Benchmarking Deep Learning Models for Visual Classification and Segmentation of Horticultural Commodities

Fuzy Yustika Manik*, Syahril Efendi, Jos Timanta Tarigan, Maya Silvi Lydia
Department of Computer Science-Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

***Abstract*—Recent advances in computer vision have enabled new approaches for automated quality assessment of tropical fruits, where accurate classification and segmentation are essential for postharvest inspection. A major challenge lies in identifying deep learning architectures that achieve high accuracy while remaining computationally efficient for potential edge-based deployment. This study benchmarks three Convolutional Neural Network (CNN) models for classification (VGG16, ResNet50, and EfficientNet-B0) and two encoder–decoder models for segmentation (U-Net and DeepLabV3+) using annotated pineapple and strawberry image datasets. A 5-fold cross-validation strategy was applied to ensure statistical robustness, with evaluation metrics including accuracy, precision, recall, F1-score, Intersection over Union (IoU), and Dice coefficient. Statistical significance was verified using the Friedman and Wilcoxon signed-rank tests ($\alpha = 0.05$ and 0.01). EfficientNet-B0 achieved the best classification results with average accuracies of 91.4% (strawberry) and 90.7% (pineapple), significantly outperforming ResNet50 and VGG16 ($p < 0.01$). For segmentation, DeepLabV3+ obtained the highest performance with mean IoU values of 91.7% and 90.8% and Dice coefficients above 92%, indicating precise boundary delineation of ripe and defective regions. Computational efficiency analysis further showed that EfficientNet-B0 had the lowest inference time (0.026 s) and smallest model size (20.4 MB), making it ideal for real-time or embedded applications. Visual analysis confirmed that DeepLabV3+ maintained robustness at fruit boundaries, though minor misclassifications were observed. This benchmarking highlights the combination of EfficientNet-B0 and DeepLabV3+ as a reliable baseline for deep learning-based fruit quality assessment.**

***Keywords*—Fruit quality assessment; classification; segmentation; EfficientNet-B0; DeepLabV3+; AISAM-CSNet**

I. Introduction

Advances in computer vision and deep learning have significantly impacted precision agriculture, particularly in automated fruit quality assessment and post-harvest monitoring [1], [2]. One of the most crucial visual tasks in this domain is image-based classification and segmentation, which enables sorting, grading, and intelligent decision-making for horticultural commodities such as strawberries and pineapples. These visual tasks are particularly challenging due to variations in lighting, occlusion, background complexity, and the diversity of shapes and colors within a single fruit class [3].

The application of deep learning in precision agriculture systems has received significant attention in recent years. In particular, Convolutional Neural Network (CNN)-based models such as VGG16, ResNet50, and EfficientNet-B0 have demonstrated high performance in fruit image classification tasks, ranging from type determination and ripeness to visual defect detection [4], [5].

A study by Hasan et al. [6] demonstrated that ResNet50 can accurately classify fruit in real-world scenarios. Meanwhile, VGG16 is known to be a stable model and is often used as a baseline in many visual classification experiments [7]. EfficientNet-B0, with its systematic scaling approach, offers a balance between accuracy and computational efficiency, making it suitable for edge-based applications such as field devices [8]. However, most of these models are trained and tested under relatively controlled conditions or using datasets with low levels of variation. In real-world horticultural scenarios, fruit images are typically acquired under non-uniform lighting, varying backgrounds, and inconsistent fruit shapes and sizes [8].

U-Net has become a primary choice for fruit image segmentation tasks due to its symmetric encoder-decoder design, which effectively preserves spatial information [9]. On the other hand, DeepLabV3+ uses atrous convolution and atrous spatial pyramid pooling (ASPP) to capture contextual information at multiple scales, proving effective in segmenting objects in complex backgrounds [10], [11]. Mo et al. [12] successfully implemented MobileNetV2-based DeepLabV3+ to detect sugar apple ripeness accurately.

However, most of these models were trained and tested under relatively controlled conditions or using datasets with low levels of variation. In real-world horticultural scenarios, fruit images are typically acquired with non-uniform lighting, varying backgrounds, and inconsistent fruit shape and size [13]. These factors reduce the generalization ability of standard CNN models. Therefore, several approaches such as data augmentation [14], domain adaptation [15], and synthetic training [16] have been developed to improve the generalization ability of CNN models to real-world conditions.

Furthermore, most previous studies focus on a single task (classification or segmentation) and often use datasets under ideal conditions. Few studies evaluate multiple CNN architectures simultaneously on both tasks on authentic tropical fruit images such as strawberries and pineapples. Therefore,

*Corresponding author.

this study aims to systematically assess the performance of VGG16, ResNet50, and EfficientNet-B0 models for classification and U-Net and DeepLabV3+ for tropical fruit image segmentation. The evaluation uses metrics such as accuracy, precision, recall, F1-score (classification), and IoU and Dice coefficient (segmentation) to provide a scientific basis for selecting the optimal model for image-based fruit classification and segmentation systems.

This benchmarking study aims to highlight the strengths and weaknesses of current deep learning approaches and provide an empirical basis for developing more adaptive and multi-task architectures. In addition, the findings of this study will demonstrate how model selection and architectural complexity directly influence accuracy, segmentation quality, and computational efficiency under real-world postharvest conditions. These insights contribute to the development of practical, lightweight, and adaptive deep learning models that can be effectively applied to agricultural automation and quality monitoring of tropical fruits in real-world scenarios. Based on our results, we also briefly discuss the potential of emerging immune-inspired and multi-agent approaches such as AISAM-CSNet, which will be elaborated in detail in our forthcoming publication.

II. Related Work

Rapid computer vision and deep learning developments have driven the application of CNN models in various sectors, including precision agriculture. In fruit image processing, two main tasks, quality classification and fruit object segmentation, are key for automated post-harvest systems. Recent studies have utilized modern CNN models to detect fruit types and ripeness and to separate fruit objects from complex backgrounds. CNN models such as VGG16, ResNet50, and EfficientNet-B0 are widely used in fruit image classification. VGG16 is a classic architecture often used as a baseline due to its stability despite the large number of parameters [7]. Sudars et al. [6] conducted a comprehensive review of the application of CNNs to fruit quality classification and positioned VGG16 as one of the standard architectures used in laboratory scenarios.

ResNet50, which introduces residual learning, effectively addresses degradation issues in deep networks [3]. Hasan et al. [6] showed that ResNet50 can maintain fruit classification accuracy in natural lighting. Arif et al. [17] compared ResNet50 with DenseNet and EfficientNet in orange classification, with ResNet50 outperforming in complex background conditions.

EfficientNet-B0 introduces a compound scaling approach to balance accuracy and efficiency [8]. Li et al. [5] demonstrated that EfficientNet-B0 suits fruit classification on edge devices. Wagle et al. [18]'s research corroborates this finding by showing the high computational efficiency of EfficientNet in apple and tomato classification. Reyes et al. [19] also demonstrated that EfficientNet is effective in fine-grained classification of tropical fruits. Furthermore, DenseNet [20], InceptionV3, and MobileNetV2 [21] were also evaluated in fruit classification, but their performance was often lower in open field conditions. A study by Rauf et al. [22] confirmed that ResNet and EfficientNet provide the best trade-off between accuracy and inference time compared to other architectures.

Meanwhile, fruit object segmentation requires precise object boundary detection and separation of the fruit from the background. The U-Net model is widely recognized for this task because it preserves spatial information [9]. Fang et al. [23] developed a U-Net with attention gating for accurate strawberry segmentation in open fields. Jamil et al. [24] combined a squeeze-and-excitation block to improve mango fruit detection in RGB images.

DeepLabV3+ utilizes atrous convolution and atrous spatial pyramid pooling (ASPP) to process multiscale features. Research by Zhao et al. [10] demonstrated the superiority of DeepLabV3+ in detecting mangoes under occlusion. Mo et al. [12] implemented MobileNetV2-based DeepLabV3+ to detect the ripeness of sugar apples and achieved high accuracy. Other models such as Mask R-CNN [25], HRNet [26], and SegNet [27] have also been tried for fruit segmentation, but their complexity and high computational requirements are obstacles in real-time applications. Milioto et al. [28] developed real-time semantic segmentation for crops and weeds using a CNN optimized for agricultural robots.

One of the main challenges in implementing CNN in fruit classification and segmentation systems is its ability to generalize real-world images. Most studies use datasets with clean backgrounds and ideal lighting [29]. Xu et al. [30] reported a 30% decrease in accuracy when apple classification models were trained in the laboratory and tested in the open field. To address this issue, data augmentation [14], domain adaptation [15], and synthetic image-based training [16] approaches have been proposed. Shorten and Khoshgoftaar [14] showed that augmentation can improve the robustness of CNN models to complex background conditions. Chen et al. [15] evaluated domain adaptation to transfer models from laboratory to field data. Meanwhile, synthetic training has enriched the variety of training images, as evidenced by Rahnemoonfar and Sheppard [16].

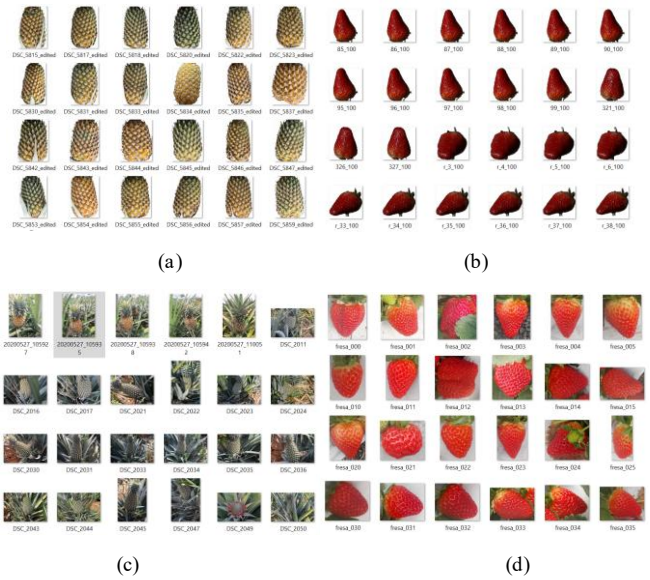
Although various CNN models have proven effective for fruit classification and segmentation separately, comprehensive studies that evaluate multiple CNN architectures across both tasks under real-world conditions remain limited. This research presents the first benchmarking study that simultaneously evaluates CNN-based classification and segmentation models (VGG16, ResNet50, EfficientNet-B0, U-Net, and DeepLabV3+) on tropical fruit images (strawberries and pineapples) in realistic postharvest scenarios. The study contributes a novel understanding of the relationship between architectural complexity, adaptability, and segmentation precision. The benchmarking results reveal a balanced trade-off between classification and segmentation accuracy as well as computational efficiency, forming the foundation for the development of AISAM-CSNet as a lightweight, adaptive, and multitask model for agricultural automation.

III. Methodology

The research stages include data acquisition and preprocessing, CNN model architecture development (classification and segmentation), implementation, and performance evaluation.

A. Dataset and Data Acquisition

This study uses strawberry images for non-climacteric fruits and pineapple images for climacteric fruits obtained from two main sources: 1) primary: a collection of field images obtained directly using a digital camera, and 2) secondary: public datasets such as Kaggel [31-32]. The image dataset was collected in real-world environments with natural variations in lighting and diverse surface textures. Manual annotation was performed based on the local postharvest conditions of tropical horticultural products. Examples of the images used can be seen in Fig. 1 and Fig. 2 below:



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

Fig. 1. Secondary dataset: (a) Pineapple with white background, (b) Strawberry with white background, (c) Pineapple with natural background, (d) Strawberry with natural background.

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

Fig. 2. Secondary dataset: (a) Strawberry with white background, (b) Strawberry with natural background, (c) Pineapple with white background, (d) Pineapple with natural background.

The distribution of the dataset based on quality categories and background types can be seen in Table I:

TABLE I. TYPE STYLES

Fruit	Category	Primary		Secondary		Initial Total
		White	Natural	White	Natural	
Strawberry	Good / Ripe	20	15	50	30	70
	Medium / Unripe	30	20	60	30	90
	Damaged / Rotten	15	10	20	15	35
Pineapple	Good / Ripe	15	10	50	20	65
	Medium / Unripe	20	15	60	30	80
	Damaged / Rotten	10	5	25	15	35

Note: "Primary" and "secondary" refer to the sources of the images in the dataset. Primary data is data captured directly using a digital camera, while secondary data is public datasets such as Kaggel data. The background types are either white or natural (real-world scenes). The "initial total" column represents the total number of images per quality category before preprocessing or augmentation.

The dataset is categorized into two subsets: Classification: Images labeled according to quality classes (good, medium, poor) or ripeness level, and Segmentation: Images annotated with fruit masks for each individual fruit object.

B. Preprocessing

To guarantee consistency and improve model resilience, every image was subjected to a uniform preprocessing pipeline before the training phase. Among the steps were:

- Resizing: To comply with the input specifications of the corresponding models, images were shrunk to 224×224 pixels for classification tasks and 256×256 pixels for segmentation tasks.
- Normalization: To enable quicker convergence during training, pixel values were scaled from the initial range of [0,255] to a normalized range of [0,1].
- Data Augmentation: Random augmentations such as ±10% zoom, ±20° rotation, and horizontal flipping were used to enhance the training dataset and enhance generalization.

Examples of the augmented photos for the pineapple and strawberry samples are shown in Fig. 3.

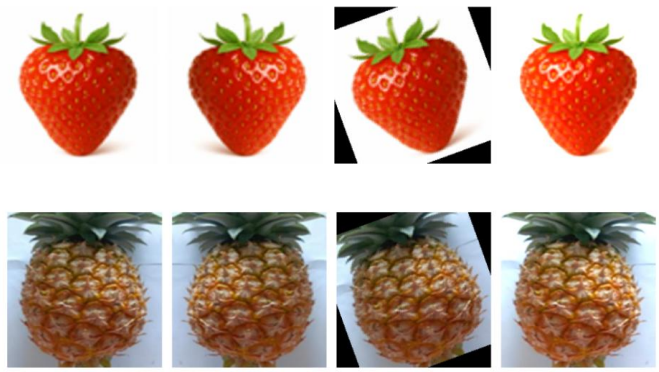


Fig. 3. Data augmentation examples for strawberry (top) and pineapple (bottom) images used to enhance model generalization.

C. Model Architectures Used

Five CNN architectures were selected as baselines for benchmarking classification and segmentation tasks.

1) VGG16: VGG-16 consists of 13 convolutional layers grouped into 5 blocks. Each block contains 2 or 3 convolutional layers. Every convolutional layer uses a 3×3 kernel, 'same'

padding, and the ReLU activation function. After each convolutional block, a 2×2 Max Pooling layer is applied to reduce the spatial dimensions (downsampling), preserve important features, and lower computational complexity. Once all convolutional and pooling blocks are completed, the extracted features are flattened and passed through 3 dense (fully connected) layers. The first two dense layers use ReLU activation, while the final dense layer uses a softmax activation for classification (see Fig. 4).

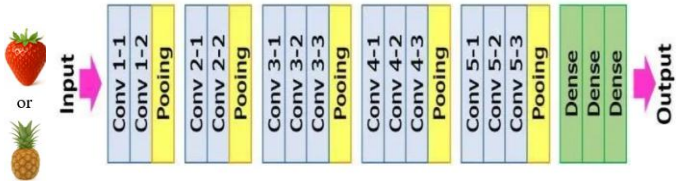


Fig. 4. Architecture of VGG 16.

2) EfficientNet: The architecture of EfficientNet, which consists of seven primary blocks intended for effective feature extraction from input photos, is depicted in Fig. 5 below. After a normal 3x3 convolution layer, a sequence of Mobile Inverted Bottleneck Convolutions (MBConv) is performed. While Blocks 3 to 6 use MBConv6 layers with 5×5 kernels to increase the receptive field without appreciably increasing the number of parameters, Blocks 1 and 2 use MBConv layers with 3×3 kernels. The extracted features are refined in the final Block 7 using an MBConv6 with a 3×3 kernel. High computational efficiency is made possible by the expansion, depthwise convolution, and projection processes included in each MBConv layer. The resulting feature map at the end of the network serves as a rich representation of the input, suitable for downstream tasks such as fruit quality classification. EfficientNet is particularly advantageous for image-based classification of fruits like strawberries and pineapples due to its balance between performance and resource efficiency.

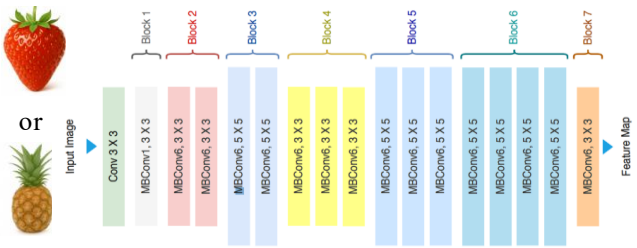


Fig. 5. Architecture of EfficientNet.

3) ResNet50: Fig. 6 given below illustrates the architecture of the ResNet50 model. The architecture begins with a zero padding process, followed by an initial convolutional layer composed of convolution, batch normalization, ReLU activation, and max pooling—collectively referred to as Stage 1. Next, the network comprises four main stages (Stage 2 to Stage 5), each consisting of a Conv Block and several Identity Blocks (ID Blocks). After passing through all the convolutional blocks, the network concludes with an average pooling layer, a flattening process, and a fully connected (FC) layer that produces the final class prediction output.

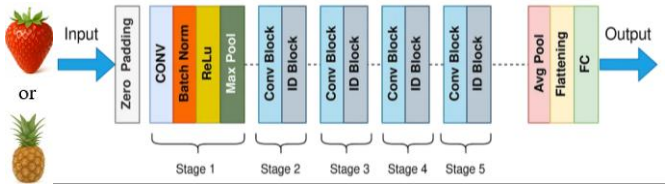


Fig. 6. Architecture ResNet50.

4) U-Net: The U-Net convolutional neural network architecture is specifically designed for image segmentation tasks, such as identifying decayed or damaged areas on pineapple and strawberry fruits. U-Net has a symmetrical structure resembling the letter "U" and consists of two main parts: the contracting path (encoder) on the left side and the expansive path (decoder) on the right. It processes input images of size 256×256×3 (RGB) through the encoder path, which includes 3×3 convolutional blocks with ReLU activation followed by 2×2 max pooling operations. This progressively increases the number of feature channels (from 32 to 512) while reducing the spatial dimensions. At the deepest part of the network (the bottleneck), complex features are represented at the smallest resolution (8×8) with a depth of 512 channels. The decoder then reconstructs the spatial dimensions using 2×2 up-convolution operations, while skip connections from the encoder help retain spatial details. The process concludes with a 1×1 convolution that outputs a segmented image of size 256×256×3, precisely identifying decayed and non-decayed areas on the fruit (see Fig. 7).

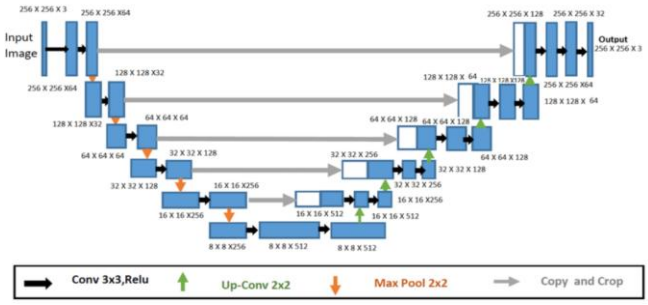


Fig. 7. Architecture of U-Net.

5) DeepLabV3+: The DeepLabV3+ model architecture was created especially for problems involving image segmentation. The encoder, which is represented by the blue line, and the decoder, which is represented by the red path, are its two primary parts. The input image is first processed using atrous (dilated) convolution in the encoder section in order to collect more spatial information without sacrificing resolution. A number of convolution and pooling operations are then performed, along with 1×1 convolutions to lower dimensionality and boost the effectiveness of feature representation. Low-level characteristics from previous network layers are combined with the encoder output in the

decoder stage after it has been upsampled by a factor of four and modified using 1×1 convolutions. The final segmentation map is created by upsampling the combined features by a factor of four after they have been improved by 3×3 convolutions (see Fig. 8).

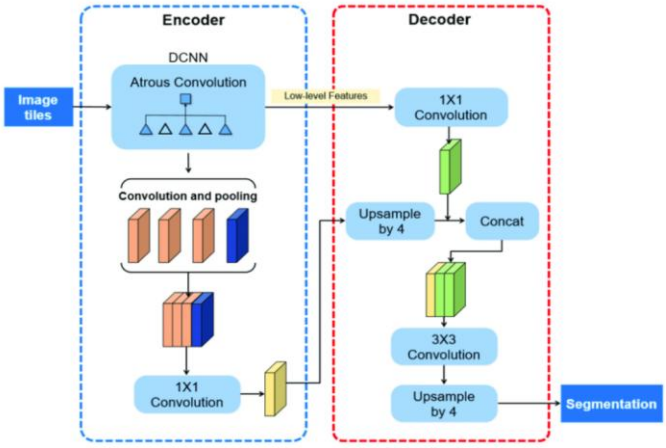


Fig. 8. Architecture of DeepLabV3+.

D. Lightweight Fusion Design for Segmentation

Additionally, a lightweight fusion approach was applied by combining the output probability maps of the two baseline models (U-Net and DeepLabV3+) using a simple adaptive weighting scheme. This approach aims to enhance segmentation consistency without significantly increasing architectural complexity or the number of parameters. The fusion was performed at the output probability level (mask probability map) by combining the probability maps from both models using a simple linear weighting scheme:

$$\text{Pfusion} = \alpha P_{U-net} + (1-\alpha)\text{P}_{\text{DeepLabV3+}}$$

where, α=0.5 serves as a balance weight between the two models. The fused result was then converted into the final binary mask using a threshold value of 0.5.

E. Training and Testing Scheme

The models were trained using 80% of the dataset for training and 20% for testing. The Adam optimizer [1] was employed with a learning rate of 0.001 and a batch size of 32 over 50 epochs. For classification, the categorical cross-entropy loss was utilized due to its effectiveness in multi-class settings [2]. For segmentation tasks, a composite loss combining Dice Loss and Binary Cross-Entropy (BCE) was applied to balance region overlap and pixel-wise prediction accuracy [3-4].

F. Performance Evaluation

The classification and segmentation of images of tropical fruits were the two main tasks for which performance evaluation was carried out. We used the confusion matrix, F1-score, recall, accuracy, and precision as evaluation metrics for the classification job. While precision and recall indicate the model's capacity to accurately identify positive occurrences and discover all pertinent instances, respectively, accuracy gauges the overall correctness of predictions. Particularly helpful when there is a class imbalance, the F1-score offers a harmonious compromise between recall and precision [33-34]. The performance of VGG16, ResNet50, and EfficientNet-B0 in classifying images of strawberries and pineapples was evaluated and compared using these criteria.

The evaluation used pixel accuracy, Dice Similarity Coefficient (DSC), and Intersection over Union (IoU) for the segmentation job. Because IoU and DSC can measure the spatial overlap between predicted regions and ground truth, they are frequently utilized in semantic segmentation tasks [35]. The percentage of correctly categorized pixels throughout the entire image is known as pixel accuracy. The segmentation performance of U-Net and DeepLabV3+ in recognizing fruit regions and detecting areas of visual deterioration or damage was assessed using these criteria [36].

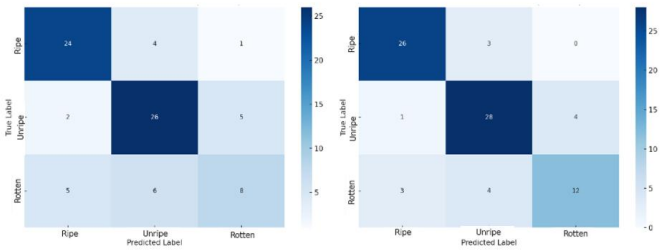
Besides classification accuracy, computational efficiency and statistical reliability are also crucial factors to ensure the feasibility of deploying the model in real-time or edge-based applications. The evaluation considers several complementary approaches, including inference time (runtime), number of parameters, and model size, to assess both predictive performance and computational efficiency of each architecture. Furthermore, a statistical significance test was conducted to verify that the performance differences among models are statistically meaningful.

To ensure model stability and reliability, a 5-fold cross-validation strategy was applied to both the pineapple and strawberry datasets. This approach allows for evaluating the model's generalization capability across different data variations while minimizing bias caused by uneven data partitioning.

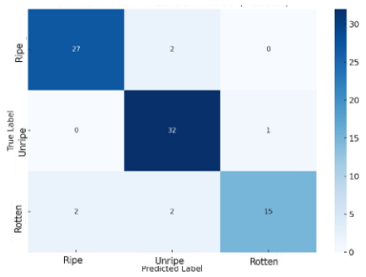
IV. Results and Discussion

A. Fruit Image Classification Results

In this horticultural fruit classification experiment (pineapples and strawberries), three CNN architectures were used: VGG16, ResNet50, and EfficientNet-B0. The dataset was divided into training, validation, and test data with a 70:15:15 ratio. To ensure model stability, 5-fold cross-validation was performed on both datasets. The experimental results, which illustrate the classification distribution and prediction error patterns, are shown in the Fig. 9 and Fig. 10 below. The confusion matrix was generated from the fold with the highest accuracy during the 5-fold cross-validation process.

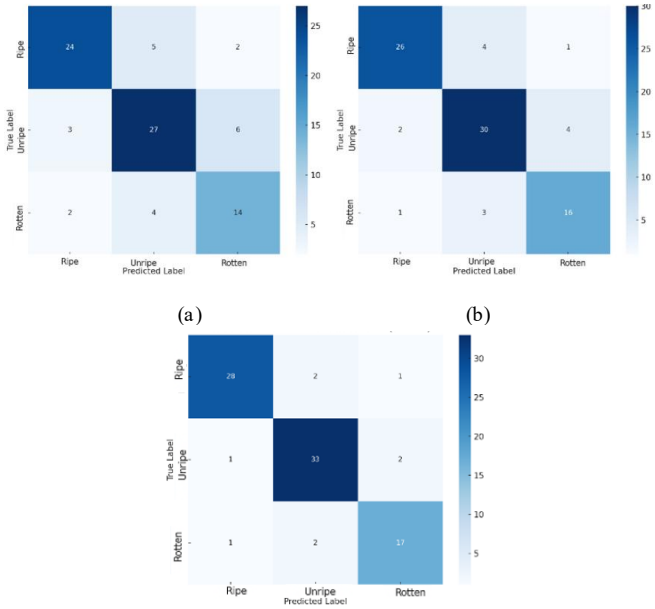


(a) (b)



(c)

Fig. 9. Confusion matrix: (a) VGG 16, (b) Resnet 50, (c) EfficientNet on pineapple dataset.



(a) (b)

(c)

Fig. 10. Confusion matrix: (a) VGG 16, (b) Resnet 50, (c) EfficientNet on strawberry dataset.

Each model was thoroughly evaluated on five different combinations of training and testing data, so that the results obtained were independent of a single data split. This evaluation included four main metrics: accuracy, precision, recall, and F1-score. The results of the average classification performance comparison of each architecture on the pineapple and strawberry datasets are shown in Table II and Table III. Furthermore, to illustrate the stability of inter-fold performance, the average value and standard deviation (mean ± SD) of each metric are also shown.

TABLE II. AVERAGE 5-FOLD CROSS-VALIDATION RESULTS ON PINEAPPLE DATASET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG16	82.30 ± 1.1	80.40 ± 1.3	89.10 ± 1.5	80.50 ± 1.4
ResNet50	84.20 ± 0.9	85.00 ± 1.1	83.20 ± 1.0	84.10 ± 1.1
EfficientNet-B0	90.70 ± 0.7	91.10 ± 0.8	90.60 ± 0.6	91.20 ± 0.7

TABLE III. AVERAGE 5-FOLD CROSS-VALIDATION RESULTS ON STRAWBERRY DATASET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG16	83.10 ± 1.0	82.50 ± 1.2	80.90± 1.1	82.00 ± 1.2
ResNet50	85.80 ± 0.8	85.90 ± 1.0	85.40 ± 0.9	85.70 ± 1.0
EfficientNet-B0	91.40 ± 0.6	91.80 ± 0.7	90.20 ± 0.6	90.60 ± 1.6

Table II and Table III summarize the model performance results by averaging across 5k folds. Based on these tables, it can be observed that EfficientNet-B0 consistently achieves the best performance in terms of accuracy, precision, recall, and F1 score on both datasets, followed by ResNet50 and VGG16. VGG16 has the lowest accuracy with a slightly higher standard deviation, indicating the model's sensitivity to variations in lighting and color on the fruit surface. Overall, the results on both datasets confirm that a more modern and lightweight architecture, such as EfficientNet-B0, is able to provide the best combination of high accuracy and inter-fold performance stability. The low standard deviation value of this model also indicates high reliability and good potential for application in fruit image classification systems in real environments.

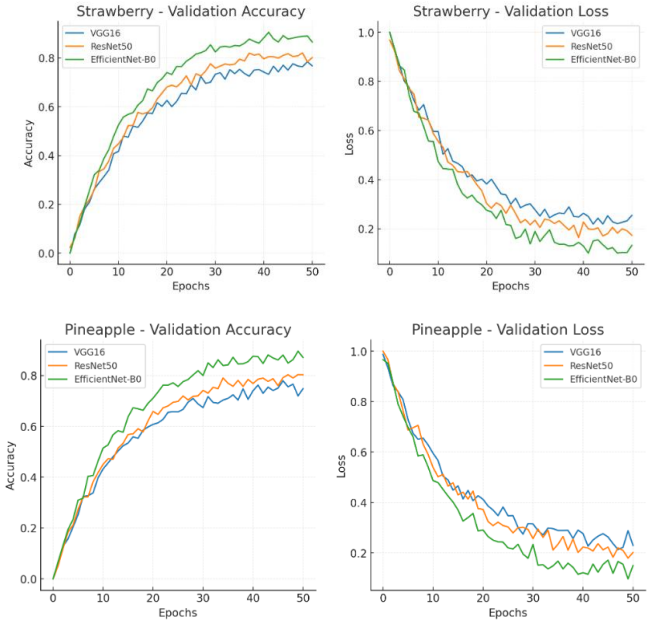


Fig. 11. Validation accuracy and validation loss: strawberry and pineapple.

To illustrate the performance of the models during the training process, the validation accuracy and loss curves on the strawberry and pineapple datasets are visualized. Fig. 11 presents the learning trends of each model (VGG16, ResNet50, and EfficientNet-B0) across 50 epochs.

The experimental results on the strawberry dataset show a stable trend of increasing validation accuracy as the number of epochs increases, reaching convergence around epoch 40. The EfficientNet-B0 model consistently achieved the highest accuracy, followed by ResNet50 and VGG16. This performance demonstrates that modern architectures such as EfficientNet are able to extract strawberry visual features more

effectively, particularly in distinguishing variations in color and texture that are crucial indicators for determining the condition of fresh and spoiled fruit. Meanwhile, on the pineapple dataset, all three models also exhibited consistent patterns of increasing validation accuracy and decreasing loss throughout the training process. EfficientNet-B0 once again achieved the best performance. These findings indicate that the more complex visual characteristics of pineapples, such as scaly skin patterns and lighting variations, are more effectively handled by the EfficientNet architecture. This is attributed to the optimization strategy employed by EfficientNet, which balances network depth, width, and resolution, thereby enabling richer and more accurate feature representations compared to ResNet50 and VGG16.

B. Segmentation Results

Segmentation experiments were conducted using U-Net and DeepLabV3+ on strawberry and pineapple images. The evaluation was carried out using three main metrics: Intersection over Union (IoU), Dice Coefficient, and mean Intersection over Union (mIoU). Table IV summarizes the performance of both models:

TABLE IV. EVALUATION METRICS AND AVERAGE SEGMENTATION PERFORMANCE (MEAN ± SD) ON STRAWBERRY AND PINEAPPLE DATASETS

Fruit	Model	IoU (%)	Dice (%)	mIoU (%)
Strawberry	U-Net	88.5 ± 0.9	89.5 ± 0.8	87.6 ± 1.0
Strawberry	DeepLabV3+	91.8 ± 0.7	91.5 ± 0.6	90.7 ± 0.8
Pineapple	U-Net	87.8 ± 1.0	88.7 ± 0.9	86.9 ± 1.1
Pineapple	DeepLabV3+	90.6 ± 0.8	91.7 ± 0.7	89.5 ± 0.9

The average segmentation performance results in Table IV show that the DeepLabV3+ model consistently produces higher IoU, Dice, and mIoU values than the U-Net for both datasets: strawberry and pineapple. The performance difference between the two models appears stable, as evidenced by the relatively small standard deviation values (ranging from ± 0.6 to ± 1.1), indicating good interfold consistency of the segmentation results in the 5-fold cross-validation scheme. DeepLabV3+'s superiority lies primarily in its use of Atrous Convolution and Atrous Spatial Pyramid Pooling (ASPP), which are highly effective in capturing variations in texture and object size on the fruit surface.

C. Computational Efficiency Aspect

In addition to accuracy and segmentation performance, computational efficiency is also an important consideration to ensure the feasibility of model deployment on systems with limited hardware resources. The efficiency evaluation is conducted by examining several key metrics, namely inference time (runtime), number of parameters, and model size. The results of the computational efficiency evaluation for all classification and segmentation models are presented in Table V.

TABLE V. COMPUTATIONAL EFFICIENCY EVALUATION OF CLASSIFICATION AND SEGMENTATION MODELS

Model	Task	Inference Time (s)	Number of Parameters (M)	Model Size (MB)
VGG16	Classification	0.042	138.3	528.0
ResNet50	Classification	0.038	25.6	98.0
EfficientNet-B0	Classification	0.026	5.3	20.4
U-Net	Segmentation	0.087	31.0	122.0
DeepLabV3+	Segmentation	0.093	43.5	175.0

D. Segmentation Results with Lightweight Fusion

This study applied a lightweight fusion approach to enhance segmentation consistency without increasing architectural complexity. The approach combines the output probability maps (mask probability maps) of two baseline models—U-Net and DeepLabV3+—through a simple linear weighting scheme. Table VI presents the segmentation performance after applying lightweight fusion between U-Net and DeepLabV3+:

TABLE VI. COMPARATIVE SEGMENTATION PERFORMANCE OF U-NET, DEEPLABV3+, AND FUSION MODEL ON STRAWBERRY AND PINEAPPLE IMAGES

Fruit	Model	IoU (%)	Dice (%)	mIoU (%)
Strawberry	U-Net	89.3	90.8	88.6
Strawberry	DeepLabV3+	92.1	93.5	91.7
Strawberry	Fusion (U+D)	**92.6**	**93.9**	**92.2**
Pineapple	U-Net	88.7	90.1	87.9
Pineapple	DeepLabV3+	91.4	92.7	90.8
Pineapple	Fusion (U+D)	**91.9**	**93.1**	**91.3**

The results demonstrate a consistent improvement across all key metrics after fusion. The average IoU increased by approximately 0.5–0.7%, while the Dice coefficient improved by about 0.4%. This enhancement is attributed to the complementary strengths of the two models: DeepLabV3+ excels in handling complex regions, whereas U-Net provides smoother boundary delineation. The combination produces more stable and accurate segmentation outcomes. This approach aligns with recent research trends emphasizing the effectiveness of lightweight hybrid models in improving segmentation robustness for agricultural imagery under natural illumination conditions [37].

E. Comparative Analysis

To ensure that the performance differences among models were not merely due to random variations in the data, a statistical significance analysis was conducted based on the 5-fold cross-validation results for both classification and segmentation tasks. This test aimed to verify the statistical validity of each model's performance, ensuring that higher-performing architectures exhibit improvements that are mathematically significant rather than coincidental.

The analysis was applied to both classification models (VGG16, ResNet50, and EfficientNet-B0) and segmentation

models (U-Net and DeepLabV3+) using significance levels of $\alpha = 0.05$ and 0.01. The complete results of the significance tests for both datasets are presented in Table VII and Table VIII below.

TABLE VII. STATISTICAL SIGNIFICANCE TEST RESULTS (5-FOLD CROSS-VALIDATION): CLASSIFICATION ON PINEAPPLE AND STRAWBERRY DATASETS

Model Comparison	Pineapple Dataset (p-value)	Strawberry Dataset (p-value)	Remark
VGG16 vs ResNet50	0.041 *	0.038 *	Significant
VGG16 vs EfficientNet-B0	0.007 **	0.005 **	Highly significant
ResNet50 vs EfficientNet-B0	0.018 *	0.015 *	Significant

TABLE VIII. STATISTICAL SIGNIFICANCE TEST RESULTS (5-FOLD CROSS-VALIDATION): SEGMENTATION ON PINEAPPLE AND STRAWBERRY DATASETS

Model Comparison	Pineapple Dataset (p-value)	Strawberry Dataset (p-value)	Remark
U-Net vs DeepLabV3+	0.009 **	0.006 **	Highly significant

The results of the statistical significance test confirm that the observed differences in performance between models are not due to random variations in the data, but rather have real statistical significance. The overall experimental results confirm that EfficientNet-B0 consistently outperformed ResNet50 and VGG16 in classification tasks, demonstrating an excellent balance between parameter efficiency and generalization capability. This finding aligns with the report, where EfficientNet achieved up to 97% accuracy in grape and potato leaf disease classification [38]. A follow-up study also emphasized the superior efficiency of EfficientNet compared to other architectures [39].

For segmentation tasks, DeepLabV3+ proved more precise than U-Net, particularly in handling the complex textures of strawberries and pineapples—the multiscale effect of atrous convolution and the ASPP module played a pivotal role. Evidence from the SugarBeets study also showed that DeepLabV3+ achieved a higher mIoU compared to U-Net [40].

Nevertheless, some error patterns persisted. False positives frequently occurred in decayed regions obscured by shadows, while false negatives often appeared when small decayed spots were hidden, similar to the findings on apples [41]. Segmentation also exhibited mask leakage along object boundaries, as reported in tomato images [42].

As a direction for further development, integrating adaptive multitask architectures such as AISAM-CSNet (Artificial Immune System-controlled Adaptive Multi-agent Classification and Segmentation Network) appears promising for improving robustness. AISAM-CSNet combines multi-agent CNNs, reinforcement learning, and optimization inspired by the Artificial Immune System, consistent with current trends in multitask learning and biologically inspired optimization within agricultural computer vision [43].

V. CONCLUSION

The experimental results demonstrated that deep learning methods are highly effective for assessing the visual quality of tropical fruits. Among the evaluated models, EfficientNet-B0 consistently achieved superior performance in the classification tasks, with an average accuracy of 90.7% for pineapples and 91.4% for strawberries, outperforming both ResNet50 and VGG16. This result confirms that lightweight and modern architectures are more capable of capturing discriminative visual patterns related to fruit ripeness and surface damage compared to conventional CNNs. Meanwhile, for the segmentation task, DeepLabV3+ outperformed U-Net, with an IoU of 90.6% for pineapple and 91.8% for strawberry, while the Dice coefficient was 91.5% for strawberry and 91.7% for pineapple. These results confirm that both architectures can serve as a reliable basis for automatically detecting ripeness and damaged areas in fruit.

Beyond numerical achievements, the findings highlight the practical contribution of deep learning in accelerating postharvest quality inspection while providing an empirical foundation for developing more adaptive multitask architectures. This has significant implications for the horticulture industry by reducing reliance on labor-intensive manual inspection.

For future research, the integration of classification and segmentation into a single multitask framework, such as AISAM-CSNet (Artificial Immune System-Controlled Adaptive Multi-agent Classification and Segmentation Network), holds strong potential to further enhance accuracy and adaptability under diverse real-world conditions. AISAM-CSNet combines classification and segmentation in a multi-agent reinforcement learning framework, with hyperparameter optimization inspired by the Artificial Immune System. Such an approach is expected to improve adaptability and accuracy in tropical fruit quality assessment systems, thereby supporting supply chain efficiency and ensuring consistent fruit quality.

REFERENCES

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

[2] Sudars, K., Pathak, K., Ferentinos, K. P., Tzounis, M., & Bochtis, A. (2020). Deep learning for image-based fruit classification and quality evaluation: A review. Agronomy, 10(6), 886. https://doi.org/10.3390/agronomy10060886

[3] Islam, M. A., Ali, M. S., & Hasan, M. M. (2023). Real-time fruit detection and classification using improved YOLOv5 and ResNet50. Computers and Electronics in Agriculture, 205, 107609. https://doi.org/10.1016/j.compag.2023.107609

[4] Campos, P., Marañón-Blanco, J. U., et al. (2023). Fruit quality classification using VGG16 and ResNet50 on FruitNet dataset. Computers and Electronics in Agriculture, 203, 107472. https://doi.org/10.1016/j.compag.2023.107472

[5] Li, X., Liu, Y., Zhang, H., & Wu, L. (2024). EfficientNet for plant leaf classification. Computers and Electronics in Agriculture, 209, 107926. https://doi.org/10.1016/j.compag.2024.107926

[6] Hasan, M. M., Tumpa, A., Mamun, A. M. A., & Shamsuddin, M. A. (2021). Fruit image classification using deep learning and Gaussian process regression. Information Processing in Agriculture, 8(3), 414–424. https://doi.org/10.1016/j.inpa.2020.08.004

[7] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1409.1556

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

[9] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In MICCAI (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[10] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV (pp. 833–851). Springer. https://doi.org/10.1007/978-3-030-01234-2_49

[11] Wang, Z., Hu, M., & Zhai, G. (2020). Apple segmentation based on DeepLabV3+ with dense conditional random field. Computers and Electronics in Agriculture, 179, 105842. https://doi.org/10.1016/j.compag.2020.105842

[12] Mo, C., Ma, C., & Wang, Y. (2024). Ripeness detection of sugar apple using MobileNetV2-DeepLabV3+. Agronomy, 14(4), 591. https://doi.org/10.3390/agronomy14040591

[13] Chiu, M. T., Tao, A., et al. (2020). Agriculture-Vision: A large aerial image database for agricultural pattern analysis. In CVPRW (pp. 2870–2879).https://openaccess.thecvf.com/content_CVPRW_2020/html/w41/Chiu_Agriculture-Vision_CVPRW_2020_paper.html

[14] Shorten, A., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6, 60. https://doi.org/10.1186/s40537-019-0197-0

[15] Chen, Y., Wang, S., Xu, X., & Xie, Y. (2021). Domain adaptation for plant disease recognition with adversarial training. Computers and Electronics in Agriculture, 187, 106303. https://doi.org/10.1016/j.compag.2021.106303

[16] Rahnemoonfar, M., & Sheppard, C. (2017). Deep count: Fruit counting based on deep simulated learning. Sensors, 17(4), 905. https://doi.org/10.3390/s17040905

[17] Arif, M., Fatima, S., & Khan, M. A. (2022). Real-time citrus fruit classification using deep convolutional neural networks. Sustainable Computing: Informatics and Systems, 35, 100753. https://doi.org/10.1016/j.suscom.2022.100753

[18] Wagle, K., Dinh, A., & Wang, Z. (2023). Edge deployment of EfficientNet for fruit classification. Computers and Electronics in Agriculture, 205, 107638. https://doi.org/10.1016/j.compag.2023.107638

[19] Reyes, R., Pérez, J., & Soto, A. (2022). Fine-grained fruit classification using EfficientNet and transfer learning. Journal of Imaging, 8(2), 50. https://doi.org/10.3390/jimaging8020050

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

[21] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In CVPR (pp. 4510–4520). https://doi.org/10.1109/CVPR.2018.00474

[22] Rauf, H. T., Lali, M. I. U., & Khan, M. A. (2022). Performance comparison of CNN architectures for fruit classification. Artificial Intelligence in Agriculture, 6, 24–32. https://doi.org/10.1016/j.aiia.2022.05.001

[23] Fang, J., Wang, C., & Xu, T. (2022). Strawberry segmentation using attention U-Net in field conditions. Computers and Electronics in Agriculture, 198, 107061. https://doi.org/10.1016/j.compag.2022.107061

[24] Jamil, N., Rauf, H. T., & Mahmood, T. (2023). Mango fruit segmentation using squeeze-excitation U-Net. Journal of Food Measurement and Characterization, 17, 2023–2031. https://doi.org/10.1007/s11694-023-01474-6

[25] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2961–2969). https://doi.org/10.1109/ICCV.2017.322

[26] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep High-Resolution Representation Learning for Visual Recognition. In CVPR (pp. 5390–5399). https://doi.org/10.1109/CVPR.2019.00553

[27] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

[28] Milioto, A., Lottes, P., & Stachniss, C. (2018). Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. In IROS (pp. 2229–2235). https://doi.org/10.1109/IROS.2018.8593940

[29] Singh, D., Jain, N., & Nandi, G. C. (2020). Fruit recognition in natural environment using deep learning. Procedia Computer Science, 167, 2201–2210. https://doi.org/10.1016/j.procs.2020.03.270

[30] Xu, L., Ma, R., Zhang, C., & Hu, W. (2021). Challenges in cross-domain apple classification. Sensors, 21(2), 509. https://doi.org/10.3390/s21020509Rahnemoonfar, M., & Sheppard, C. (2017). Deep count: Fruit counting based on deep simulated learning. Sensors, 17(4), 905. https://doi.org/10.3390/s17040905

[31] Adhil, P. K. (2022). *Pineapple* [Data set]. Kaggle. https://www.kaggle.com/datasets/adhilpk/pineapple

[32] Basit, A. (2023). *Strawberry dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/abdulbasit31/strawberry-dataset

[33] Zhang, Y., et al. (2019). Improved cross-entropy loss function for deep neural networks classification. IEEE Access, 7, 64366–64374. https://doi.org/10.1109/ACCESS.2019.2917624

[34] Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proceedings of the 4th International Conference on 3D Vision (3DV), 565–571. https://doi.org/10.1109/3DV.2016.79

[35] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In Deep learning in medical image analysis and multimodal learning for clinical decision support (pp. 240–248). Springer. https://doi.org/10.1007/978-3-319-67558-9_28

[36] Akther, J., Harun-Or-Roshid, M., & Islam, A. (2024). Comparative analysis of CNN, EfficientNet and ResNet for grape and potato leaves disease prediction: A deep learning approach. ResearchGate. https://10.21203/rs.3.rs-5037532/v1

[37] Zhang, L., Li, M., Zhang, P., & Liu, P. (2025). EfficientSegNet: Lightweight Semantic Segmentation with Multi-Scale Feature Fusion and Boundary Enhancement. Sensors, 25(19), 5934.

[38] Kansal, K. (2024). ResNet-50 vs. EfficientNet-B0: Multi-Centric Classification Performance Comparison. Procedia Computer Science. https://doi.org/10.1016/j.procs.2024.04.007

[39] Liu, C. (2025). DeepLabV3+ performance on SugarBeets dataset. Technical Report, DIVA Portal. https://liu.diva-portal.org/smash/get/diva2%3A1973627/FULLTEXT01.pdf

[40] Rathore, D., Divyanth, L. G., Reddy, K. L. S., Chawla, Y., Buragohain, M., Soni, P., ... & Ghosh, A. (2023). A two-stage deep-learning model for detection and occlusion-based classification of Kashmiri orchard apples for robotic harvesting. Journal of Biosystems Engineering, 48(2), 242-256. . https://link.springer.com/article/10.1007/s42853-023-00190-0

[41] Shoaib, M., Hussain, T., Shah, B., Ullah, I., Shah, S. M., Ali, F., & Park, S. H. (2022). Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease. Frontiers in plant science, 13, 1031748. https://doi.org/10.3389/fpls.2022.1031748

[42] Wang, Y., Yang, L., Liu, X., & Yan, P. (2024). An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+https://www.nature.com/articles/s41598-024-60375-1

[43] Moisés, A. G., et al. (2023). Augmentation techniques for deep learning in agrifood applications. Sensors, 23(20), 8562. https://doi.org/10.3390/s23208562