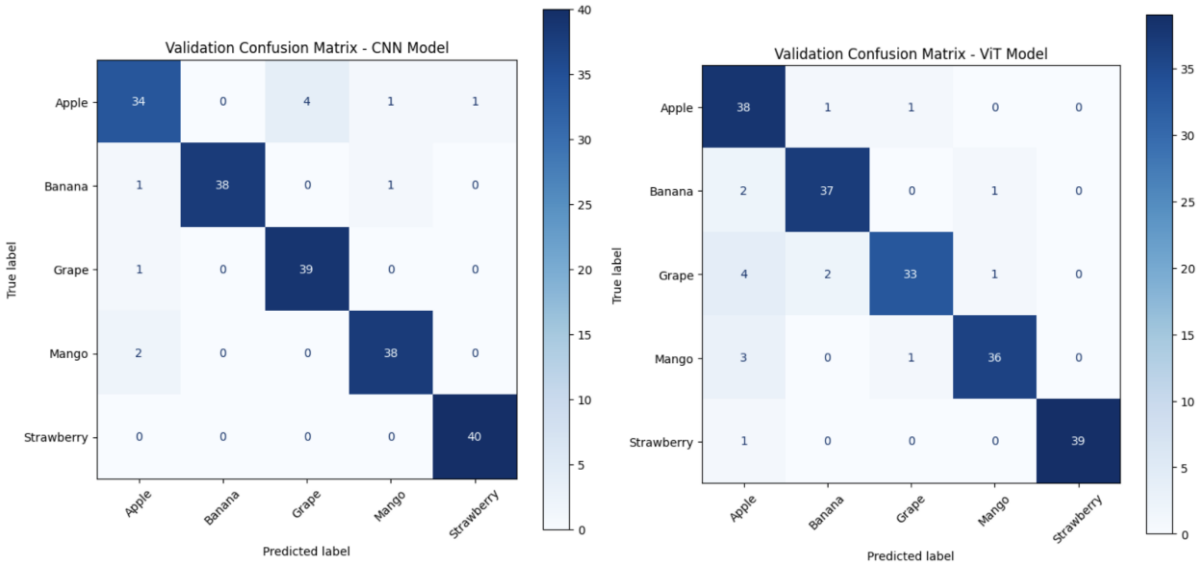

Figure 6. The confusion matrices for each of the different models

The confusion matrices for the EfficientNet-B0 (CNN) and ViT-B/16 (Vision Transformer) models provide further insights into their classification performance. Both models demonstrate strong class-wise differentiation and highlight areas where misclassifications occur.

The confusion matrices highlight that both models have strengths and weaknesses. While EfficientNet-B0 (CNN) excels in overall accuracy and precision, ViT-B/16 shows great promise for tasks that require capturing long-range dependencies and handling finer distinctions between certain fruit classes. Combining both models could mitigate these weaknesses and improve overall performance in future implementations.

3.2 Discussion

In this study, we compared the performance of EfficientNet-B0 (CNN) and ViT-B/16 (Vision Transformer) models on a fruit classification task involving five fruit classes: Apples, Bananas, Grapes, Mangoes, and Strawberries. The results reveal distinct strengths and weaknesses for each model based on accuracy, precision, recall, F1-score, and confusion matrix analysis.

The EfficientNet-B0 (CNN) model outperformed the ViT-B/16 (Vision Transformer) model in terms of overall accuracy, achieving 94% compared to 92% for the ViT model. The CNN model also demonstrated high precision and recall across most fruit classes, particularly bananas and strawberries, where near-perfect classification results were observed. The confusion matrix for the CNN model revealed minimal misclassifications, which suggests that CNN are highly effective for this task, especially when fruit classes are visually distinct.

However, while slightly less accurate overall, the ViT-B/16 (Vision Transformer) model exhibited strong performance in certain areas, notably with strawberries. This model achieved a perfect precision score of 1.00 for strawberries, indicating its ability to classify the class correctly with high reliability. In this case, the ViT model's strength likely stems from its self-attention mechanism, which excels at capturing long-range

dependencies within images. Nevertheless, the ViT model struggled with differentiating apples from grapes, as shown in the confusion matrix, highlighting the challenge of distinguishing between visually similar fruit classes.

The precision, recall, and F1-score metrics further emphasise that the EfficientNet-B0 (CNN) model excels in general accuracy across all fruit types, with high F1-scores indicating balanced performance. In contrast, the ViT-B/16 model, despite achieving lower overall accuracy, showed significant promise in handling more intricate visual features, particularly in the case of strawberries.

Both models also presented some limitations. While the EfficientNet-B0 (CNN) model performed well in most cases, it occasionally confused apples and grapes, which suggests that it may not fully capture subtle visual differences between these two classes. The ViT-B/16 model, while effective for certain fruit classes, showed inconsistent results and lower accuracy for other classes. This indicates that while ViTs have the potential for capturing global relationships in images, they may require further optimisation to perform well in fruit classification tasks where local feature extraction is essential.

3.2.1 Implications

This study has significant implications for the implementation of automated fruit classification systems in various industries, including agriculture, food processing, and retail supply chains. The findings indicate that while CNN models (EfficientNet-B0) offer higher overall accuracy and consistent performance across various fruit classes, ViT models (ViT-B/16) show promise in handling finer-grained classification tasks and capturing long-range dependencies within images. The practical implication is that the choice of model architecture should be informed by the specific characteristics of the classification task; for visually distinct differentiation, CNNs may be more efficient, whereas for more intricate details, ViTs could offer an advantage. Furthermore, the potential of hybrid or ensemble approaches combining the strengths of both models could lead to more robust, generalizable, and efficient fruit classification systems. This is particularly relevant for advancing smarter supply chains and automated food quality systems.

3.2.2 Research contribution

This research makes a significant contribution to the literature by systematically evaluating and comparing Convolutional Neural Networks (CNN) and Vision Transformers (ViT) under consistent experimental conditions for a realistic fruit classification task. Unlike many studies that focus on hybridization, this research directly analyzes the strengths and weaknesses of each architecture. Specifically, its contributions include: empirical demonstration that EfficientNet-B0 (CNN) generally outperforms ViT-B/16 in overall accuracy for fruit classification on a moderate-scale dataset; identification of cases where ViT, despite slightly lower overall accuracy, excels in classifying specific fruit classes like strawberries due to its ability to capture long-range dependencies within images ; and insights into specific misclassification patterns for both models through confusion matrix analysis, highlighting challenges in distinguishing visually similar fruits like apples and grapes. These contributions provide practical recommendations on architecture selection and optimization for practitioners working with moderate-scale image datasets in agricultural or retail settings.

3.2.3 Limitations

Despite providing valuable insights, this study has several limitations. The dataset used consists of images of five different types of fruit: Apples, Bananas, Grapes, Mangoes, and Strawberries, totaling 10,000 images, with each class containing 2,000 images. While this dataset represents real-world variability in terms of resolution, lighting conditions, angles, and backgrounds, generalizing the findings to larger and more diverse datasets or to other fruit types may require further validation. Additionally, the models were trained for 30 epochs , which might not fully explore the optimal convergence potential for ViTs, as ViTs are generally known to require larger datasets and longer training times to outperform CNNs. The study also only selected two specific models: ViT Base Patch16/224 and EfficientNet-B0, implemented using the timm library with pretrained weights. Exploring other CNN and ViT architectures, or hybrid variations, could provide a more comprehensive understanding. Finally, while data augmentation techniques such as random horizontal flipping and rotation (15 degrees) were applied for the training set , more advanced augmentation strategies like Mixup, CutMix, and RandAugment, which are known to significantly benefit ViTs, were mentioned in the introductory discussion but not detailed as to their full application within the experiments. This could impact the ViT's performance relative to the CNN.

3.2.4 Suggestions

Based on the findings and limitations of this study, several suggestions can be made for future research. First, exploring hybrid approaches that integrate the strengths of both CNNs and ViTs is highly recommended. This could involve methods such as model ensembling or shared feature encodings. Second, testing the models on larger and more diverse fruit classification datasets would help to further test the generalization capabilities of both architectures. Third, investigating the impact of more advanced data augmentation and regularization strategies, particularly those known to benefit ViTs (such as Mixup, CutMix, and RandAugment), could potentially close the performance gap between CNNs and ViTs on moderate-scale datasets. Finally, future research could focus on further fine-tuning these models and exploring their potential for broader applications in agricultural automation, food quality control, and related industries ,

including implications for deployment on edge devices, in greenhouses, or on production lines.

4. CONCLUSION

This study compares EfficientNet-B0 (CNN) and ViT-B/16 (Vision Transformer) models for fruit classification. The results show that the EfficientNet-B0 (CNN) model offers higher overall accuracy and performs well across all fruit classes, making it a strong candidate for fruit classification tasks. However, despite its slightly lower accuracy, the ViT-B/16 (Vision Transformer) model excels in capturing long-range dependencies and handles finer-grained classification tasks well, particularly in distinguishing strawberries.

The findings suggest that both models have advantages and could be combined to achieve optimal performance. The CNN model is ideal for handling well-defined, visually distinct fruit types, while the ViT model may be more effective for tasks requiring attention to complex visual cues. Future work could explore hybrid approaches that integrate the strengths of both models, such as through model ensembling or fine-tuning ViTs for fruit-specific characteristics.

The EfficientNet-B0 and ViT-B/16 are promising fruit classification approaches, each offering unique strengths. Further research is needed to fine-tune these models and explore their potential for broader applications in agricultural automation, food quality control, and related industries.

5. ACKNOWLEDGEMENT

The authors would like to thank all individuals and institutions that have supported this research. We greatly appreciate the contributions of dataset providers, particularly the publicly available fruit classification dataset on Kaggle, which was an important resource for our experiments. We would also like to express our gratitude for every form of contribution that was instrumental in realizing the results of this research.

6. AUTHOR CONTRIBUTION STATEMENT

MJ, AM, DS, and MK contributed to the conception, design, analysis, and writing of this study. All authors participated in the discussion of results, review of the manuscript, and approval of the final version for publication.

AUTHOR INFORMATION

Corresponding Authors

Malik Jawarneh, Oman College of Management and Technology, Muscat, Oman
https://orcid.org/0000-0001-6894-2756
Email: mjawarneh@omancollege.edu.om

Authors

Arief Marwanto, Universitas Islam Sultan Agung, Semarang, Indonesia
https://orcid.org/0000-0001-6873-5108
Email: arief@unissula.ac.id

Dedy Syamsuar, Universitas Bina Nusantara, Jakarta, Indonesia
https://orcid.org/0000-0002-2374-9546
Email: dedy.syamsuar@binus.ac.id

Maivi Kusnandar, Politeknik Negeri Sriwijaya, Palembang, Indonesia
https://orcid.org/0009-0007-8119-7253
Email: maivi_kusnandar_mi@polsri.ac.id

REFFERENCE

Altalak, M., Uddin, M. A., Alajmi, A., & Rizg, A. (2022). Smart Agriculture Applications Using Deep Learning Technologies: A Survey. *Applied Sciences (Switzerland)*, *12*(12). https://doi.org/10.3390/app12125919

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00444-8

Darwin, B., Dharmaraj, P., Prince, S., Popescu, D. E., & Hemanth, D. J. (2021). Recognition of bloom/yield in crop images using deep learning models for smart agriculture: A review. *Agronomy*, *11*(4), 1–22. https://doi.org/10.3390/agronomy11040646

Dewi, D. A., Kurniawan, T. B., Thinakaran, R., Batumalay, M., Habib, S., & Islam, M. (2024). Efficient Fruit Grading and Selection System Leveraging Computer Vision and Machine Learning. *Journal of Applied Data Sciences*, *5*(4), 1989–2001. https://doi.org/10.47738/jads.v5i4.443

Elharrouss, O., Akbari, Y., Almadeed, N., & Al-Maadeed, S. (2024). Backbones-review: Feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision. *Computer Science Review*, *53*, 1–23. https://doi.org/10.1016/j.cosrev.2024.100645

Espinoza, S., Aguilera, C., Rojas, L., & Campos, P. G. (2024). Analysis of Fruit Images With Deep Learning: A Systematic Literature Review and Future Directions. *IEEE Access*, *12*, 3837–3859. https://doi.org/10.1109/ACCESS.2023.3345789

Ghazal, S., Qureshi, W. S., Khan, U. S., Iqbal, J., Rashid, N., & Tiwana, M. I. (2021). Analysis of visual features and classifiers for Fruit

classification problem. *Computers and Electronics in Agriculture*, *187*, 1–9. https://doi.org/10.1016/j.compag.2021.106267

Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores. *Applied Sciences (Switzerland)*, *14*(21), 1–21. https://doi.org/10.3390/app14219863

Ismail, W. N., Alsalamah, H. A., Hassan, M. M., & Mohamed, E. (2023). AUTO-HAR: An adaptive human activity recognition framework using an automated CNN architecture design. *Heliyon*, *9*(2), e13636. https://doi.org/10.1016/j.heliyon.2023.e13636

Kanadath, A., Angel Arul Jothi, J., & Urolagin, S. (2024). CViTS-Net: A CNN-ViT Network With Skip Connections for Histopathology Image Classification. *IEEE Access*, *12*(August), 117627–117649. https://doi.org/10.1109/ACCESS.2024.3448302

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*, *54*(10), 1–30. https://doi.org/10.1145/3505244

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, *151*, 107398. https://doi.org/10.1016/j.ymssp.2020.107398

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. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences (Switzerland)*, *13*(9). https://doi.org/10.3390/app13095521

Miller, C., Portlock, T., Nyaga, D. M., & O'Sullivan, J. M. (2024). A review of model evaluation metrics for machine learning in genetics and genomics. *Frontiers in Bioinformatics*, *4*(September), 1–13. https://doi.org/10.3389/fbinf.2024.1457619

Momeny, M., Asghar, A., Arafat, M., & Kia, S. (2021). Learning-to-augment strategy using noisy and denoised data: Improving generalizability of deep CNN for the detection of COVID-19 in X-ray images. *Computers in Biology and Medicine*, *136*, 1–13. https://doi.org/10.1016/j.compbiomed.2021.104704

Oliullah, K., Islam, M. R., Babar, J. I., Quraishi, M. A. N., Rahman, M. M., Mahbub-Or-Rashid, M., & Bhuiyan, T. M. A. U. H. (2025). FruVeg_MultiNet: A hybrid deep learning-enabled IoT system for fresh fruit and vegetable identification with web interface and customized blind glasses for visually impaired individuals. *Journal of Agriculture and Food Research*, *19*(April 2024), 101623. https://doi.org/10.1016/j.jafr.2024.101623

Pandey, V. K., Srivastava, S., Dash, K. K., Singh, R., Mukarram, S. A., Kovács, B., & Harsányi, E. (2023). Machine Learning Algorithms and Fundamentals as Emerging Safety Tools in Preservation of Fruits and Vegetables: A Review. *Processes*, *11*(6), 1–17. https://doi.org/10.3390/pr11061720

Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., & Ye, Q. (2021). Conformer: Local Features Coupling Global Representations for Visual Recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 357–366. https://doi.org/10.1109/ICCV48922.2021.00042

Shabir, A., Ahmed, K. T., Mahmood, A., Garay, H., González, L. E. P., & Ashraf, I. (2025). Deep image features sensing with multilevel fusion for complex convolution neural networks & cross domain benchmarks. *PLoS ONE*, *20*(3 March), 1–42. https://doi.org/10.1371/journal.pone.0317863

Shi, X., Wang, S., Zhang, B., Ding, X., Qi, P., Qu, H., Li, N., Wu, J., & Yang, H. (2025). Advances in Object Detection and Localization Techniques for Fruit Harvesting Robots. *Agronomy*, *15*(1), 1–19. https://doi.org/10.3390/agronomy15010145

Trigka, M., & Dritsas, E. (2025). A Comprehensive Survey of Deep Learning Approaches in Image Processing. *Sensors*, *25*(2). https://doi.org/10.3390/s25020531

Wang, W., Zhang, J., Cao, Y., Shen, Y., & Tao, D. (2022). Towards Data-Efficient Detection Transformers. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13669 LNCS*, 88–105. https://doi.org/10.1007/978-3-031-20077-9_6

Zhang, W., Belcheva, V., & Ermakova, T. (2025). Interpretable Deep Learning for Diabetic Retinopathy: A Comparative Study of CNN, ViT, and Hybrid Architectures. *Computers*, *14*(5), 1–24. https://doi.org/10.3390/computers14050187