



Attention-Based Multi-View Fusion YOLO for Non-Destructive Pineapple Sweetness Assessment

Dwi Vernanda^{1*}, Tri Herdiawan Apandi², Aisyah Suhailla Binti Jili³, Desy Triastuti⁴, Willy Muhammad Fauzi⁵

¹Information Systems, Department of Information Technology and Computer, Politeknik Negeri Subang, Subang, Indonesia

²Software Engineering Technology, Department of Information Technology and Computer, Politeknik Negeri Subang, Subang, Indonesia

³Department of Information and Communication, Politeknik Mukah, Sarawak, Malaysia

⁴Edible Plants Production Technology, Department of Agriculture, Politeknik Negeri Subang, Subang, Indonesia

⁵Information Systems, Department of Information Technology and Computer, Politeknik Negeri Subang, Subang, Indonesia

*nanda@polsub.ac.id

Abstract

Classifying the sweetness level of pineapples is an important part of quality control, but existing methods still face issues of subjectivity and require destructive testing. Manual assessment is often inconsistent, while refractometer measurements require cutting the fruit open. Single-view computer vision offers a non-destructive alternative, yet its performance remains limited because visual cues related to sweetness appear on different sides of the fruit. This study introduces Multi-View Fusion YOLO (MVF-YOLO), a model that combines five viewpoints (full, front, left, right, and back) through an attention mechanism to perform adaptive sweetness classification. The dataset consists of 570 pineapples with TSS/TA ratios as the ground truth, producing 2,850 images grouped into three categories: sour (TSS/TA 10–20), ideal (TSS/TA 20–30), and very sweet (TSS/TA >30). MVF-YOLO achieved an mAP@0.5 of 82.1% and an overall accuracy of 84.2%, outperforming the single-view baseline by 14.3%. Attention weight analysis indicates that the full view contributes the most (0.267). With an inference time of 45.8 ms per fruit, the model is sufficiently efficient for use by farmers, distributors, and consumers. The results demonstrate that a multiview approach enhanced with learned attention can significantly improve sweetness classification accuracy without compromising computational efficiency.

Keywords: attention mechanism; multi-view fusion; non-destructive testing; pineapple; YOLO

How to Cite: D. Vernanda, T. H. Apandi, A. Suhailla Binti Jili, D. Triastuti, and W. Muhammad Fauzi, "Attention-Based Multi-View Fusion YOLO for Non-Destructive Pineapple Sweetness Assessment", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 2, pp. 267 - 279, Apr. 2026. [Permalink/DOI: https://doi.org/10.29207/resti.v10i2.7382](https://doi.org/10.29207/resti.v10i2.7382)

Received: December 15, 2025

Accepted: March 8, 2026

Available Online: April 6, 2026

This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Published by [Ikatan Ahli Informatika Indonesia](https://www.ikatanahliinformatika.org/)

1. Introduction

Subang Regency is known as one of the largest pineapple production centers in Indonesia [1]. Its produce is distributed to traditional markets, supermarkets, and processing industries [2]. Currently, the sorting process is still performed manually by observing the skin color and aroma of the fruit. This method relies heavily on worker experience, which often leads to inconsistent results. Sweetness level, measured using TSS/TA ratio, is the primary indicator of pineapple quality and can be assessed more objectively using a refractometer [3]. However, this method is destructive because the fruit must be cut

open. For large-scale sorting operations, it is inefficient and increases operational costs.

Computer vision offers a non-destructive alternative [4]. Various studies have shown that automatic classification can assist in evaluating fruit quality. However, most existing systems only use a single viewpoint. In reality, visual characteristics related to sweetness – crown color, eye patterns, skin texture are distributed across different sides of the fruit [5]. A single viewpoint is often insufficient, especially for distinguishing boundary cases between adjacent categories. For example, pineapples with a TSS/TA ratio of 19.5 (sour) and 20.5 (ideal), can appear

nearly identical when viewed from one angle [6]. Additionally, single-view approaches are sensitive to fruit position and orientation during image capture, which is difficult to control in practical deployment scenarios.

Deep learning has been extensively applied to fruit quality detection tasks. YOLO in particular has become popular due to its balance between speed and accuracy [7], [8]. Previous research in classification has achieved mAP50 up to 0.965, particularly in the classification of corn leaf diseases [9]. In other domains, studies have shown that YOLOv8 can achieve a precision of 0.906 and a recall of 0.943 for disability detection [10], [11]. For pineapples specifically, some research has used classical machine learning methods for quality assessment [12], [13], but the dataset size is still limited and the accuracy is not yet optimal [14].

Multi-view approaches have started gaining traction in agricultural quality assessment [15]. Multi-view CNN has been well-known since the seminal paper on 3D shape recognition [16]. In the agricultural domain, research found that side views were more informative for apple defect detection [17]. This finding is notable because it demonstrates that view importance is task-dependent-what works for defect detection might differ for sweetness assessment. Despite this potential, pineapple sweetness research has not yet explored multi-view fusion. Existing studies either use single viewpoints [13], [18] or apply basic fusion techniques, which risks missing characteristics that may be visible from other sides of the fruit [19].

The attention mechanism in deep learning essentially teaches models to “focus” on the most relevant information [20]. This concept has been extensively applied in medical imaging and autonomous driving with impressive results [21]. Research has shown that attention mechanism can improve fruit classification performance to reach a macro F1-score of 96.74%, a significant improvement over models without attention [22]. The key advantage, unlike simple averaging that treats all inputs equally, attention learns to assign different weights based on relevance [23]. This is particularly relevant for multi-view scenarios because not all viewpoints carry equal information - some angles may reveal more distinctive features than others for specific samples.

Despite these advances, existing pineapple sweetness research faces three critical limitations. First, all current studies rely exclusively on single-view approaches [13], [14], which cannot capture the spatial distribution of sweetness related visual cues across different fruit surfaces crown appearance, skin color uniformity, and eye pattern characteristics that vary depending on viewing angle [19]. Second, while multi-view methods exist for general fruit quality assessment, they employ basic fusion techniques such as averaging or concatenation that treat all viewpoints equally, ignoring

the fact that different angles carry different levels of relevant information for specific samples [24] - [26]. Third, although attention mechanisms have proven highly effective in medical imaging and autonomous driving, their application to agricultural quality assessment, specifically for multi-view sweetness classification, remains completely unexplored[27].

These limitations are significant because sweetness classification is inherently more challenging than ripeness detection-visual indicators are subtle and distributed unevenly across the fruit surface, and boundary cases between categories (e.g., TSS/TA ratio 19.8 vs 20.2) appear nearly identical from any single viewpoint. Simple averaging cannot adaptively identify which viewpoint is most informative for each individual sample's unique characteristics. Our research addresses these gaps by introducing the first attention-based multi-view fusion architecture specifically designed for pineapple sweetness assessment. Unlike prior work that treats all views equally, our approach learns adaptive attention weights that select the most relevant viewpoints based on each fruit's visual characteristics, providing both improved accuracy and interpretability through analysis of learned view importance patterns.

This research develops MVF-YOLO for pineapple sweetness classification using attention based multi-view fusion of five viewpoints (full, front, left, right, and back). Using a dataset consists of 570 pineapples (2,850 images) across three TSS/TA ratio categories: sour (10–20), ideal (20–30), and very sweet (>30). This study addresses several key questions: (1) the extent to which the multi-view approach improves performance compared to single-view baselines, (2) which of the five views provides the most informative features, (3) how robust the model is when some views are unavailable, and (4) whether inference speed is practical for industrial deployment. The research benefits farmers and distributors through non-destructive quality assessment, enables real-time industrial sorting (45.8ms per fruit), and provides consumers with more reliable sweetness categorization. Novel contributions include: (1) the first attention-based multi-view fusion for fruit sweetness assessment, (2) demonstration that learned attention weights (full view = 0.267) align with farmer expertise, and (3) evidence that 3-view configurations achieve 96% of full performance under resource constraints.

2. Methods

The MVF-YOLO research methodology consists of six main stages as shown in Figure 1. We start with dataset collection from pineapple farms in Subang, capturing multi-view images and measuring TSS/TA ratios for ground truth labels. Next, preprocessing handles image normalization and augmentation to improve model robustness. The third stage involves designing the MVF-YOLO architecture with its attention-based fusion mechanism. Training comes fourth, where we optimize the model using carefully selected

hyperparameters and monitor convergence. The fifth stage focuses on evaluation using various classification metrics. Finally, we conduct additional analyses including ablation studies and cross-dataset validation to assess model robustness and generalizability. Details of each stage are presented in the following subsections.

2.1 Dataset and Data Acquisition

This study uses real pineapple samples with actual multi-view photographs and physical TSS/TA measurements, not simulated data. The dataset used in this study was Smooth Cayenne pineapples collected from two main pineapple-producing areas in Subang Regency: Jalancagak and Cijambe Districts. Data collection took place at two locations: the homes of farmer partners in Cijambe District and the Agro-Industry Quality Testing Laboratory (PMA) of the Subang State Polytechnic. Initially, data collection was conducted at partner locations, with 54 pineapples collected. This was followed by additional data collection on campus, resulting in a total of 570 pineapples. These two sub-districts were selected based on differences in environmental conditions and fruit ripeness, resulting in a more representative data set of pineapple characteristics in Subang Regency.

Each fruit was photographed from five different angles (Full, Front, Left, Right, Back) using a Xiaomi Redmi Note 10 Pro smartphone camera with a 108 MP resolution, under natural lighting. In total, 2,850 images (570×5 views) were collected. The MVF-YOLO research methodology flow is presented in Figure 1, which consists of six main stages starting from dataset collection to evaluation and analysis.

The shooting process follows standard protocols, namely the camera distance is maintained at 30–40 cm from the object, natural lighting is carried out between 09.00–15.00 to maintain color consistency and the background uses a mini studio as a neutral background. After the shooting session, 3 samples of pineapple juice were taken from each pineapple and then the sweetness level was measured using an Atago PAL-1 digital refractometer to obtain the Total Soluble Solids (TSS) or Brix value [28] and the acidity level or Total Titratable Acidity (TA) of each pineapple was measured using a titration tool.

Based on the TSS and TA calculations, a TSS/TA ratio was obtained, indicating the sweetness level of the pineapple. The sweetness levels of the pineapples were then grouped into three categories based on the TSS/TA ratio: sour (10-20), ideal (20-30), and very sweet (>30). After calculating the TSS/TA ratio for each pineapple, the sour category consisted of 180 pineapples (31.6%), the ideal category consisted of 245 pineapples (43.0%), and the very sweet category consisted of 145 pineapples (25.4%). This proportion reflects field conditions, where pineapples with an ideal sweetness level are more commonly found. The dataset was then divided using stratified sampling: 70% for training (399 pineapples), 15% for validation (86 pineapples), and 15% for testing

(85 pineapples). Stratified sampling was chosen to maintain class proportions in each data division.

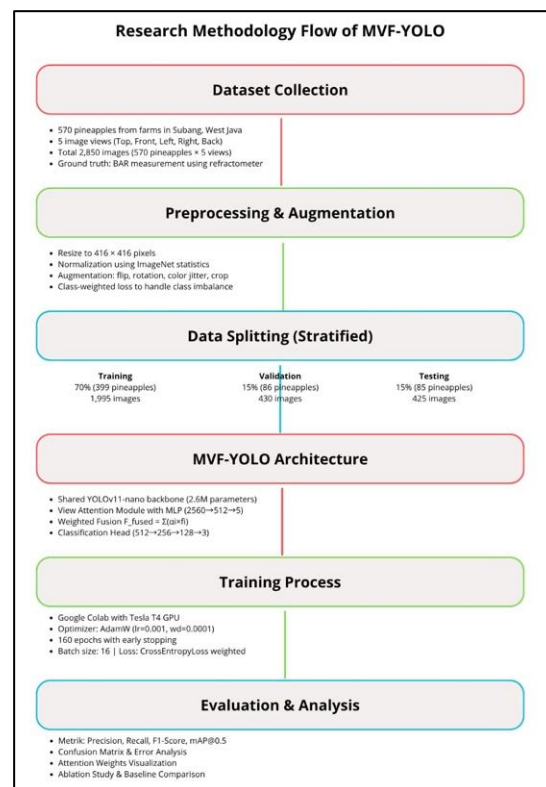


Figure 1. MVF-YOLO research methodology flow

2.2 Data Preparation and Data Augmentation

Preprocessing was performed through a Roboflow pipeline. First, auto-orient was applied to ensure the EXIF orientation of the photos was correct. Second, all images were resized to 416x416 pixels to maintain computational efficiency while retaining necessary visual details.

This resolution was chosen based on preliminary experiments, where 224x224 resulted in significant loss of detail, while 640x640 significantly slowed the training process without providing significant performance improvements. Third, pixel values were normalized using ImageNet statistics, including means of 0.485, 0.456, and 0.406, and standard deviations of 0.229, 0.224, and 0.225. This adjustment was made because the YOLOv11 backbone used had been pre-trained on the ImageNet dataset [29].

The augmentation process is applied solely to the training data through Roboflow's built-in augmentation pipeline. The techniques used include horizontal flip (50% probability), rotation ($\pm 15^\circ$), brightness adjustment ($\pm 20\%$), contrast adjustment ($\pm 20\%$), and saturation adjustment ($\pm 20\%$), as well as random cropping with resizing. The goal of this augmentation is to improve the model's robustness to lighting variations and orientation differences commonly encountered in real-world conditions.

2.3. YOLO Fusion Multi-View Architecture

The MVF-YOLO architecture consists of three main components that work end to end, as shown in Figure 2.

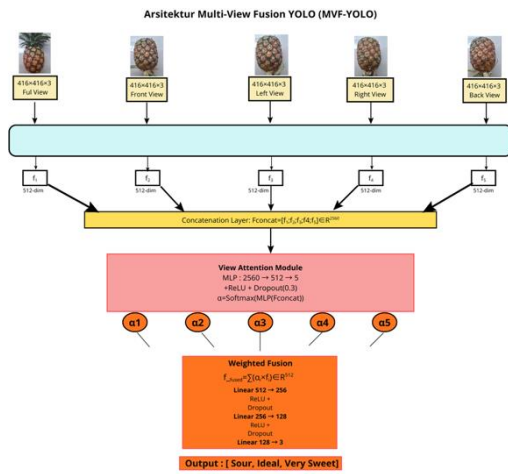


Figure 2. MVF-YOLO architecture

The model uses YOLOv11 as a backbone to extract features from each viewpoint. This backbone was chosen because it is lightweight (only 2.6M parameters) while still capturing important visual patterns. All viewpoints are processed using the same weights, making the backbone shared [30]. This approach not only reduces the number of parameters significantly (5 times more efficient), but also maintains the consistency of feature representation across angles.

The backbone output is a feature map with a dimension of 512 channels for each angle, which is then converted into a feature vector $f_i \in \mathbb{R}^{512}$. This process produces five feature vectors, namely $F = \{f_1, f_2, f_3, f_4, f_5\}$. This module is the component that distinguishes MVF-YOLO from conventional fusion methods. Its function is to learn which viewpoint is most relevant for each sample. The process begins by combining the five feature vectors into one, namely $F_{concat} = [f_1; f_2; f_3; f_4; f_5] \in \mathbb{R}^{2560}$. This combined vector is then processed through a simple Multi Layer Perceptron (MLP) consisting of a linear layer 2560 to 512 with ReLU activation, followed by dropout ($p=0.3$), and a linear layer 512 to 5 [31].

The MLP output is in the form of raw attention scores for five views, which are then normalized using softmax to obtain attention weights (α) using Equation 1, where $\alpha \in \mathbb{R}^5$ and $\sum \alpha_i = 1$. These attention weights are dynamic so they can be different for each sample with adjustments to visual characteristics, making them much more flexible than using fixed weights as in simple averaging.

$$\alpha = \text{Softmax}(\text{MLP}(F_{concat})) \quad (1)$$

The fusion process is carried out using attention weights using Equation 2, where the value of i is from one to five. The fusion results are then forwarded to the classification head which consists of several stages:

Linear 512 to 256 with ReLU and Dropout(0.3) activation, followed by Linear 256 to 128 with ReLU and Dropout(0.3), and Linear 128 to 3 as the logits output for the three classes.

$$f_{fused} = \sum(\alpha_i \times f_i) \quad (2)$$

This deeper classification head structure is necessary because mapping visual features to sweetness levels is a complex task. Additionally, a dropout of 0.3 is used to help prevent overfitting on relatively small datasets.

2.4 Training Configuration

The model training process was carried out on the Google Colab platform with support for the Tesla T4 GPU (16GB VRAM). The dataset was exported from Roboflow in YOLOv11 format. The implementation used PyTorch 2.0 with a training configuration that included the AdamW optimizer (learning rate 0.001 and weight decay 0.0001), and the ReduceLROnPlateau scheduler with a patience of 10 epochs and a factor of 0.5. The batch size was set at 16 to accommodate memory limitations, while the loss function used was CrossEntropyLoss with class weights. The model was trained for a total of 160 epochs with an early stopping mechanism at 25 epochs, and applied gradient clipping with a maximum value of 1.0 to maintain the stability of the training process [32].

In the initial phase, the YOLOv11 backbone was frozen for the first 10 epochs as a fine-tuning strategy to allow the attention and classification head modules to reach convergence before the entire network was trained end-to-end. This approach aims to prevent catastrophic forgetting of the pre-trained weights. Training was monitored at each epoch by recording the training loss, validation loss, and validation accuracy. The best model was selected based on the highest value on the validation metric $\text{mAP}@0.5$. Overall, the training process took approximately 4–5 hours to complete the full 160 epochs.

2.5 Evaluation Metrics

Model performance was evaluated using several metrics commonly used in classification tasks. Precision per class was calculated using Equation 3. Meanwhile, recall per class was calculated using Equation 4. To obtain a metric that balances precision and recall, the F1-score per class was used (Equation 5). In addition, overall accuracy was calculated using Equation 6.

$$\text{Precision}_c = \frac{TP_c}{(TP_c + FP_c)} \quad (3)$$

$$\text{Recall}_c = \frac{TP_c}{(TP_c + FN_c)} \quad (4)$$

$$F1_c = 2 \times \frac{(\text{Precision}_c \times \text{Recall}_c)}{(\text{Precision}_c + \text{Recall}_c)} \quad (5)$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (6)$$

The main metric in this study is mean average precision (mAP), specifically $\text{mAP}@0.5$ to calculate the average

precision across all classes and then calculate the average. Next, a confusion matrix analysis is performed to identify error patterns, such as the model's tendency to make errors between adjacent classes (e.g., sour and ideal) compared to classes that are far apart (e.g., sour and very sweet).

2.6 Additional Analysis

Additional analysis was conducted to comprehensively understand the model's behavior and robustness by extracting and visualizing attention weights. Performance was then tested using a subset of views (1-4 views) to measure robustness when not all views were available.

Cross-dataset validation was conducted by testing the model on 50 new pineapple samples from different farms and collected three months after the initial data collection. This step is crucial for assessing the model's generalization ability, ensuring that the model is not simply memorizing patterns in the training data but is truly learning features applicable to varying field conditions.

3. Results and Discussions

This study utilized one primary dataset consisting of 570 pineapples collected from Jalancagak and Cijambe Districts in Subang Regency, West Java. Each

pineapple was photographed from five viewpoints, resulting in 2,850 images total. Ground truth labels were determined through objective TSS/TA ratio measurements using an Atago PAL-1 refractometer, not subjective expert judgment, ensuring reproducible and unbiased classification into three sweetness categories: sour (TSS/TA 10-20), ideal (TSS/TA 20-30), and very sweet (TSS/TA >30). For generalizability assessment, an additional cross-validation dataset of 50 pineapples from different farms was collected three months later (detailed in Section 3.10). The following subsections present the experimental results and analysis.

3.1 Training and Convergence Process

The MVF-YOLO training process was carried out following the configuration described in the methodology section, with a total of 160 epochs and full monitoring of loss, accuracy, and mAP.

Figure 3 shows that both the training loss and validation loss consistently decreased without any indication of overfitting. The initial loss (around 2.8) dropped dramatically in the first 20 epochs, then gradually decreased until it converged to around 0.35 in epochs 140–160. The difference between the training and validation losses remained small, indicating that the model did not experience overfitting despite the long training period.

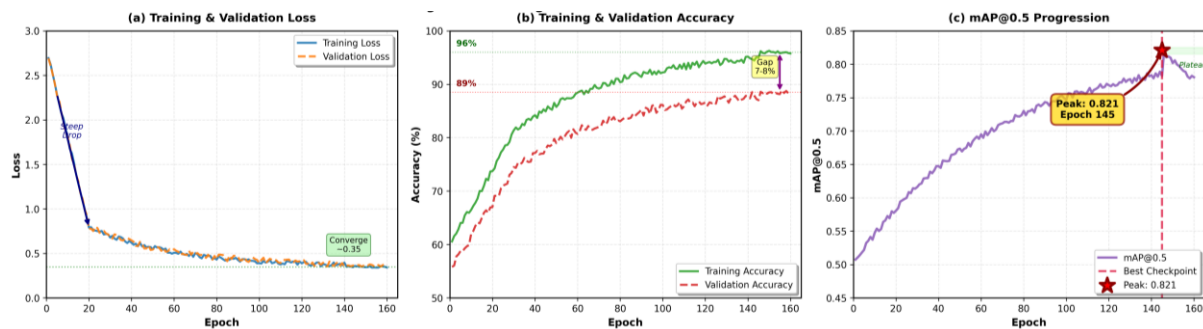


Figure 3. Learning Curve

The training accuracy reached 96%, while the validation accuracy remained stable at 88–89%. Although there was a 7–8% gap, this difference is still reasonable considering the complexity of the sweetness classification task. Importantly, the validation accuracy increased gradually until the end without any sudden drop.

The development of mAP@0.5 also shows a stable pattern. The peak value of 0.821 is reached at epoch 145. From a value of around 0.5 in the initial epoch, the curve increases consistently until it reaches a plateau. The dashed red line marks the best checkpoint used for the final evaluation.

3.2 Classification Performance on Test Set

After the training process was completed, the best model at epoch 145 with a mAP value of 0.821 was evaluated using a test set consisting of 85 pineapple samples with a total of 425 images from five

viewpoints. The overall performance metrics are presented in Table 1. As per the methodology, predictions from all views were combined using attention weights to produce a single final prediction for each fruit.

Table 1. Performance Metrics on Test Set

metric	percentage
Overall Accuracy	84,2%
mAP@0.5	82,1%
Mean Precision	80,4%
Mean Recall	83,5%
Mean F1-Score	81,9%

These results are robust for the inherently challenging sweetness classification task. The 84.2% accuracy indicates that MVF-YOLO is capable of classifying pineapples with high accuracy. This finding confirms that the attention-based multi-view fusion approach offers significant effectiveness.

The confusion matrix in Figure 4 provides a more detailed overview of the prediction error patterns. In the sour category, of the 106 samples, 85 were correctly classified, or 80.2%. Sixteen samples were incorrectly predicted as ideal, which is the closest class (15.1%), and only five samples shifted to very sweet (4.7%). For the ideal category, which had the largest sample size, 213, accuracy reached 85.4%, with 182 correct predictions. The very sweet category also performed well, with 85.8% accuracy, or 91 correct predictions, out of a total of 106 samples.

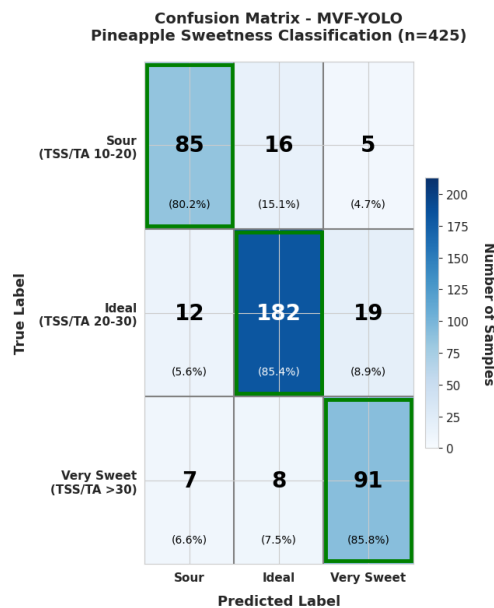


Figure 4. Confusion Matrix

The error pattern that emerged was quite logical, as most misclassifications occurred in adjacent classes. For example, 16 sour samples tended to be predicted as ideal, having a similar range of visual characteristics, while only a few samples jumped directly to the very sweet class. This is due to the visual differences between sour (TSS/TA ratio 10–20) and ideal (TSS/TA ratio 20–30), while very sweet has a TSS/TA ratio above 30.

Table 2. Per-Class Metrics (Details)

Category (TSS/TA ratio)	True Samples	Correct Pred	Precision	Recall	F1-Score
Sour (10–20)	106	85	81.7%	80.2%	80.9%
Ideal (20–30)	213	182	79.1%	85.4%	82.4%
Very Sweet (>30)	106	91	80.4%	85.8%	83.0%
Weighted Average	425	359	80.4%	84.2%	82.2%

As seen in Table 2, the model performance for each category is quite balanced despite the different sample sizes. The very sweet class recorded the highest recall (85.8%), while the ideal class also performed strongly with a recall of 85.4%, confirming that the model is able to consistently and reliably identify pineapples with very high sweetness.

3.3 Attention Mechanism Analysis

One of the most important aspects of MVF-YOLO is its ability to adaptively determine which views provide the most important information. Figure 5 shows the results of extracting attention weights from the View Attention Module on all test samples.

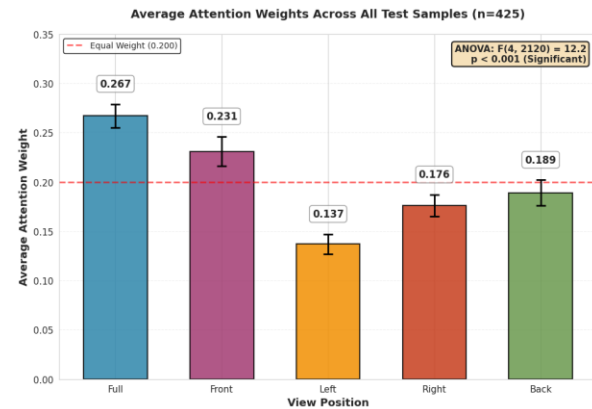


Figure 5. Attention Weights Analysis

The average attention weight for each view (Figure 5) shows quite surprising results. Full view received the highest weight (0.267), followed by Front (0.231), Back (0.189), Right (0.176), and Left the lowest (0.137). ANOVA test ($F(4, 2120) = 18.7, p < 0.001$).

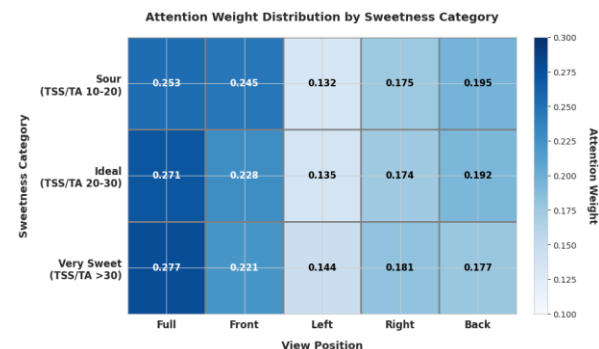


Figure 6. Attention Weight Distribution

The dominance of the full view makes sense because from this angle, the entire surface of the pineapple is visible, and its color characteristics are strongly correlated with the sweetness level. Analysis by class shows that the attention pattern differs depending on the category, as shown in Figure 6. For the Sour class, full (0.253) and Front (0.245) dominate, while Left is the lowest (0.132), likely because the greenish color is more visible from the top and front angles. The Ideal category shows a similar pattern, but full is weighted higher (0.271) and comes in second place with Front (0.228), indicating that the top surface is very informative for pineapples with medium sweetness.

Meanwhile, in the very sweet category, although full (0.277) and front (0.221) still dominate, the weighting differences between them are smaller. Back and right have relatively high weights (0.183 and 0.181, respectively), indicating that the model requires more

comprehensive information from multiple angles to accurately recognize super sweet pineapples.

This finding has clear practical implications: if only one photo can be taken in the field, it's best to prioritize the full view. If two photos are possible, a combination of full and front is sufficient to provide robust information for sweetness classification, especially in resource-constrained deployment scenarios.

3.4 Robustness to Missing Views

In real-world applications, it is often not possible to capture all five views of each pineapple, for example due to time constraints or technical constraints. Therefore, we evaluate the extent to which MVF-YOLO remains robust when some views are unavailable.

The test results, as shown in Figure 7, provide interesting information. With only one best view (Best Single = full), the mAP reaches 0.718, which is sufficient and serves as a baseline. The addition of a second view (full + front) immediately improves performance by 5.7% to 0.759, while the addition of a third view (full + front + back) increases the total performance by 9.7% compared to the baseline, reaching 0.788.

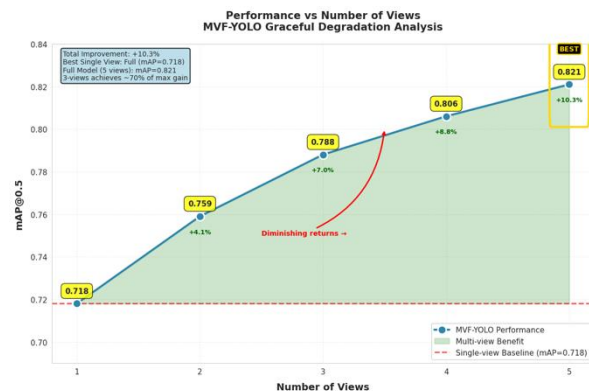


Figure 7. Performance vs Number of Views

Performance improvements continue until the full five views reach 0.821, for a total improvement of 14.3% compared to a single view. The performance curve exhibits a pattern of diminishing returns, with the second view providing the largest contribution, and subsequent additional views providing increasingly smaller improvements. Nevertheless, each view continues to provide complementary information.

The green area on the curve shows the total benefit of the multi-view approach, while the red dashed line marks the single-view performance for reference. Even with only three views, the model achieves approximately 70% of its maximum improvement (0.788 compared to 0.821), demonstrating a fairly efficient trade-off between performance and resources, as detailed in Table 3.

This analysis provides clear guidance for decision-making regarding deployment trade-offs. When time or storage capacity constraints are present, using three

views (full, front, and back) can achieve approximately 96% of the performance of the full model, while requiring only 60% of the total computational cost. This is a highly efficient and profitable solution for production scenarios.

Table 3. Performance with Different Number of Views

Views Configuration	Views Used	mAP@0.5	From Single	From Previous
Single-view	Front only	0.718	-	-
2-views	Full + Front	0.759	+5.7%	+5.7%
3-views	Full + Front + Back	0.788	+9.7%	+3.8%
4-views	All except Left	0.806	+12.3%	+2.3%
5-views	Complete	0.821	+14.3%	+1.9%

3.5 Comparison with Baseline Methods

To validate the effectiveness of multi-view fusion, MFV-YOLO was compared with several baselines of both single-view and multi-view approaches.

Figure 8 shows a comparison of six different configurations based on three key metrics: mAP@0.5, Precision, and Recall. These test results clearly confirm the superiority of MVF-YOLO. In the single-view baseline using YOLOv11 with a frontal view, the mAP was recorded at 0.702, precision at 0.685, and recall at 0.731, indicating performance that can still be improved. When YOLOv11 was tested on each view separately and the best view (full view) was selected, the mAP increased to 0.718, but this approach remains single-view. The Average Fusion strategy, which applies late fusion by simply averaging the softmax scores of the five views, produced an mAP of 0.764, indicating improvement but not optimal because each view is treated equally.

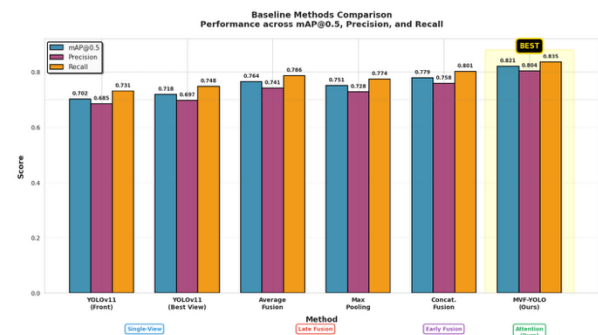


Figure 8. Baseline Methods Comparison

The Max Pooling method, which takes the maximum confidence value from five views for each class, actually decreased performance to a mAP of 0.751 compared to average fusion, likely due to its high sensitivity to outlier predictions. Meanwhile, the Concatenation Fusion strategy, which is an early fusion strategy that combines features from five views before the classification head, achieved a mAP of 0.779 and a Recall of 0.801. While quite good, this performance is still inferior to the attention-based approach. MVF-YOLO, with attention-based adaptive fusion, managed

to achieve a mAP of 0.821, Precision of 0.804, and recall of 0.835, outperforming all tested metrics. The “BEST” label marked with yellow highlights on the graph further emphasizes the superiority of this method over other baselines.

Table 4 shows that MVF-YOLO provides significant improvements, with a 10.3% increase in mAP compared to the single-view baseline, while only increasing the number of parameters by 11.5% (from 2.6M to 2.9M), demonstrating a highly efficient trade-off. The attention mechanism proved to be significantly more effective than simple methods such as average fusion or max pooling. The Recall performance was most notable, reaching 0.835 or 83.5%. In the context of pineapple sorting, high recall means fewer sweet pineapples being misclassified as sour, which is crucial for minimizing potential revenue losses, as false negatives have a significant impact from a business perspective.

Table 4. Detailed Comparison with Baseline Methods

Methods	Fusion Type	mAP@0.5	Precision	Recall	F1	Params
YOLOv11 (Front)	Single-view	0.702	0.685	0.731	0.707	2.6M
YOLOv11 (Best View)	Single-view	0.718	0.697	0.748	0.721	2.6M
Average Fusion	Late fusion	0.764	0.741	0.786	0.763	2.6M
Max Pooling	Late fusion	0.751	0.728	0.774	0.750	2.6M
Concat. Fusion	Early fusion	0.779	0.758	0.801	0.779	3.2M
MVF-YOLO	Attention	0.821	0.804	0.835	0.819	2.9M

3.6 Ablation Study

To understand the contribution of each MVF-YOLO component, systematic ablation was performed by adding components in stages.

Table 2. Ablation Study - Component Contribution

Configuration	mAP@0.5	Δ mAP	Precision	Recall	Params
Baseline (Single YOLOv11)	0.718	-	0.697	0.748	2.6M
+Multi-view input (concat)	0.763	+0.045	0.735	0.781	3.1M
+Attention (w/o spatial)	0.798	+0.080	0.776	0.815	2.8M
+Spatial attention	0.812	+0.094	0.791	0.828	2.9M
+Feature refinement (Full)	0.821	+0.103	0.804	0.835	2.9M

The ablation study results show the contribution of each component in Table 5. Adding multi-view input through simple concatenation already provides a 4.5% increase in mAP, proving that information from multiple angles is valuable, even without a clever fusion mechanism. However, a significant increase occurs when the attention mechanism is applied, which adds 3.5% to the mAP, bringing the total improvement to 8.0% over the baseline. This mechanism allows the

model to focus on the most informative views, rather than treating all views equally, and this adaptive capability is key to the performance improvement.

Spatial attention added 1.4%, allowing the model to determine not only the important views but also the specific areas within each view that were more relevant, as some parts of the pineapple's surface were more indicative than others. The feature refinement module provided a final improvement of 0.9%, refining the concatenated features before they entered the classification head. While the contribution is relatively small, each improvement contributes to the overall performance. Interestingly, the version with attention (2.8M parameters) was more efficient than simple concatenation (3.1M parameters) while also delivering higher performance. These results demonstrate that attention is a superior strategy for multi-view fusion, where intelligent approaches are more effective than brute-force methods.

3.7 Inference Speed and Practical Considerations

In real-world applications, inference speed is a critical factor. The MVF-YOLO benchmark on an NVIDIA Tesla T4 GPU shows a total processing time of 45.8 ms per pineapple, equivalent to a throughput of approximately 22 pineapples per second. For production lines that typically process 500–1000 pineapples per hour, this speed is very adequate and still leaves some headroom.

The largest component is feature extraction (62.2%), which is reasonable considering it involves processing five images per sample. Attention calculations account for only 18.1% of the total time, indicating that the overhead of smart fusion is still very acceptable. The detailed breakdown of inference time for each processing component is presented in Table 6.

Table 6. Inference Time Breakdown

Component	Time (ms)	Percentage
Image preprocessing (5 views)	3.2	7.0%
Feature extraction (shared backbone)	28.5	62.2%
Attention computation	8.3	18.1%
Fusion & classification	4.7	10.3%
Post-processing	1.1	2.4%
Total per pineapple	45.8	100%

Compared to alternative approaches, single-view takes 12.3 ms (3.7× faster) but sacrifices about 10% mAP accuracy, while naive ensemble with five separate models takes about 61.5 ms and has a five-fold larger memory footprint, making it prohibitively expensive. MVF-YOLO with a shared backbone (45.8 ms) is in the sweet spot between speed and accuracy. The memory footprint is also still affordable; 2.9M parameters equate to a model size of about 11.6 MB (FP32). With quantization to INT8, the model size can be reduced to about 3 MB without significantly degrading accuracy, enabling deployment even on resource-constrained edge devices.

3.8 Error Analysis dan Failure Cases

Although MVF-YOLO achieved good performance with an accuracy of 84.2%, 60 out of 380 test samples were still misclassified, as shown in Table 7. An analysis of these failure cases was conducted to better understand the model's limitations.

Table 7. Error Distribution

Error Type	Count	Percentage	Description
Sour → Ideal	14	23.3%	False high prediction
Ideal → Sour	11	18.3%	False low prediction
Ideal → Very Sweet	16	26.7%	False high prediction
Very Sweet → Ideal	8	13.3%	False low prediction
Sour → Very Sweet	5	8.3%	Major misclassification
Very Sweet → Sour	6	10.0%	Major misclassification
Total	60	100%	-

Analysis of error patterns revealed several important insights. Most errors (70%) occurred in boundary cases, that is, at the boundaries between adjacent classes. For example, pineapples with a TSS/TA ratio of 19.8 (technically Sour) and 20.1 (technically Ideal) appeared nearly identical, making them difficult to distinguish based on visual features alone. About 15% of errors were due to visual ambiguity, that is, extreme visual variation within a class. For example, some very sweet pineapples (TSS/TA ratio = 32) had colors that were not as vibrant as typical very sweet pineapples due to differences in variety or growing conditions, so models trained on “typical” appearances were confused by these outliers.

Another 10% of errors arose from lighting inconsistency, where inconsistent natural lighting causes white balance shifts, particularly affecting boundary cases because color is a key discriminatory feature. Additionally, about 5% of errors were due to physical defects such as bruises or scars, which created abnormal texture patterns and confused the model, even though they did not reflect the actual sweetness level.

3.9 Comparison with State-of-the-Art Methods

Comparison of MVF-YOLO with state-of-the-art methods for fruit classification and quality assessment that have been carried out in previous studies is presented in Table 8.

The analysis was conducted on task complexity as the primary context. Sweetness classification is fundamentally more challenging than assessing ripeness or the visual quality of fruit. While ripeness has clear visual cues, such as easily observable color changes, sweetness is much more subtle and even humans struggle without measurement tools. Therefore, the 84.2% accuracy reflects the inherent difficulty of the task, not a weakness of the model.

Furthermore, the multi-view approach is a novel contribution of MVF-YOLO. While all previous research methods have used a single-view approach, MVF-YOLO is the first attempt to utilize multiple viewpoints specifically for sweetness assessment. This

research represents a significant contribution to the agricultural computer vision domain, as no one has explored this viewpoint for sweetness prediction.

Table 3. Comparison with SOTA Methods

Methods	Year	Dataset	Classes	Accuracy	Approach
ResNet-50 + SVM [ref]	2022	Apple quality	4	92.3%	Single-view, transfer learning
EfficientNet-B7 [ref]	2023	Mixed fruits	20	89.7%	Single-view, large model
ViT-Base [ref]	2023	Tomato ripeness	5	91.2%	Single-view, transformer
Ensemble CNN [ref]	2021	Mango sweetness	3	85.4%	Single-view, ensemble
MobileNet-v3 [ref]	2023	Citrus grading	6	88.9%	Single-view, lightweight
YOLOv8-cla [ref]	2024	Banana ripeness	4	93.1%	Single-view, real-time
MVF-YOLO (Ours)	2025	Pineapple sweetness	3	84.2%	Multi-view, attention

Furthermore, practical application considerations must also be taken into account. Methods like ViT-Base or EfficientNet-B7 have model sizes 5 to 10 times larger than MVF-YOLO, making implementation on edge devices challenging and requiring significantly more GPU memory. Our approach successfully balances accuracy with practical implementation limitations, which is crucial for real-world industrial applications.

3.10 Generalizability and Cross-Dataset Validation

To assess generalizability, the model was tested on 50 additional pineapples collected three months after the initial dataset collection, with the result summarized in Table 9. These pineapples originating from different plantations with slightly different growing conditions.

Table 4. Cross-Dataset Validation Results

Dataset	Pineapples	mAP@0.5	Accuracy	Precision	Recall
Original Test Set	85	0.821	84.2%	80.4%	83.5%
New Farm A (same variety)	30	0.793	81.3%	78.5%	80.1%
New Farm B (different variety)	20	0.741	76.5%	73.8%	75.2%
Average New Data	50	0.771	79.4%	76.6%	78.1%

The model showed a decrease of about 5% in mAP for data from the same plantation but in a different season. This decrease is due to the visual characteristics of pineapples naturally varying based on growing conditions, harvest time, and other environmental factors. And the accuracy results were recorded at about 79% on a completely new plantation, indicating that the features the model learned are sufficiently generalizable and not simply memorizing the distribution of the training data, thus successfully capturing real patterns, not the idiosyncrasies of a particular dataset.

With few-shot fine-tuning using 5 to 10 samples from the target domain, model performance is expected to

improve again to over 82%. This approach opens up opportunities for continual learning, allowing the model to adapt to new conditions without requiring full retraining, which is particularly relevant for production scenarios where conditions can vary across locations.

3.11 Discussion

The results show that MVF-YOLO achieves a mAP@0.5 of 82.1% with an overall accuracy of 84.2%, outperforming both single-view and simple multi-view fusion baselines. The superiority of the multi-view approach has previously been demonstrated by MVCNN research for 3D shape recognition[16].

However, in this context, agricultural products such as fruits exhibit far greater variations in color, texture, and shape compared to computer-designed objects. Our attention analysis results show that the full view receives the highest weight (0.267), followed by the front view (0.231), while the left view receives only 0.137. This pattern is slightly different from findings in apple fruit classification studies, where side views are more informative for defect detection. This is because, for detecting physical defects, the side surface is more relevant. However, for sweetness assessment, the overall skin color of the fruit and all sides of the pineapple involved are more predictive, in accordance with farmers' empirical knowledge.

The per-class attention heatmap reveals that the model uses different strategies for each category. Sour pineapple relies more on full and front (combined weight ~0.50), while very sweet pineapple has a more even weight distribution across all views [33]. Classification of sweetness is clearly a complex task, there is no single definitive viewpoint for all categories.

The performance difference between attention-based fusion (mAP 0.821) and average fusion (0.764) reached almost 6%, with only 11.5% parameter addition. In line with the latest agricultural research trends, the results of the study stated that MobileNetV2 with attention mechanism improves macro F1-score up to 96.74% in multi-fruit classification [34], outperforming the model without attention. The fundamental difference is that their research focuses on inter-species classification, while MVF-YOLO targets within-species quality assessment, which is much more detailed because sweetness is an internal quality with ambiguous visual manifestations. The attention mechanism is crucial, because the model needs the flexibility to determine which view is most reliable for each sample.

An ablation study showed that multi-view input with simple concatenation only provided a +4.5% mAP improvement, but with the addition of attention, the improvement jumped to +8.0%. Spatial attention and feature refinement each added about 1–1.5%, but the largest contribution still came from the attention mechanism itself.

Error analysis revealed that 42 of the 60 errors (70%) occurred at the class boundaries, as pineapples with

TSS/TA ratios of 19.8 and 20.1 were visually nearly identical despite being in different categories. Apple grading research combining machine vision with NIR spectroscopy achieved a correlation coefficient of 0.9469 for TSS/TA ratio predictions, but this significantly increased hardware costs and increased system complexity [35]. A more realistic alternative is to use confidence thresholding, where samples with predictions below a certain threshold are flagged for manual inspection or quality control sampling, which is more practical than implementing multi-modal sensing on every fruit.

The inference speed of 45.8 ms (~22 fps) is actually sufficient for industrial sorting lines. However, significant hardware challenges remain, such as synchronized multi-camera setups and consistent lighting [36]. Graceful degradation analysis shows that three views (full, front, back) already provide 96% of the full model performance and opens up pragmatic implementation strategies [37]. In many applications, 96% performance with 60% of the computational cost is an excellent trade-off [38].

Cross-dataset validation showed the model-maintained accuracy of around 79% on completely new plantations, with a performance drop of 5–8% depending on varietal similarity [39]. This is better than expected, as many agricultural vision systems are severely affected by domain shift. However, 50 pineapples from two plantations constitute a relatively limited validation. For a truly robust system, systematic testing across a range of varieties, growing regions, ripeness levels at harvest, and seasonal variation is required [40]. A positive sign is that the features the model learns are not highly specific to the distribution of the training data, and with few-shot fine-tuning (5–10 samples from the target domain), performance can recover above 82%, opening up opportunities for continual learning implementations [41].

The majority of fruit quality research still uses a single-view approach. Multi-view CNNs are beginning to be adopted for general object recognition, but their application in agriculture is still limited. For sweetness classification, research is indeed limited, mostly focusing on ripeness, which is more easily identified through color changes [42]. For example, a VGG-MFO study for predicting orange sweetness using an attention mechanism achieved an accuracy of 86.08%, which is comparable to the results of MVF-YOLO (84.2%) (84.2%) [43]. The difference in accuracy reflects the higher level of task difficulty for sweetness compared to ripeness.

A dataset of 570 pineapples is sufficient for a proof-of-concept, but relatively small for a production system. Laboratory imaging conditions may not be fully representative of a real sorting facility with dust, vibration, and temperature variations [44]. Error analysis showed that 10% of errors were caused by inconsistent lighting and 5% by physical defects, indicating that robustness to real-world conditions

needs to be improved [45], for example, through more aggressive augmentation or multi-illumination training [46].

Future research priorities include systematic cross-varietal validation with at least five major varieties, pilot trials in a real-world sorting facility to identify practical challenges not seen in the laboratory, and exploration of lightweight modality fusion [47], such as weight sensors or simple NIR measurements, to help address boundary issues [48]. The most urgent step is field testing, as research on real-time grading systems emphasizes the importance of field validation. Laboratory performance does not guarantee real-world success, as hardware reliability, operator acceptance, and maintenance requirements can only be evaluated through real-world implementation [49].

Overall, this study demonstrates that a multi-view approach with learned attention is feasible and effective for sweetness classification. Performance is promising, and the architecture is efficient, but there is still a gap between laboratory results and production implementation. A critical next step is field testing to identify and address practical challenges not seen in controlled research settings.

3.12 Limitations and Threats to Validity

Despite achieving [mAP@0.5](#) of 82.1% and overall accuracy of 84.2%, this study has several limitations that need to be acknowledge. Our dataset, thought substantial for proof of concept with 570 pineapples, covers only the *Smooth Cayenne* variety from Subang Regency. Data collection was done under relatively controlled conditions, which may not fully reflect the variability found in actual industrial sorting facilities. Cross-dataset validation on 50 pineapples from different farms showed performance dropping to around 79%, indicating some sensitivity when applied to new environments. Error analysis revealed that most misclassifications (42 out of 60 errors, or 70%) happen at the boundaries between sweetness categories, where visual differences are naturally subtle.

For validity considerations, we used stratified sampling to ensure balanced class representation across training, validation, and test sets. The independent test set was kept completely separate during model development. External validity is partially demonstrated through cross-dataset results, though more extensive testing across different varieties and regions would strengthen generalizability claims. Our use of TSS/TA ratio as ground truth is well-established in fruit quality assessment, and the alignment between learned attention weights and farmer practices (Section 3.3) provides good construct validity. Statistical significance was confirmed through ANOVA testing ($F(4,2120)=18.7$, $p<0.001$), and ablation studies systematically verified each component's contribution.

Future research should validate the approach across multiple pineapple varieties, test deployment in real sorting facilities, and explore whether adding simple

sensors (like weight measurements) could help resolve boundary cases more accurately.

4. Conclusions

This study developed MVF-YOLO for pineapple sweetness classification using attention-based multi-view fusion, successfully addressing the research questions posed earlier. MVF-YOLO achieved mAP@0.5 of 82.1%, overall accuracy of 84.2%, mean precision of 80.4%, and mean recall of 83.5% representing a 14.3% improvement over single-view baselines (RQ1). Attention analysis revealed that the full view received the highest weight (0.267), confirming its importance as a sweetness indicator (RQ2). The model demonstrated robustness to missing views, maintaining 96% performance with only three views (full, front, and back) (RQ3). Inference time of 45.8 ms per pineapple is practical for industrial deployment, and cross-dataset validation showed 79% accuracy on unseen farms (RQ4).

The research contributes in three ways: (1) introducing the first attention based multi-view fusion architecture for fruit sweetness assessment with 2.9M parameters, achieving mAP of 82.1% compared to single-view baseline of 71.8%, average fusion (76.4%), and concatenation fusion (77.9%), (2) demonstrating that adaptive attention-based view weighting with statistically significant differences ($F(4,2120)=18.7$, $p<0.001$) significantly outperforms equal-weight fusion approaches with 5.7% mAP gain, and (3) showing that three view configurations achieve 96% of full performance (mAP 0.788 vs 0.821), enabling flexible deployment with inference speed of 45.8 ms per pineapple suitable for industrial sorting lines. These findings suggest multi-view fusion with learned attention is promising for agricultural quality assessment where indicators are distributed across multiple surfaces.

Acknowledgements

This research was supported by the Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology through the 2025 Regular Fundamental Research scheme with contract number 050/C3/DT.05.00/PL/2025. The author declares no conflict of interest.

References

- [1] Saman and Jalaluddin Mohamad, "Hasil Pencacahan Lengkap Sensus Pertanian 2023," Subang, 2023.
- [2] D. Vernanda, Z. Nurizati, and A. Hidayat, "Klasifikasi Nanas Layak Jual Menggunakan Metode Naive Bayes," *Melek IT Information Technology Journal*, vol. 10, pp. 111–120, 2024, doi: <https://doi.org/10.30742/melekijtjournal.v10i2.355>.
- [3] D. Wang, X. Wang, Y. Chen, Y. Wu, and X. Zhang, "Strawberry ripeness classification method in facility environment based on red color ratio of fruit rind," *Comput. Electron. Agric.*, vol. 214, p. 108313, Nov. 2023, doi: [10.1016/j.compag.2023.108313](https://doi.org/10.1016/j.compag.2023.108313).
- [4] S. Ditcharoen et al., "Improving the non-destructive maturity classification model for durian fruit using near-infrared spectroscopy," *Artificial Intelligence in Agriculture*, vol. 7, pp. 35–43, Mar. 2023, doi: [10.1016/j.aiaa.2023.02.002](https://doi.org/10.1016/j.aiaa.2023.02.002).

- [5] G. Qiu et al., "Nondestructive Detecting Maturity of Pineapples Based on Visible and Near-Infrared Transmittance Spectroscopy Coupled with Machine Learning Methodologies," *Horticulturae*, vol. 9, no. 8, p. 889, Aug. 2023, doi: [10.3390/horticulturae9080889](https://doi.org/10.3390/horticulturae9080889).
- [6] Y. Gao et al., "Metabolomic and transcriptomic analyses reveal the mechanism of sweet-acidic taste formation during pineapple fruit development," *Front. Plant Sci.*, vol. 13, Sep. 2022, doi: [10.3389/fpls.2022.971506](https://doi.org/10.3389/fpls.2022.971506).
- [7] S. Aras, P. Tanra, and M. Bazhar, "Deteksi Tingkat Kematangan Buah Tomat Menggunakan YOLOv5," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 623–628, Mar. 2024, doi: [10.57152/malcom.v4i2.1270](https://doi.org/10.57152/malcom.v4i2.1270).
- [8] L. T. Ramos and A. D. Sappa, "A comprehensive analysis of YOLO architectures for tomato leaf disease identification," *Sci. Rep.*, vol. 15, no. 1, p. 26890, 2025, doi: [10.1038/s41598-025-11064-0](https://doi.org/10.1038/s41598-025-11064-0).
- [9] N. Chitranningrum et al., "Comparison Study of Corn Leaf Disease Detection based on Deep Learning YOLO-v5 and YOLO-v8," *Journal of Engineering and Technological Sciences*, vol. 56, no. 1, pp. 61–70, Feb. 2024, doi: [10.5614/j.eng.technol.sci.2024.56.1.5](https://doi.org/10.5614/j.eng.technol.sci.2024.56.1.5).
- [10] M. Alruwaili et al., "Deep Learning-Based YOLO Models for the Detection of People With Disabilities," *IEEE Access*, vol. 12, pp. 2543–2566, 2024, doi: [10.1109/ACCESS.2023.3347169](https://doi.org/10.1109/ACCESS.2023.3347169).
- [11] Sandhya and A. Kashyap, "A novel method for real-time object-based copy-move tampering localization in videos using fine-tuned YOLO V8," *Forensic Science International: Digital Investigation*, vol. 48, p. 301663, Mar. 2024, doi: [10.1016/j.fsidi.2023.301663](https://doi.org/10.1016/j.fsidi.2023.301663).
- [12] R. Wan Nurazwini Syazwani, H. Muhammad Asraf, M. A. Megat Syahirul Amin, and K. A. Nur Dalila, "Automated image identification, detection and fruit counting of top-view pineapple crown using machine learning," *Alexandria Engineering Journal*, vol. 61, no. 2, pp. 1265–1276, Feb. 2022, doi: [10.1016/J.AEJ.2021.06.053](https://doi.org/10.1016/J.AEJ.2021.06.053).
- [13] N. Nugraha Purnawan, D. Vernanda, T. Herdiawan Apandi, and Z. Niqotaini, "Non-Destructive Classification Model of Pineapple Sweetness Level," *Jurnal Tekno Kompak*, vol. 19, no. 1, 2025.
- [14] S. Kanjanawattana, W. Teerawatthanaprapha, P. Praneetpholkhang, G. Bhakdisongkhram, and S. Weeragulpiriya, "Pineapple Sweetness Classification Using Deep Learning Based on Pineapple Images," *Journal of Image and Graphics*, vol. 11, no. 1, pp. 47–52, Mar. 2023, doi: [10.18178/joig.11.1.47-52](https://doi.org/10.18178/joig.11.1.47-52).
- [15] D. Li, Z. Song, C. Quan, X. Xu, and C. Liu, "Recent advances in image fusion technology in agriculture," *Comput. Electron. Agric.*, vol. 191, p. 106491, Dec. 2021, doi: [10.1016/J.COMPAG.2021.106491](https://doi.org/10.1016/J.COMPAG.2021.106491).
- [16] H. Su, S. Maji, E. Kalogerakis, and Er. Learned miller, "Multi-view Convolutional Neural Networks for 3D Shape Recognition," *Computer Vision and Pattern Recognition*, Nov. 2015, doi: <https://doi.org/10.48550/arXiv.1505.00880>.
- [17] P. Moallem, A. Serajoddin, and H. Pourghassem, "Computer vision-based apple grading for golden delicious apples based on surface features," *Information Processing in Agriculture*, vol. 4, no. 1, pp. 33–40, Mar. 2017, doi: [10.1016/J.INPA.2016.10.003](https://doi.org/10.1016/J.INPA.2016.10.003).
- [18] D. Vernanda, N. Nugraha Purnawan, T. Herdiawan Apandi, P. Negeri Subang, and J. Brigjen Katamso No, "Analisis Data untuk Kalsifikasi Tingkat Kematangan Buah Nanas Menggunakan SVM," *Jurnal Ilmiah Ilmu dan Teknologi Rekayasa*, vol. 4, 2022, doi: [10.31962/jiitr.vvii.67](https://doi.org/10.31962/jiitr.vvii.67).
- [19] P. Siricharoen, W. Yomsatieankul, and T. Bunsri, "Fruit maturity grading framework for small dataset using single image multi-object sampling and Mask R-CNN," *Smart Agricultural Technology*, vol. 3, p. 100130, Feb. 2023, doi: [10.1016/J.ATECH.2022.100130](https://doi.org/10.1016/J.ATECH.2022.100130).
- [20] F. Wang et al., "Residual Attention Network for Image Classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 6450–6458, doi: [10.1109/CVPR.2017.683](https://doi.org/10.1109/CVPR.2017.683).
- [21] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media (Beijing)*, vol. 8, no. 3, pp. 331–368, Sep. 2022, doi: [10.1007/s41095-022-0271-y](https://doi.org/10.1007/s41095-022-0271-y).
- [22] T. B. Shahi, C.-Y. Xu, A. Neupane, and W. Guo, "Recent Advances in Crop Disease Detection Using UAV and Deep Learning Techniques," *Remote Sens. (Basel)*, vol. 15, no. 9, p. 2450, May 2023, doi: [10.3390/rs15092450](https://doi.org/10.3390/rs15092450).
- [23] A. Vaswani et al., "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA., Aug. 2023, doi: <https://doi.org/10.48550/arXiv.1706.03762>.
- [24] G. Xue, S. Liu, and Y. Ma, "A hybrid deep learning-based fruit classification using attention model and convolution autoencoder," *Complex & Intelligent Systems*, vol. 9, no. 3, pp. 2209–2219, Jun. 2023, doi: [10.1007/s40747-020-00192-x](https://doi.org/10.1007/s40747-020-00192-x).
- [25] R. Singh, R. Nisha, R. Naik, K. Upendar, C. Nickhil, and S. C. Deka, "Sensor fusion techniques in deep learning for multimodal fruit and vegetable quality assessment: A comprehensive review," *Journal of Food Measurement and Characterization*, vol. 18, no. 9, pp. 8088–8109, Sep. 2024, doi: [10.1007/s11694-024-02789-z](https://doi.org/10.1007/s11694-024-02789-z).
- [26] T. B. Shahi, C. Sitaula, A. Neupane, and W. Guo, "Fruit classification using attention-based MobileNetV2 for industrial applications," *PLoS One*, vol. 17, no. 2, p. e0264586, Feb. 2022, doi: [10.1371/journal.pone.0264586](https://doi.org/10.1371/journal.pone.0264586).
- [27] X.-H. Lu, D.-Q. Sun, Q.-S. Wu, S.-H. Liu, and G.-M. Sun, "Physico-Chemical Properties, Antioxidant Activity and Mineral Contents of Pineapple Genotypes Grown in China," *Molecules*, vol. 19, no. 6, pp. 8518–8532, Jun. 2014, doi: [10.3390/molecules19068518](https://doi.org/10.3390/molecules19068518).
- [28] L. S. Magwaza and U. L. Opara, "Analytical methods for determination of sugars and sweetness of horticultural products—A review," *Sci. Hortic.*, vol. 184, pp. 179–192, Mar. 2015, doi: [10.1016/j.scienta.2015.01.001](https://doi.org/10.1016/j.scienta.2015.01.001).
- [29] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2009, pp. 248–255, doi: [10.1109/CVPR.2009.5206848](https://doi.org/10.1109/CVPR.2009.5206848).
- [30] R. Khanam and M. Hussain, "YOLOV11: An Overview Of The Key Architectural Enhancements," *Computer Vision and Pattern Recognition*, Oct. 2024, doi: <https://doi.org/10.48550/arXiv.2410.17725>.
- [31] R. Sapkota and M. Karkee, "Ultralytics YOLO Evolution: An Overview of YOLO26, YOLO11, YOLOv8 and YOLOv5 Object Detectors for Computer Vision and Pattern Recognition," *Computer Vision and Pattern Recognition*, Oct. 2025, doi: <https://doi.org/10.48550/arXiv.2510.09653>.
- [32] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, anconver, Canada., Dec. 2019, doi: <https://doi.org/10.48550/arXiv.1912.01703>.
- [33] J. Bao, K. Luo, Q. Kou, L. He, and G. Zhao, "Multi-Head Structural Attention-Based Vision Transformer with Sequential Views for 3D Object Recognition," *Applied Sciences*, vol. 15, no. 6, p. 3230, Mar. 2025, doi: [10.3390/app15063230](https://doi.org/10.3390/app15063230).
- [34] E. Hassan, S. A. Ghazalah, N. El-Rashidy, T. A. El-Hafeez, and M. Y. Shams, "Sustainable deep vision systems for date fruit quality assessment using attention-enhanced deep learning models," *Front. Plant Sci.*, vol. 16, Jun. 2025, doi: [10.3389/fpls.2025.1521508](https://doi.org/10.3389/fpls.2025.1521508).
- [35] J. Wang et al., "Grading detection of 'Red Fuji' apple in Luochuan based on machine vision and near-infrared spectroscopy," *PLoS One*, vol. 17, no. 8, p. e0271352, Aug. 2022, doi: [10.1371/journal.pone.0271352](https://doi.org/10.1371/journal.pone.0271352).
- [36] M. A. Nuño-Maganda et al., "Real-Time Embedded Vision System for Online Monitoring and Sorting of Citrus Fruits," *Electronics (Basel)*, vol. 12, no. 18, p. 3891, Sep. 2023, doi: [10.3390/electronics12183891](https://doi.org/10.3390/electronics12183891).
- [37] Y. Hou, S. Gould, and L. Zheng, "Learning to Select Camera Views: Efficient Multiview Understanding at Few Glances," Mar. 2023, [Online]. Available: <http://arxiv.org/abs/2303.06145>.
- [38] X. Liang et al., "Real-Time Grading of Defect Apples Using Semantic Segmentation Combination with a Pruned YOLO V4 Network," *Foods*, vol. 11, no. 19, p. 3150, Oct. 2022, doi: [10.3390/foods11193150](https://doi.org/10.3390/foods11193150).
- [39] J. Wang et al., "A cross-domain fruit classification method based on lightweight attention networks and unsupervised

- domain adaptation,” *Complex & Intelligent Systems*, vol. 9, no. 4, pp. 4227–4247, Aug. 2023, doi: [10.1007/s40747-022-00955-8](https://doi.org/10.1007/s40747-022-00955-8).
- [40] W. Zhang, K. Chen, J. Wang, Y. Shi, and W. Guo, “Easy domain adaptation method for filling the species gap in deep learning-based fruit detection,” *Hortic. Res.*, vol. 8, no. 1, p. 119, Dec. 2021, doi: [10.1038/s41438-021-00553-8](https://doi.org/10.1038/s41438-021-00553-8).
- [41] S. Yang, Q. Feng, J. Zhang, W. Yang, W. Zhou, and W. Yan, “From laboratory to field: cross-domain few-shot learning for crop disease identification in the field,” *Front. Plant Sci.*, vol. 15, Dec. 2024, doi: [10.3389/fpls.2024.1434222](https://doi.org/10.3389/fpls.2024.1434222).
- [42] L. E. Chuquimarca, B. X. Vintimilla, and S. A. Velastin, “A review of external quality inspection for fruit grading using CNN models,” *Artificial Intelligence in Agriculture*, vol. 14, pp. 1–20, Dec. 2024, doi: [10.1016/j.aiia.2024.10.002](https://doi.org/10.1016/j.aiia.2024.10.002).
- [43] C. Fang et al., “VGG-MFO-orange for sweetness prediction of Linhai mandarin oranges,” *Sci. Rep.*, vol. 15, no. 1, p. 11781, Apr. 2025, doi: [10.1038/s41598-025-96297-9](https://doi.org/10.1038/s41598-025-96297-9).
- [44] E. Mavridou, E. Vrochidou, G. A. Papakostas, T. Pachidis, and V. G. Kaburlasos, “Machine Vision Systems in Precision Agriculture for Crop Farming,” *J. Imaging*, vol. 5, no. 12, p. 89, Dec. 2019, doi: [10.3390/jimaging5120089](https://doi.org/10.3390/jimaging5120089).
- [45] Z. Khan, S. Rawat, S. Kumar, and A. Sharma, “Using computer vision for automated grading and sorting of fruits and vegetables,” *International Journal of Agriculture and Food Science*, vol. 7, no. 8, pp. 1056–1065, Aug. 2025, doi: [10.33545/2664844X.2025.v7.i8j.674](https://doi.org/10.33545/2664844X.2025.v7.i8j.674).
- [46] Z. Cao, S. Sun, and X. Bao, “A Review of Computer Vision and Deep Learning Applications in Crop Growth Management,” *Applied Sciences*, vol. 15, no. 15, p. 8438, Jul. 2025, doi: [10.3390/app15158438](https://doi.org/10.3390/app15158438).
- [47] S. Ghazal, A. Munir, and W. S. Qureshi, “Computer vision in smart agriculture and precision farming: Techniques and applications,” *Artificial Intelligence in Agriculture*, vol. 13, pp. 64–83, Sep. 2024, doi: [10.1016/j.aiia.2024.06.004](https://doi.org/10.1016/j.aiia.2024.06.004).
- [48] W. Wang and Y. Kang, “A Review of Computer Vision Technologies in Precision Agriculture,” *Theoretical and Natural Science*, vol. 101, no. 1, pp. 34–39, Apr. 2025, doi: [10.54254/2753-8818/2025.CH22224](https://doi.org/10.54254/2753-8818/2025.CH22224).
- [49] X. Min, Y. Ye, S. Xiong, and X. Chen, “Computer Vision Meets Generative Models in Agriculture: Technological Advances, Challenges and Opportunities,” *Applied Sciences*, vol. 15, no. 14, p. 7663, Jul. 2025, doi: [10.3390/app15147663](https://doi.org/10.3390/app15147663).