Optimized Dense Dilated Convolutional Attention Vision Transformer (ODCAViT)-Based Multi-Fruit Disease Detection

P. Sajitha
Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India
sajithap@karunya.edu.in

Diana A. Andrushia
Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India
diana@karunya.edu

N. Anand
Department of Civil Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India
nanand@karunya.edu

S. S. Suni
Department of Electronics & Communication Engineering, Ilahia College of Engineering & Technology, Ernakulam, India
suni.ss@gmail.com

Christine Dewi
Department of Information Technology, Satya Wacana Christian University, Salatiga City, Indonesia
christine.dewi@uksw.edu

Abbott Po Shun Chen
Department of Marketing and Logistics Management, Chaoyang University of Technology, Taichung City, Taiwan | Artificial Intelligence Department, Honchita Co. Ltd., Changhua County, Taiwan
chprosen@gm.cyut.edu.tw (corresponding author)

Received: 24 January 2026 | Revised: 23 March 2026, 4 April 2026, and 10 April 2026 | Accepted: 17 April 2026

 DOI: https://doi.org/10.48084/etasr.17747

ABSTRACT

The early detection of fruit diseases plays an important role in modern agriculture, as it helps enhance crop quality and significantly reduces post-harvest losses. This work proposes a unified deep-learning framework for the detection of diseases across different types of fruits such as oranges, mangoes, and pomegranates. Specifically, an Optimized Dense Dilated Convolutional Attention Vision Transformer (ODCAViT) is introduced, which incorporates densely connected dilated convolutional layers into an attention-enhanced Vision Transformer (ViT) architecture. These dense dilated convolutions preserve spatial hierarchies and allow the model to recognize subtle details and scattered disease symptoms in various fruits. In addition, the model captures long-range dependencies and global contextual information, whereas feature responses are refined by focusing on the most discriminative regions using channel attention and Squeeze-and-Excitation (SE) modules. To further improve computational efficiency, the Chaotic Puma Optimization Algorithm (CPOA) is utilized to obtain optimal parameter settings. Experimental results demonstrate that ODCAViT achieves high performance, with 99% accuracy on the pomegranate and mango datasets and 98% accuracy on the orange dataset. Overall, the proposed model demonstrates strong potential for precision agriculture and intelligent fruit disease monitoring.

Keywords-fruit disease detection; deep learning; dilated convolution; Vision Transformer (ViT)

I. INTRODUCTION

In precision agriculture applications, early detection of fruit diseases is important because it directly influences crop quality, yield, and reduction of post-harvest losses. Fruits such as oranges, mangoes, and pomegranates are cultivated worldwide for their commercial value and nutritional benefits. However, these fruits are highly vulnerable to various diseases, many of which are triggered by environmental factors such as humidity, rainfall, and sudden temperature fluctuations. If these diseases are not identified at an early stage, they can lead to severe economic damage. Therefore, developing intelligent, automated, and reliable disease detection systems is essential for timely intervention and effective crop management [1-3].

Mango (Mangifera indica), orange (Citrus sinensis), and pomegranate (Punica granatum) are commercially important fruit crops widely cultivated across tropical and subtropical regions and are highly nutritious. Despite their high economic value, these fruits are highly susceptible to several diseases that significantly reduce their quality. For example, pomegranates commonly suffer from anthracnose, bacterial blight, cercospora, and alternaria, which lead to internal decay, fruit lesions, and leaf spot formation. These infections are often triggered or intensified by environmental conditions such as high humidity, heavy rainfall, and sudden fluctuations in temperature [4].

Computer vision technologies have remarkably advanced fruit disease detection by enabling automated and accurate image-based classification. Convolutional Neural Networks (CNNs), in particular, have shown strong performance due to their ability to automatically learn meaningful visual features from large datasets. For example, a recent study introduced a Deep Ensemble Learning Model (DELM) that uses super-resolution images to identify tomato leaf diseases. The model combines GoogleNet, InceptionV3, and VGG16 with transfer learning and data augmentation, outperforming individual models and other ensembles. With an F1-score of 93.25% and an accuracy of 98%, this approach demonstrates the potential of deep learning for reliable and efficient plant disease detection [5].

Computer vision–based systems are well suited for real-world agricultural applications because they can handle variations in fruit size, shape, color, and lighting conditions. When integrated effectively, such intelligent models offer a scalable and reliable approach for identifying diseases in oranges, mangoes, and pomegranates, thereby supporting efficient and accurate disease management. Recently, authors in [6] proposed You Only Look Once (YOLO)-based deep learning models for disease detection in agriculture, which show outstanding performance. They achieved real-time performance and enhanced detection accuracy by integrating architectural optimizations to minimize complexity while maintaining accuracy, making them appropriate for agricultural applications with limited resources.

By leveraging the strengths of CNNs and Vision Transformers (ViTs), a number of studies have investigated hybrid models for fruit disease detection. ViTs offer global contextual understanding, which improves classification accuracy, whereas CNNs efficiently capture local spatial features. Compared to conventional CNNs, this hybrid approach has demonstrated better performance, particularly in challenging disease identification tasks involving minute visual differences [8].

The literature indicates that CNN-fused ViT models for fruit disease detection exhibit encouraging results; however, they have significant drawbacks, including high computational costs, substantial data requirements, and unstable training due to complex feature fusion [9]. These limitations make them less suitable for edge-based or real-time agricultural applications. Recently, researchers have suggested dilated convolution architectures combined with attention modules as a solution to these problems. These architectures efficiently capture scale-invariant features while requiring fewer parameters and lower computational power. Even with small datasets, these approaches improve spatial feature extraction while preserving robustness and efficiency.

Authors in [10] developed a method using Integration Nonlocal Means (INLM) filtering to reduce noise and improve feature clarity. A deep learning model named DCCAM-MRNet, based on ResNeXt50, was used to detect small and scattered disease spots. This model achieved 94.3% accuracy and used a lower number of parameters than the original ResNeXt50. Authors in [11] presented a deep learning approach that automatically learns and extracts features to detect plant leaf diseases more accurately. A new model called Dilated Convolutional Neural Network (DCNN) with Global Average Pooling (GAP) was proposed, which uses dilated convolutions to reduce computation and GAP to avoid overfitting. The method was tested on four types of diseases and showed a 5.49% improvement in training accuracy over traditional CNN.

Authors in [12] presented a model called Dilated Convolution Capsule Network (DCCapsNet), consisting of capsule networks combined with dilated Inception modules for better feature extraction. DCCapsNet extracts scale-invariant features and uses dynamic routing to further enhance training and reduce overfitting. Authors in [13] used an Optimal Attention Capsule Network for the detection of diseases on pomegranate fruits.

Authors in [14] proposed a Squeeze-and-Excitation Vision Transformer (SEViT) model to address challenges in non-traditional large-scale disease recognition. SEViT integrates a ResNet backbone with a channel attention module and a ViT. The model was designed to improve classification when the number of disease types is large and the visual similarity between categories is high. The experiments demonstrated that SEViT yields 88.34% classification accuracy, which is significantly better than competing models and provides a 5.15% accuracy improvement over the baseline classification model.

Recent research on disease detection and quality detection has shown great promise in capturing multi-scale contextual

features while maintaining consistent training performance through different deep learning methods [15]. The use of dilated convolution with attention mechanisms, by expanding the receptive field without significantly increasing computational complexity, successfully overcomes the spatial resolution constraints of traditional CNNs. A gap in fine-grained feature representation and class discrimination is highlighted by the literature, which shows that average classification accuracy frequently ranges between 85% and 90%. For subtle or closely related disease patterns, these models often fail to achieve high precision, even though they retain training stability.

This study proposes an Optimized Dense Dilated Convolutional Attention Vision Transformer (ODCAViT) enhanced with a Squeeze-and-Excitation (SE) fusion module. By enhancing channel-wise feature weighting and more effectively incorporating multi-scale context, this proposed architecture achieves improved feature representation and higher classification performance compared to existing methods.

II. PROPOSED METHODOLOGY

Figure 1 shows the proposed system, which uses ODCAViT for fruit disease identification. The input images are pre-processed using the Enhanced Wiener Filtering (EWF) method. Wiener filtering reduces noise while preserving important structural details, thereby improving image quality. It minimizes the mean square error between the estimated and observed intensities to reconstruct the original image. The filter adapts its behavior based on local image statistics, specifically the local mean and local variance within a predefined neighborhood surrounding each pixel. Due to this adaptive nature, it effectively preserves edges and fine details in high-variance regions while more strongly suppressing noise in low-variance (smooth) regions.

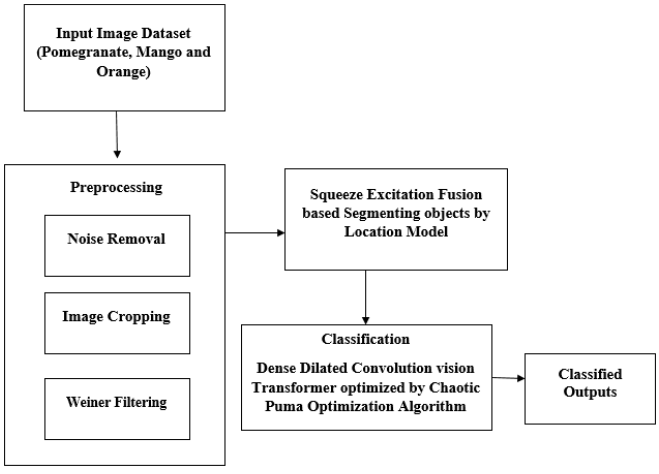

Fig. 1. Block diagram of the ODCAViT-based multi-fruit disease detection and classification system.

SE channel attention [16] is used to segment the diseased regions of the pre-processed inputs. By explicitly modeling inter-channel interdependencies, the SE module dynamically re-calibrates the feature responses at the channel level. Irrelevant features are thus suppressed, and informative parts of the fruit images—such as texture patterns or disease spots—are further emphasized. The network can better identify meaningful regions that are crucial for accurate disease detection or quality grading by incorporating SE modules into the segmentation pathway.

Dilated convolutions [17, 18], also known as atrous convolutions, are a powerful enhancement over standard convolutional operations that enable the network to exponentially expand the receptive field without adding parameters or sacrificing resolution. In an environment where disease symptoms may appear at different sizes and scales within fruits such as oranges, mangoes, and pomegranates, employing dilated convolutions allows fine-grained and large-scale features to be captured simultaneously. Dilated convolutions introduce gaps ("holes") or dilation rates between kernel elements so that the model can encompass more contextual information over a larger spatial extent, which is critical for detecting diffuse or sparsely located disease symptoms. When used in a densely connected model, dilated convolutions enable scale-invariant information to be reused across layers, producing a robust model that can classify morphological patterns of disease that are sometimes similar or identical but differ in intensity, texture, and/or extent of spread.

A. Squeeze-and-Excitation Fusion–Based Object Segmentation Using a Location-Aware Model

To capture fine-grained details in the input features, an SE attention module is integrated into the proposed framework. This module enhances feature representation by modeling inter-channel dependencies and adaptively recalibrating channel-wise responses.

The SE mechanism consists of two main stages: squeeze and excitation. In the squeeze stage, GAP is applied to each feature channel to generate a compact descriptor that captures global contextual information. In the excitation stage, this descriptor is passed through a lightweight fully connected bottleneck network with ReLU and Sigmoid activations to learn channel-wise importance weights. These weights are then applied to the original feature maps via channel-wise multiplication, enabling the network to emphasize informative features while suppressing less relevant ones. For multi-fruit disease detection, this mechanism allows the model to focus on discriminative visual cues such as subtle variations in lesion texture, color changes, and vein patterns, which may vary across different fruit types and disease categories [19].

The SE module is a channel attention mechanism designed to enhance the representational capacity of CNNs. Its key advantage lies in adaptively reweighting channel-wise feature responses, enabling the network to emphasize important features while suppressing less relevant ones [19-23].

Let the input feature map D have dimensions $L \times B \times H$, where L and B represent spatial dimensions and H denotes the number of channels. The GAP operation compresses each feature map into a single scalar value, producing a channel descriptor Z_h, defined in (1):

$$Z_h = \frac{1}{L \times B} \sum_{i=1}^{L} \sum_{j=1}^{B} X_h(i,j) \quad (1)$$

After obtaining the channel descriptors, the SE module performs feature recalibration using a gating mechanism. The descriptor vector Z_h is passed through a fully connected network to generate channel attention weights, formulated in (2):

$$S = \sigma\big(L_2 \delta(L_1 Z_h)\big) \quad (2)$$

where L_1, L_2 are learnable weight matrices, δ denotes the ReLU activation function, and σ represents the Sigmoid activation function. The output S corresponds to the learned channel-wise attention weights.

These weights are then used to rescale the original feature maps through channel-wise multiplication, as expressed in (3):

$$Y_h = S_h \times X_h \quad (3)$$

where X_h is the original feature map, S_h is the learned channel weight, and Y_h is the refined feature representation.

To further enhance spatial awareness, the SE-refined features are combined with positional encodings or location priors. These priors may be derived from gradient maps, edge information, or learned positional embeddings that encode typical object locations, such as centrally located fruits or clustered diseased regions. By integrating spatial priors with channel-wise recalibrated features, the model achieves improved segmentation accuracy and becomes more robust to cluttered backgrounds and low-contrast disease patterns.

B. Optimized Dense Dilated Convolutional Attention Vision Transformer

The proposed method is designed to integrate local and global feature learning in a balanced manner for reliable fruit disease detection. It captures both fine-grained details and global contextual information by using dense dilated convolutions, which enable efficient multi-scale feature extraction and feature reuse. In addition, the ViT module models long-range dependencies across the fruit surface, whereas the SE attention mechanism adaptively emphasizes the most informative channels, resulting in more discriminative feature representations.

The feature maps are first processed using a 1×1 convolution with a dilation rate of 2, which increases the receptive field without increasing the number of parameters. This helps the model capture both local details and broader contextual information. The output is then passed through a ReLU activation function, which introduces non-linearity by suppressing negative values and preserving positive responses, thereby improving feature learning capability. After activation, another 1×1 convolution is applied to further refine and integrate the extracted features, producing enhanced spatial and semantic representations for downstream processing.

The SE attention mechanism selectively focuses on the most relevant channels, whereas dilated convolutions capture rich spatial and multi-scale information. The resulting feature representation is then passed to a lightweight ViT module, which models contextual relationships and long-range dependencies across the image [22, 23].

This combination is particularly effective for multi-fruit and multi-disease datasets, where disease patterns vary significantly not only across different fruits but also within the same fruit class due to environmental and biological variations. Therefore, the proposed hybrid framework ensures that both local symptoms and global contextual information are effectively captured, leading to a more accurate and generalizable disease recognition system suitable for real-world agricultural applications [24-26].

ViT architecture plays an important role in extracting and interpreting visual features related to disease presence and severity on fruit surfaces. In the encoder stage, the input image is divided into fixed-size patches, which are then flattened and projected into embedding vectors. Since transformers do not inherently preserve spatial information, positional encodings are added to maintain spatial structure. These patch embeddings are then processed through multiple transformer layers using multi-head self-attention, allowing the model to learn relationships between different regions of the image. This is particularly important for detecting subtle symptoms such as color changes, texture variations, and lesion distribution.

In the decoder stage, a set of learnable query embeddings is used to attend to the encoded features through cross-attention layers. This allows the model to focus on disease-specific regions and produce accurate outputs such as class labels, bounding boxes, or segmentation masks, depending on the task. Overall, the ViT architecture is highly effective for fine-grained fruit disease detection, as the encoder captures rich global representations of the image, whereas the decoder translates them into meaningful predictions.

C. Chaotic Puma Optimization Algorithm

An improved version of the Puma Optimization Algorithm is used to enhance the performance of the proposed model. To avoid premature convergence and improve global search capability, chaotic maps are introduced into both stages of the optimization process, resulting in dynamic and non-repetitive behavior. The Chaotic Puma Optimization Algorithm (CPOA) is used to optimize internal weights and tune hyperparameters of deep learning models, including hybrid ViT architectures, CNNs, and dense dilated convolutional ViTs for fruit disease classification [27].

CPOA combines randomness with controlled chaotic behavior using logistic maps from chaos theory. This helps the algorithm avoid local optima and efficiently explore complex, high-dimensional search spaces. This is particularly important in fruit disease detection, where subtle patterns such as early infections, color changes, and surface irregularities must be identified under noisy, occluded, and uneven lighting conditions [28, 29]. The optimization process begins by generating a diverse population of candidate solutions, such as different model configurations. These solutions are then updated iteratively using chaotic search behavior inspired by puma hunting strategies. During this process, chaotic dynamics help the algorithm explore new regions while also refining

promising solutions, ensuring a balance between exploration and exploitation.

Since real-world fruit datasets are nonlinear and multi-modal, this two-stage strategy provides stable and adaptive parameter tuning. CPOA can optimize important parameters such as learning rate, number of attention heads, convolution dilation rates, layer depth, and dropout ratios. It can also support feature selection and pruning of redundant neurons, making the final model more compact and computationally efficient while maintaining high detection accuracy [30-32].

III. RESULTS AND DISCUSSION

This section presents the dataset details, implementation settings, simulation results, and overall system performance. The experimental setup is as follows. The initial learning rate of the model is set to 0.001. Training is performed for a maximum of 300 epochs with a batch size of 16. The input image size is 224 × 224, and the Leaky ReLU activation function is used. Each input image of size 224 × 224 is divided into 49 non-overlapping patches, which helps the network learn meaningful feature representations effectively. The proposed implementation is carried out on a system with an Intel® Core™ i5-3570 CPU running at 3.40 GHz, 8 GB of RAM (7.89 GB usable), and a 64-bit operating system.

The CPOA is used to optimize the learning rate, which is an important hyperparameter affecting model convergence and performance. The optimization is performed within a search space of $[10^{-4}, 10^{-2}]$. An initial population of 10 candidate learning rates is randomly generated. These candidates are updated over 20 iterations using a sinusoidal chaotic function, which improves the balance between exploration and exploitation during the search process. In each iteration, candidate solutions are modified using a chaos-driven update strategy and evaluated using a fitness function based on the final training loss of a neural network trained for 3 epochs using the Adam optimizer. The best solution is selected by minimizing the mean squared error loss. Compared to a fixed learning rate of 0.001, the CPOA-based adaptive approach identifies a more suitable learning rate, leading to better convergence and lower training loss. The optimized learning rate obtained by CPOA is 0.00073, which improves training stability and overall model performance.

A. Dataset Description

The orange dataset consists of 14 folders and a total of 1,090 images [33]. The pomegranate dataset, available at [34] and further discussed in [35], includes five folders containing 5,099 labeled and categorized images of healthy and diseased pomegranate fruits. The Mango FruitDDS dataset, available at [36] and further discussed in [37], contains 1,700 images in 224 × 224 pixel format and covers mango diseases such as anthracnose, black mold rot, stem end rot, and alternaria. This dataset also includes a healthy fruit category. For model training and evaluation, each dataset is divided into training and validation sets using an 80:20 ratio.

B. Performance Evaluation and Comparative Analysis

The performance of the proposed model is evaluated using standard performance metrics, and a comparative analysis is conducted to assess its effectiveness against existing methods. The proposed approach is compared with state-of-the-art deep learning models, including YOLOv8, GoogLeNet, MobileNet, EfficientNet, and VGG16 [35, 38-40], using three datasets comprising pomegranate, orange, and mango fruit images.

YOLOv8 is a real-time object detection model that provides a strong balance between accuracy and speed, making it suitable for tasks requiring simultaneous object localization and classification. GoogLeNet (Inception v1) introduces the Inception module, which enables multi-scale feature extraction within a single layer, allowing the network to capture diverse spatial patterns efficiently. MobileNet is a lightweight architecture designed for resource-constrained environments, using depthwise separable convolutions to significantly reduce computational cost and model parameters. EfficientNet improves performance by applying a compound scaling strategy that uniformly scales network depth, width, and input resolution in a balanced manner. VGG16 is a deep convolutional network characterized by its simple and uniform architecture, which uses stacked 3×3 convolutional layers to learn hierarchical feature representations.

1) Confusion Matrix Analysis

The performance of the proposed classifier is evaluated using confusion matrices, which show the number of correctly and incorrectly classified samples for each class. The results indicate that most predictions lie on the diagonal, confirming strong classification performance with only a few misclassifications across all datasets.

For the pomegranate dataset, five classes are considered. In the bacterial blight class, 40 images are correctly classified, whereas one image is misclassified as healthy. In the alternaria class, one image is incorrectly classified as healthy. For the normal class, 40 out of 41 images are correctly classified, whereas one image is misclassified as cercospora. Overall, alternaria, anthracnose, and bacterial blight achieve 100% classification accuracy, whereas cercospora and healthy classes show minor confusion. The results indicate high classification performance with very limited misclassification. The confusion matrix for the pomegranate dataset is shown in Figure 2.

For the orange dataset, four classes are evaluated. In the black spot class, 36 images are correctly classified, whereas one image is misclassified as fresh and another as canker. In the canker class, one image is incorrectly classified as greening. For the normal class, 53 out of 54 images are correctly classified, whereas one image is misclassified as canker. These results indicate consistent performance across all classes, with only a few misclassifications among closely related categories.

For the mango dataset, five classes are evaluated. In the alternaria class, 34 images are correctly classified. In the black mold rot class, one image is misclassified as anthracnose. In the normal class, 40 out of 41 images are correctly classified, whereas one image is misclassified as black mold rot. Alternaria and stem end rot achieve perfect classification, indicating strong discriminative capability across different disease patterns.

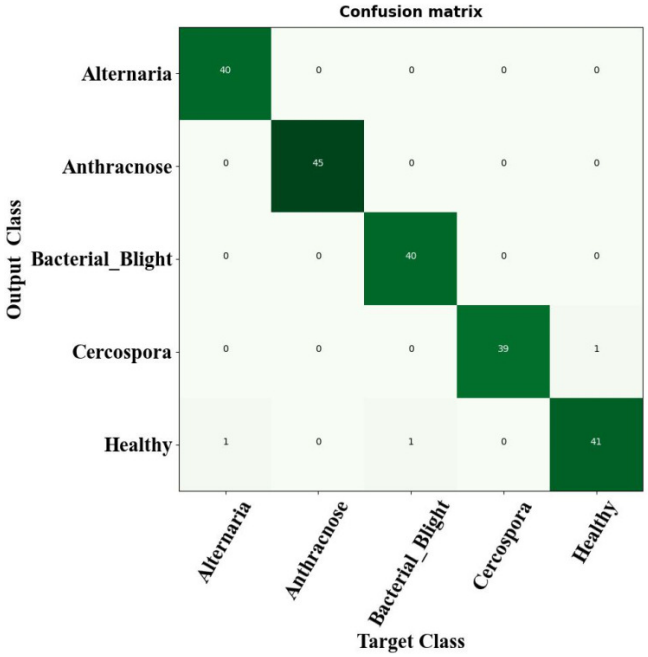

Fig. 2. Confusion matrix for the pomegranate dataset.

Overall, the confusion matrices across all three datasets show that most predictions lie on the diagonal, indicating high classification accuracy. The proposed model effectively distinguishes visually similar disease classes with minimal confusion, demonstrating strong generalization for multi-fruit disease detection.

2) Accuracy and Loss Curves

Figure 3 shows the accuracy curves for the proposed model on the orange dataset, including both training and testing accuracy. The curves illustrate the learning progress of the model across epochs. The accuracy gradually increases and stabilizes after approximately 250 epochs, indicating that the model has reached convergence and maintains stable performance. A similar trend is observed for the mango dataset, whereas for the pomegranate dataset, the accuracy stabilizes between the 200th and 300th epochs.

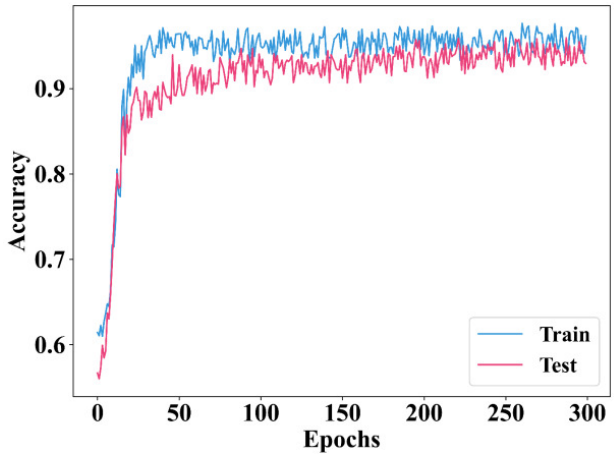

Fig. 3. Accuracy curves for the orange dataset.

Figure 4 shows the training and testing loss curves for the mango dataset using the proposed method. The results show that the loss decreases steadily during training. A significant reduction in loss is observed around the 150th epoch, after which the curve stabilizes, indicating effective learning and improved model optimization. Similar behavior is observed for the orange and pomegranate datasets.

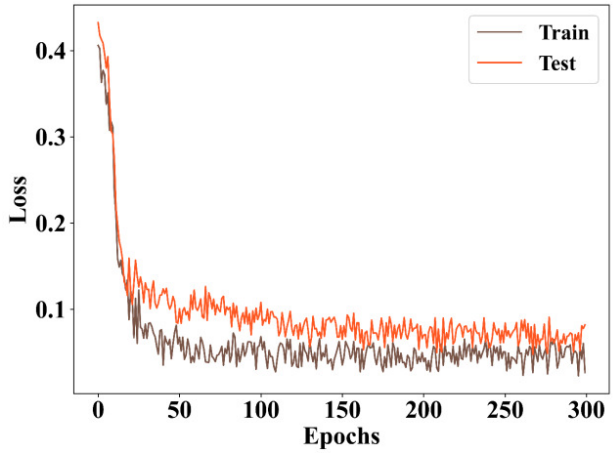

Fig. 4. Loss curves for the mango dataset.

3) Receiver Operating Characteristic Analysis

The Receiver Operating Characteristic (ROC) curve is used to evaluate the classification performance of the proposed model. Figure 5 presents the ROC curves for the orange dataset. The curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different classification thresholds, providing a comprehensive measure of class separability.

As observed, the proposed model achieves a higher ROC curve compared to the other models, consistently remaining closer to the top-left region of the plot. This indicates better classification performance, higher sensitivity, and improved discriminative ability across all threshold values.

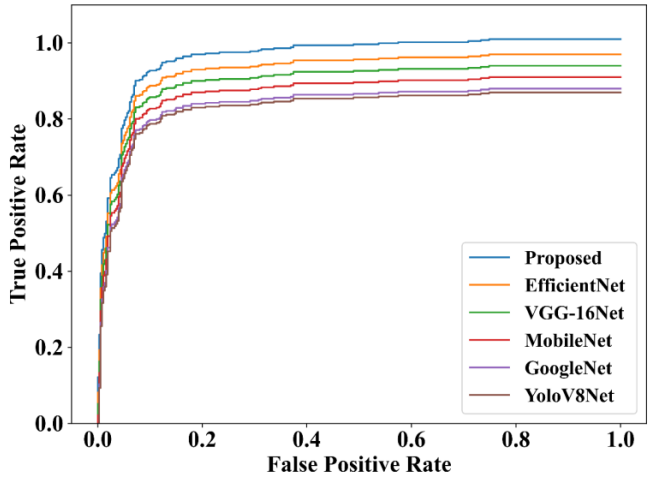

Fig. 5. ROC curve analysis for the orange dataset.

4) Performance Metrics

Tables I–III present the performance comparison of the proposed model with state-of-the-art deep learning approaches for the pomegranate, orange, and mango datasets. The results demonstrate that the proposed method consistently achieves superior performance across all evaluation metrics, including accuracy, precision, recall, and F1-score.

The improved performance is mainly due to the use of CPOA, which reduces the loss function during training by optimizing key hyperparameters. This enhances the model's ability to achieve stable convergence and improves overall classification accuracy.

For the pomegranate dataset, the proposed model achieves an accuracy of 99.5%, with precision of 99.0%, recall of 99.5%, and F1-score of 99.0%. These results are higher than all compared models, including EfficientNet, VGG16, MobileNet, GoogLeNet, and YOLOv8Net, as shown in Table I.

TABLE I. PERFORMANCE COMPARISON USING THE POMEGRANATE DATASET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Proposed	99.5	99.0	99.5	99.0
EfficientNet	97.5	95.5	95.5	95.5
VGG16	95.0	94.0	94.0	94.0
MobileNet	92.0	90.5	91.0	91.3
GoogLeNet	90.2	90.2	90.0	90.0
YOLOv8	88.5	87.1	87.7	87.3

For the orange dataset, the proposed method achieves 98.5% accuracy, 98.7% precision, 98.15% recall, and 96.2% F1-score. These results indicate strong and balanced performance across all evaluation metrics and outperform all baseline models, as shown in Table II.

TABLE II. PERFORMANCE COMPARISON USING THE ORANGE DATASET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Proposed	98.5	98.7	98.15	96.2
EfficientNet	95.5	95.0	94.5	94.5
VGG16	93.0	92.0	91.5	91.5
MobileNet	92.5	91.0	89.0	92.5
GoogLeNet	88.0	87.5	87.0	87.0
YOLOv8	86.0	87.7	86.5	85.0

For the mango dataset, the proposed model achieves 99.4% accuracy, 99.1% precision, 99.1% recall, and 98.5% F1-score. As shown in Table III, the proposed method consistently outperforms EfficientNet, VGG16, MobileNet, GoogLeNet, and YOLOv8Net.

The proposed method's obtained accuracy for the pomegranate dataset is 99%. This method yields a precision rate of 99 %. The suggested approach gives the F1-score of 99.5% for the pomegranate dataset. The proposed classifier achieved recall of 99.5%.

TABLE III. PERFORMANCE COMPARISON USING THE MANGO DATASET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Proposed	99.4	99.1	99.1	98.5
EfficientNet	97.5	95.5	96.2	95.5
VGG16	94.5	94.0	94.1	94.0
MobileNet	92.0	91.0	91.2	91.0
GoogLeNet	90.0	89.0	88.5	88.5
YOLOv8	87.5	87.0	87.0	87.0

C. Discussion

Overall, the proposed model shows superior and stable performance across all datasets, with consistently higher and more balanced metric values compared to existing methods. The results also indicate low variance across metrics, suggesting stable learning behavior.

The strong performance is due to the integrated design of the proposed architecture. DenseNet and dilated convolutions improve multi-scale feature extraction, whereas the SE module enhances channel-wise feature discrimination, especially for low-contrast disease patterns. The ViT captures long-range dependencies that are difficult for CNN-based models to learn.

In addition, the CPOA is used only for hyperparameter tuning, which improves training stability and prevents overfitting without increasing model complexity. To ensure robustness, regularization techniques, data augmentation, early stopping, and cross-validation were applied. An ablation study further confirms that each component contributes positively to overall performance.

The consistent improvements across all datasets demonstrate that the proposed model is well-suited for fine-grained fruit disease classification tasks without showing signs of overfitting.

IV. CONCLUSION

An Optimized Dense Dilated Convolutional Attention Vision Transformer (ODCAViT) is proposed for robust and accurate detection and classification of multiple fruit diseases across different fruit types. The proposed architecture effectively integrates dilated convolutions for multi-scale feature extraction, dense connections for improved feature reuse, and Squeeze-and-Excitation (SE) attention modules for adaptive channel-wise feature recalibration. These components collectively enhance the model's ability to capture both local and global features, which is essential for identifying diverse disease patterns in fruits such as mango, orange, and pomegranate.

By combining these enhanced convolutional representations with the global dependency modeling capability of a lightweight Vision Transformer (ViT), the proposed approach addresses the limitations of conventional Convolutional Neural Networks (CNNs) and standalone transformer-based models in agricultural image analysis. The experimental results demonstrate strong performance, achieving accuracies of 99% for the pomegranate and mango datasets and 98% for the

orange dataset. These results confirm that the proposed ODCAViT model is both robust and generalizable for fruit disease detection tasks.

Despite the strong performance, some limitations are observed. The alignment between convolutional feature representations and transformer embeddings is not always optimal, which may lead to variations in attention maps and inconsistent feature importance across different samples. This can reduce interpretability and make it difficult to clearly identify the contribution of individual components to specific predictions.

Future work will focus on extending the model with additional modules for disease severity estimation and temporal disease tracking. In addition, integrating Explainable Artificial Intelligence (XAI) techniques can improve model interpretability and transparency, making the system more reliable and acceptable for practical use in agriculture.

DECLARATION OF COMPETING INTERESTS

The authors declare no conflicts of interest.

ACKNOWLEDGMENT

This work was supported by Honchita Co., Ltd. (Uniform Number: 60304896), Changhua, Taiwan.

DATA AVAILABILITY

The datasets used in this study are publicly available from the sources cited in references [33-37].

REFERENCES

[1] A. Tempelaere *et al.*, "An introduction to artificial intelligence in machine vision for postharvest detection of disorders in horticultural products," *Postharvest Biology and Technology*, vol. 206, Dec. 2023, Art. no. 112576, https://doi.org/10.1016/j.postharvbio.2023.112576.

[2] P. Sajitha, A. D. Andrushia, N. Anand, and M. Z. Naser, "A review on machine learning and deep learning image-based plant disease classification for industrial farming systems," *Journal of Industrial Information Integration*, vol. 38, Mar. 2024, Art. no. 100572, https://doi.org/10.1016/j.jii.2024.100572.

[3] R. Akhter and S. A. Sofi, "Precision agriculture using IoT data analytics and machine learning," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, Part B, pp. 5602–5618, Sept. 2022, https://doi.org/10.1016/j.jksuci.2021.05.013.

[4] D. B. Ahuja and C. Chattopadhyay, "Pests of Fruit Trees (Citrus, Banana, Mango, Pomegranate and Sapota): E-Pest Surveillance and Pest Management Advisory," ICAR-National Research Centre for Integrated Pest Management, New Delhi and State Department of Horticulture, Commissionerate of Agriculture, Pune, India, 2015.

[5] P. KAUR *et al.*, "DELM: Deep Ensemble Learning Model for Multiclass Classification of Super-Resolution Leaf Disease Images," *Turkish Journal of Agriculture and Forestry*, vol. 47, no. 5, pp. 727–745, Oct. 2023, https://doi.org/10.55730/1300-011X.3123.

[6] H. M. Zayani *et al.*, "Deep Learning for Tomato Disease Detection with YOLOv8," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13584–13591, Apr. 2024, https://doi.org/10.48084/etasr.7064.

[7] Z. Chen, J. Feng, K. Zhu, Z. Yang, Y. Wang, and M. Ren, "YOLOv8-ACCW: Lightweight Grape Leaf Disease Detection Method Based on Improved YOLOv8," *IEEE Access*, vol. 12, pp. 123595–123608, 2024, https://doi.org/10.1109/ACCESS.2024.3453379.

[8] K. Aghamohammadesmaeilketabforoosh, S. Nikan, G. Antonini, and J. M. Pearce, "Optimizing Strawberry Disease and Quality Detection with Vision Transformers and Attention-Based Convolutional Neural Networks," *Foods*, vol. 13, no. 12, June 2024, Art. no. 1869, https://doi.org/10.3390/foods13121869.

[9] A. Agarwal, A. Sarkar, and A. K. Dubey, "Computer Vision-Based Fruit Disease Detection and Classification," in *Smart Innovations in Communication and Computational Sciences: Proceedings of ICSICCS-2018*, Burla, Odisha, India, 2018, pp. 105–115, https://doi.org/10.1007/978-981-13-2414-7_11.

[10] Y. Liu, Y. Hu, W. Cai, G. Zhou, J. Zhan, and L. Li, "DCCAM-MRNet: Mixed Residual Connection Network with Dilated Convolution and Coordinate Attention Mechanism for Tomato Disease Identification," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, Apr. 2022, Art. no. 4848425, https://doi.org/10.1155/2022/4848425.

[11] S. Zhang, S. Zhang, C. Zhang, X. Wang, and Y. Shi, "Cucumber leaf disease identification with global pooling dilated convolutional neural network," *Computers and Electronics in Agriculture*, vol. 162, pp. 422–430, July 2019, https://doi.org/10.1016/j.compag.2019.03.012.

[12] C. Xu, X. Wang, and S. Zhang, "Dilated convolution capsule network for apple leaf disease identification," *Frontiers in Plant Science*, vol. 13, Nov. 2022, Art. no. 1002312, https://doi.org/10.3389/fpls.2022.1002312.

[13] P. Sajitha, A. Diana Andrushia, N. Anand, M. Z. Naser, and E. Lubloy, "A deep learning approach to detect diseases in pomegranate fruits via hybrid optimal attention capsule network," *Ecological Informatics*, vol. 84, Dec. 2024, Art. no. 102859, https://doi.org/10.1016/j.ecoinf.2024.102859.

[14] Q. Zeng, L. Niu, S. Wang, and W. Ni, "SEViT: a large-scale and fine-grained plant disease classification model based on transformer and attention convolution," *Multimedia Systems*, vol. 29, no. 3, pp. 1001–1010, June 2023, https://doi.org/10.1007/s00530-022-01034-1.

[15] S. Seilov, A. Nurzhaubayev, M. Baideldinov, B. Zhursinbek, M. Ashimgaliyev, and A. Zhumadillayeva, "Hybrid Multi-Scale Neural Network with Attention-Based Fusion for Fruit Crop Disease Identification," *Journal of Imaging*, vol. 11, no. 12, Dec. 2025, Art. no. 440, https://doi.org/10.3390/jimaging11120440.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132–7141, https://doi.org/10.1109/CVPR.2018.00745.

[17] G. Lin, Q. Wu, L. Qiu, and X. Huang, "Image super-resolution using a dilated convolutional neural network," *Neurocomputing*, vol. 275, pp. 1219–1230, Jan. 2018, https://doi.org/10.1016/j.neucom.2017.09.062.

[18] S. Senthil Pandi, A. Senthilselvi, J. Gitanjali, K. ArivuSelvan, J. Gopal, and J. Vellingiri, "Rice plant disease classification using dilated convolutional neural network with global average pooling," *Ecological Modelling*, vol. 474, Dec. 2022, Art. no. 110166, https://doi.org/10.1016/j.ecolmodel.2022.110166.

[19] S. Verma, A. Chug, R. P. Singh, A. P. Singh, and D. Singh, "SE-CapsNet: Automated evaluation of plant disease severity based on feature extraction through Squeeze and Excitation (SE) networks and Capsule networks," *Kuwait Journal of Science*, vol. 49, no. 1, pp. 1–31, 2022, https://doi.org/10.48129/kjs.v49i1.10586.

[20] A. Y. Ashurov *et al.*, "Enhancing plant disease detection through deep learning: a Depthwise CNN with squeeze and excitation integration and residual skip connections," *Frontiers in Plant Science*, vol. 15, Jan. 2025, Art. no. 1505857, https://doi.org/10.3389/fpls.2024.1505857.

[21] B. N. Naik, R. Malmathanraj, and P. Palanisamy, "Detection and classification of chilli leaf disease using a squeeze-and-excitation-based CNN model," *Ecological Informatics*, vol. 69, July 2022, Art. no. 101663, https://doi.org/10.1016/j.ecoinf.2022.101663.

[22] L. R. Ali, S. A. Jebur, M. M. Jahefer, A. K. Nawar, and Z. S. Mahdi, "Integrating Squeeze-and-Excitation Network with Pretrained CNN Models for Accurate Plant Disease Detection," *International Journal of Electrical and Computer Engineering Systems*, vol. 16, no. 8, pp. 621–632, Sept. 2025, https://doi.org/10.32985/ijeces.16.8.5.

[23] Y. Pratama, E. Rasywir, and A. Siswanto, "Optimized Non-Overlapping Multi-Object Segmentation for Palm Oil Images Using FCN with Squeeze-and-Excitation and Attention Mechanisms," *Scientific Journal of Informatics*, vol. 12, no. 1, pp. 87–98, May 2025, https://doi.org/10.15294/sji.v12i1.22212.

[24] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting Objects by Locations," in *Computer Vision – ECCV 2020: 16th European Conference*, Glasgow, UK, 2020, pp. 649–665, https://doi.org/10.1007/978-3-030-58523-5_38.

[25] T.-Y. Hsiao, Y.-C. Chang, H.-H. Chou, and C.-T. Chiu, "Filter-based deep-compression with global average pooling for convolutional networks," *Journal of Systems Architecture*, vol. 95, pp. 9–18, May 2019, https://doi.org/10.1016/j.sysarc.2019.02.008.

[26] Y. Wang, G. Wang, C. Chen, and Z. Pan, "Multi-scale dilated convolution of convolutional neural network for image denoising," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 19945–19960, July 2019, https://doi.org/10.1007/s11042-019-7377-y.

[27] C. Xu, C. Yu, S. Zhang, and X. Wang, "Multi-Scale Convolution-Capsule Network for Crop Insect Pest Recognition," *Electronics*, vol. 11, no. 10, May 2022, Art. no. 1630, https://doi.org/10.3390/electronics11101630.

[28] K. Han *et al.*, "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, Jan. 2023, https://doi.org/10.1109/TPAMI.2022.3152247.

[29] B. Abdollahzadeh *et al.*, "Puma optimizer (PO): a novel metaheuristic optimization algorithm and its application in machine learning," *Cluster Computing*, vol. 27, no. 4, pp. 5235–5283, July 2024, https://doi.org/10.1007/s10586-023-04221-5.

[30] G. Kaur and S. Arora, "Chaotic whale optimization algorithm," *Journal of Computational Design and Engineering*, vol. 5, no. 3, pp. 275–284, July 2018, https://doi.org/10.1016/j.jcde.2017.12.006.

[31] G. I. Sayed, A. Darwish, and A. E. Hassanien, "A New Chaotic Whale Optimization Algorithm for Features Selection," *Journal of Classification*, vol. 35, no. 2, pp. 300–344, July 2018, https://doi.org/10.1007/s00357-018-9261-2.

[32] M. Kmich *et al.*, "Chaotic Puma Optimizer Algorithm for controlling wheeled mobile robots," *Engineering Science and Technology, an International Journal*, vol. 63, Mar. 2025, Art. no. 101982, https://doi.org/10.1016/j.jestch.2025.101982.

[33] "Orange Quality Analysis Dataset." Kaggle. [Online]. Available: https://www.kaggle.com/datasets/shruthiiiee/orange-quality.

[34] P. B and D. H. R, "Pomegranate Fruit Diseases Dataset for Deep Learning Models." Mendeley Data, Nov. 15, 2023, https://doi.org/10.17632/b6s2rkpmvh.1.

[35] P. B. and H. R., "A comprehensive standardized dataset of numerous pomegranate fruit diseases for deep learning," *Data in Brief*, vol. 54, June 2024, Art. no. 110284, https://doi.org/10.1016/j.dib.2024.110284.

[36] "MangoFruitDDS." Kaggle. [Online]. Available: https://www.kaggle.com/datasets/warcoder/mangofruitdds.

[37] D. Faye, I. Diop, N. Mbaye, D. Dione, and M. M. Diedhiou, "MangoFruitDDS: A Standard Mango Fruit Diseases Dataset Made in Africa," in *Advanced Research in Technologies, Information, Innovation and Sustainability: Third International Conference*, Madrid, Spain, 2023, pp. 237–250, https://doi.org/10.1007/978-3-031-48930-3_18.

[38] G. T. Yehulu *et al.*, "Mango fruit disease detection and classification using MobileNetV3_large model," *Scientific African*, vol. 30, Dec. 2025, Art. no. e03061, https://doi.org/10.1016/j.sciaf.2025.e03061.

[39] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 11531–11539, https://doi.org/10.1109/CVPR42600.2020.01155.

[40] H. Wang, S. Qiu, H. Ye, and X. Liao, "A Plant Disease Classification Algorithm Based on Attention MobileNet V2," *Algorithms*, vol. 16, no. 9, Sept. 2023, Art. no. 442, https://doi.org/10.3390/a16090442.