Journal of Applied Engineering and Technological Science
Vol 6(2) 2025: 984-996

COCOA RIPENESS CLASSIFICATION USING VISION TRANSFORMER

Febryanti Sthevanie[1*], Untari Novia Wisesty[2], Gia Septiana Wulandari[3], Kurniawan Nur Ramadhani[4]
School of Computing, Telkom University, Bandung 40257, Indonesia[1234]
Center of Excellence Artificial Intelligence for Learning and Optimization, Telkom University, Bandung 40257, Indonesia[124]
sthevanie@telkomuniversity.ac.id[1], untarinw@telkomuniversity.ac.id[2], giaseptiana@telkomuniversity.ac.id[3], kurniawannr@telkomuniversity.ac.id[4]

Received: 01 December 2024, Revised: 30 April 2025, Accepted: 05 May 2025
**Corresponding Author*

ABSTRACT
The quality of manual methods for assessing the ripeness of cocoa pods is subjective and varies from one person to another because of the intense labor required and variation of light and background conditions within the field. This research implemented an automated classification approach for cocoa ripeness classification utilizing Vision Transformer (ViT) with Shifted Patch Tokenization (SPT) and Locality Self Attention (LSA) to improve classification accuracy. The model proposed in this research achieved an accuracy of 82.65% and a macro F1 score of 82.71 on the exam with 1,559 images captured under varying illumination backgrounds and complex scenes. The model also proved better than baseline CNN architectures such as VGG, MobileNet, and ResNet in identifying visually progressive stages of ripeness and demonstrated greater generalization in cocoa ripeness classification. The findings of this research indicate the benefits of reducing manual intervention with careful inspection without compromising quality assurance standards in cocoa production. This work demonstrates new ways of applying transformer models to address computer vision problems in agriculture which is a step towards precision and smart farming.
Keywords: *Cocoa Ripeness Classification, Vision Transformer, Shifted Patch Tokenization, Locality Self Attention, Agricultural Computer Vision.*

1. Introduction

In 2021, Indonesia stood third in global cocoa production with 728,046 tons produced annually(Food and Agriculture Organization (FAO), 2023). The cocoa sector is a sub-strategic industry in Indonesia as more than 95% of Sulawesi, Sumatra and Papua's cocoa production comes from smallholder farmers(International Cocoa Organization (ICCO), 2022). These farmers have very low uptake of mechanisation and agricultural technology, thus applying traditional methods. This leads to a range of quality issues as well as post-harvest inefficiency. One of the most important steps in cocoa processing is harvesting. The degree of pod matureness is a key factor in determining the fermentation quality, chemical composition, and flavor profile of chocolate products. If harvested too early, beans have a high tendency of being under fermented resulting in a bitter taste. Milder over mature pods are more susceptible to being fungal infested, internally germinated, or enveloped in mucilage, thus degrading both bean quality and yield(Siregar et al., 2023). This accentuates the need to maintain certain standards and thus, highlights the importance of timely and accurate cocoa matureness assessment.

There is still no automated classification system available for mature classification in Indonesia since it is typically done manually through observation of the pod surface color. This method relies on the use of human labor which is prone to human error especially in fields where lighting and scenery are inconsistent. There is a need to design a classification system that works accurately within the constraints of the real world and can decrease post-harvest loss. Other fruits have seen a positive outcome with using computer vision and deep learning for detecting maturity levels. Previous works included the use of color-based features papaya with k-nearest neighbor (k-NN)(Suban et al., 2020), support vector machines (SVM) for banana(Juncai et al., 2015), oil palm(Siregar et al., 2023) and durian with convolutional neural networks (CNNs)(Kharamat et al., 2020). CNNs are poorly suited for tasks involving modeling

long-range dependencies and tend to overfit smaller, unbalanced training datasets, which is typical for agricultural datasets. Moreover, CNN-based models often fail to perform adequately in uncontrolled scenarios. With these reasons, this research aims to develop an image-based cocoa maturity classification system using ViT architecture. ViT treats images as sequences of patches and employs self-attention clustering to embody patterns. This method aids in the interpretation of color gradients, texture, and pod surface details better(Dosovitskiy et al., 2020). To improve model performance on small datasets and in complicated environmental conditions, two complementary modules are added. Shifted Patch Tokenization (SPT), which rearranges the spatial order of patches and simulates data augmentation(Lee et al., 2021). Locality Self-Attention (LSA) enhances attention while minimizing focus on non-local feature tokens and strengthens local feature emphasis(Zhou et al., 2021).

The effectiveness of Transformer-based models has been explored in agriculture for fruit grading, weed and crop classification, and plant disease detection, achieving high performance even with limited data (Charco et al., 2024; Chitta et al., 2024). In the specific context of cocoa, Lopes et al. (Lopes et al., 2022) demonstrated the potential of deep learning for cocoa bean grading using CNNs. However, ViT-based models for cocoa matureness classification remain underexplored and lack integration of modules that enhance data efficiency. To better understand the matureness classes of cocoa fruits targeted in this research, Figure 1 illustrates the visual differences across the Immature, Mature, and Overmature stages. Each stage shows distinct characteristics in pod surface color and texture, providing observable cues for automated classification systems.

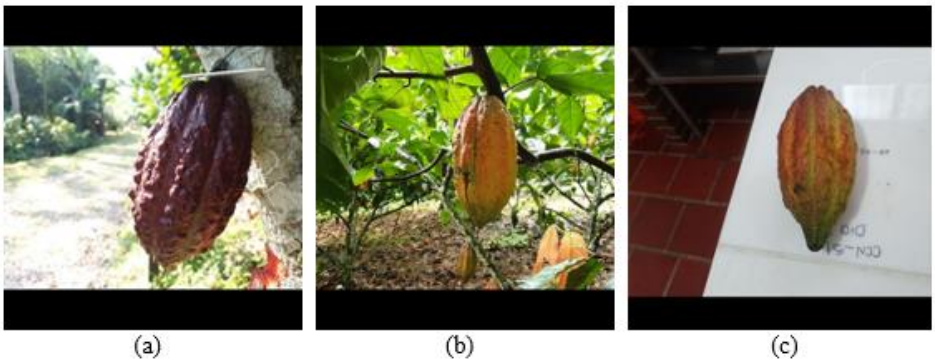

Fig. 1. Cocoa fruit matureness stages: (a) Immature, (b) Mature, (c) Overmature.

The objective of this research is to develop an image-based cocoa fruit matureness classification model using the Vision Transformer architecture, enhanced with Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA), in order to improve classification accuracy under limited-data and field-based image conditions. The contributions of this research are summarized as follows:

a. Proposing a novel ViT-based approach for classifying cocoa matureness stages from real-environment images.
b. Integrating SPT and LSA to address challenges related to small datasets and subtle feature distinctions.
c. Demonstrating that the proposed method outperforms baseline models such as ViT, ResNet, and MobileNet in terms of accuracy and class-wise performance.
d. Providing a practical solution to assist cocoa farmers in making harvesting decisions and improving post-harvest outcomes through automation.

2. Related Works

Image-based classification is a fundamental approach in smart agriculture, supporting applications such as matureness prediction (El Sakka et al., 2024; Khaki & Wang, 2019), disease diagnosis (Liu & Wang, 2021), and yield estimation (Khaki et al., 2020). CNNs have been widely used in classifying crops, detecting diseases, and grading fruit maturity due to their strong spatial feature extraction capabilities (El Sakka et al., 2024; Liu & Wang, 2021). For example, MobileNet and EfficientNet have been applied to lightweight edge devices (Paneru et al., 2024; Yasin & Fatima, 2023), while ResNet and Inception architectures were effective in

complex background conditions (Khaki et al., 2020; Liu & Wang, 2021). Initial work relied on handcrafted color features classified by traditional algorithms like SVM and k-NN (Ala'a & Ibrahim, 2024; Alimjan et al., 2018; Joshi et al., 2023), but deep learning has since replaced them with robust end-to-end CNN pipelines (El Sakka et al., 2024). Transfer learning with pretrained models such as VGG16 and ResNet50 has improved generalization under limited data conditions (Khaki et al., 2020; Khaki & Wang, 2019). Despite these advances, CNNs often overfit small datasets and struggle to capture global context (Brigato & Iocchi, 2021; Gal & Ghahramani, 2016; Yu et al., 2019).

Vision Transformers (ViTs) have gained attention as a potential solution by modeling images as sequences of patches and applying self-attention for global information encoding (Dosovitskiy et al., 2020; Khan et al., 2022). ViTs have outperformed CNNs in leaf disease classification (El Sakka et al., 2024), matureness prediction (Ergün, 2025), and weed detection via UAV (Reedha et al., 2022; Zhao et al., 2023). However, standard ViTs are data-hungry and computationally intensive (Lee et al., 2021). To address this, structural improvements like Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA) have been proposed (Lee et al., 2021; Zhou et al., 2021). SPT introduces overlapping local views, while LSA narrows the attention field for improved locality modeling (Zhou et al., 2021). These methods have enabled ViTs to generalize better on small datasets while maintaining global sensitivity (Lee et al., 2021). In agricultural domains, ViTs with SPT/LSA have shown promise in high-resolution tasks such as plant health monitoring (Borhani et al., 2022; De Silva & Brown, 2023) and tomato leaf recognition (Nahak et al., 2025). UAV-based ViT applications have also reported improved segmentation in irregular lighting (Zhao et al., 2023). ViTs are further enhanced by explainability tools like Grad-CAM (Kulkarni et al., 2022; Mishra & Malhotra, 2024). Recent researches confirm ViTs are competitive or superior to CNNs in grape quality grading (Pothen & Nuske, 2016; Shimazu et al., 2024), banana defect detection (Ergün, 2025), and rice crop classification (Ulukaya & Deari, 2025). Several works integrated multispectral inputs into ViT pipelines for more robust modeling under climate variability (Lin et al., 2023; Rad, 2024). Others employed multi-task learning to combine maturity prediction and yield estimation in parallel(Lin et al., 2023).

While these advancements are encouraging, there remains a lack of ViT-based researches targeting cocoa fruit classification. Existing CNN approaches focus on post-harvest bean sorting (Eric et al., 2023; Essah et al., 2022) and often exclude the critical field-stage matureness identification. No prior work to date has evaluated ViTs with SPT and LSA on real-world cocoa fruit images. To bridge this gap, our research introduces a ViT-SPT-LSA model specifically designed for cocoa matureness classification in uncontrolled environments. This architecture balances global attention with local detail sensitivity to enhance performance with limited image data.

3. Research Methods

3.1. Dataset Preprocessing

Our research used a dataset consisting of 1,559 labeled images of cocoa fruits captured under natural lighting conditions with varying backgrounds, angles, and matureness levels. The dataset includes three maturity classes: immature (45.73%), mature (23.04%), and overmature (31.23%). Input images were resized to 224x224 pixels to match the size of Vision Transformer (ViT) input layer. Data augmentation techniques such as horizontal flipping, rotation, brightness manipulation, and zoom were applied to increase training set diversity and reduce overfitting. All images were normalized using ImageNet mean and standard deviation. We used 1000 images for training set and 559 images for testing set. In the training process, we took 10% of the training set for validation process. The class-wise distribution across each subset is presented in Table 1.

Table 1 - Class distribution across dataset subsets.

Label	Training Set	Validation Set	Testing Set
Immature	405	45	263
Mature	180	20	159
Overmature	315	35	137

Total	900	100	559

3.2 Model Architecture

Figure 2 shows the design of the system that was built in this research. The system starts with image input. The first process in the system is patch embedding which breaks the image into a collection of patches. From the collection of existing patches, an encoding process is then carried out to produce a series of features. This series of features is then processed by the Transformer Encoder in the form of multihead Attention blocks. The output from the Transformer Encoder is a feature that will be used as input for the Softmax Classifier process at the end of the system.

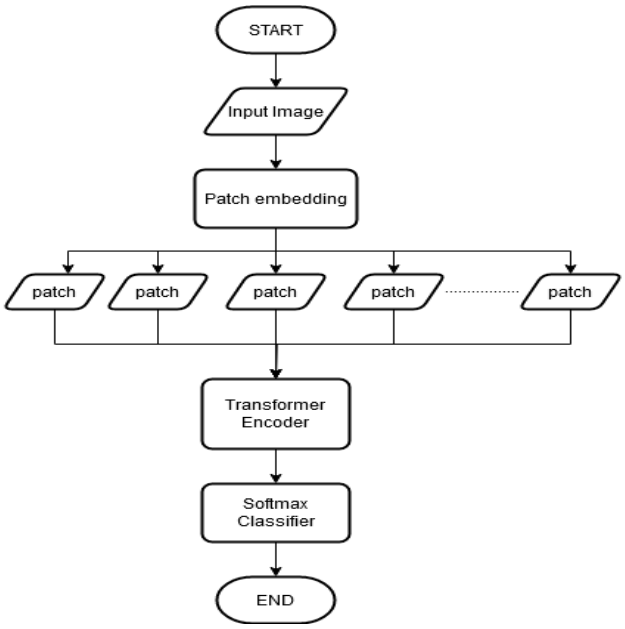

Fig. 2. Proposed Cocoa Ripeness Classification System

3.2 Vision Transformer

The Vision Transformer (ViT) architecture was selected due to its ability to model long-range dependencies via self-attention, which is beneficial for detecting subtle matureness cues spread across fruit surfaces. CNNs, although effective in local pattern recognition, struggle with capturing global context, especially in images with diverse lighting and background conditions. To enhance ViT performance in limited-data settings, we integrated two architectural improvements: Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA). SPT modifies the patch embedding process by shifting input images in fixed directions before tokenization. This encourages the model to capture more diverse and local contextual patterns within the input, thereby improving feature representation. Meanwhile, LSA restricts the self-attention mechanism to only focus on local neighborhoods rather than the entire image, embedding inductive bias that is useful for learning localized features. This modification reduces overfitting and increases model robustness in datasets with limited training samples. Patch size was empirically set to 16x16 pixels, balancing spatial granularity and computational efficiency. This patch size ensures sufficient detail is retained while maintaining compatibility with standard ViT pretraining checkpoints.

Figure 3 depicts the architecture of Vision Transformer (ViT), in which an input image is separated into a series of fixed-size patches, linearly embedded, and coupled with positional encoding before being sent through several transformer encoder layers. These layers are made up of multi-head self-attention and feedforward networks, which record global visual dependencies and enable robust categorization.

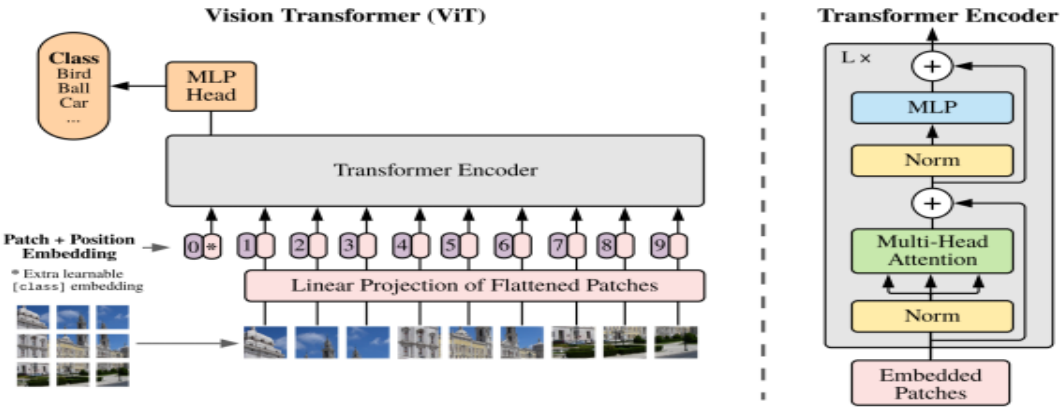

Fig. 3. Vision Transformer architecture (M.-H. Guo et al., 2022).

3.3 Shifted Patch Tokenization (SPT)

Vision Transformer needs a large amount of data in the training process. This is a problem if the number of data provided is limited. To overcome the problem, we used shifted patch tokenization (Emmamuel et al., 2022). This process enrich the variation of data by shifting the patching process and added it to the training set. Figure 4 depicts the Shifted Patch Tokenization (SPT) mechanism, in which input images are spatially shifted in multiple directions (e.g., up, down, left, right). Each shifted image is then tokenized, and the resulting tokens are aggregated to produce a richer patch representation. This enhances the model's capacity to learn local variations in spatial structures. Figure 5 shows the example of SPT result.

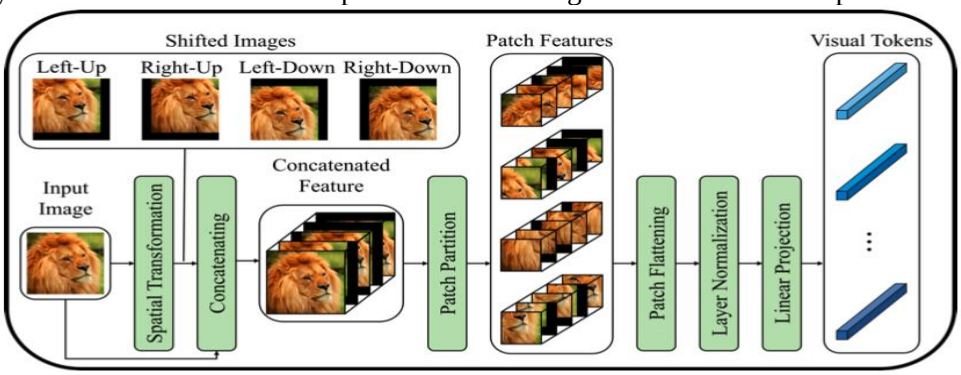

Fig. 4. Shifted Patch Tokenization process(Emmamuel et al., 2022).

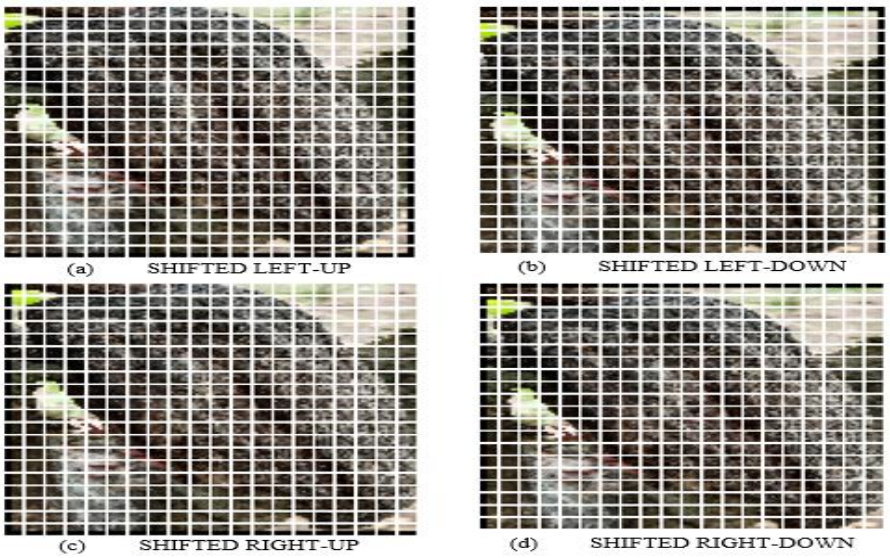

Fig. 5. Example of SPT result.

3.4 Locality Self Attention (LSA)

Another technique to overcome the limited dataset problem is by using Locality Self Attention (LSA) (Q. Guo et al., 2019). LSA improves the distribution of attention scores by determining the temperature parameters of the softmax function. The learnable temperature scaling enables ViT to determine the softmax temperature throughout the learning phase. The softmax temperature is low, which sharpens the score distribution. As a result, the learnable temperature scaling refines the distribution of attention scores. In addition, diagonal masking is applied to eliminate self-token relations by suppressing the diagonal elements of the similarity matrix generated from Query and Key computations. This technique ensures that inter-token relationships receive higher importance by excluding self-token relations from the softmax calculation. Diagonal masking achieves this by assigning $-\infty$ to the diagonal elements, directing the Vision Transformer's attention mechanism to prioritize other tokens rather than focusing on itself. This masking enhances the relative attention scores between distinct tokens, leading to a sharper distribution of attention scores. Consequently, LSA strengthens the locality inductive bias by encouraging the Vision Transformer's attention to concentrate on local regions. Figure 6 shows a schematic comparison between standard self-attention and Locality Self-Attention (LSA). In the standard version, attention is computed globally across all patches. In contrast, LSA applies a localized attention mask, limiting the scope to adjacent patches and encouraging the model to focus on spatially relevant regions, which is particularly beneficial in small datasets.

Fig. 6. Comparison Of Standard Self-Attention and LSA.

3.5 Experiment Setup

The model was trained using Adam optimizer for 100 epochs with an initial learning rate of 1×10^{-4} and batch size of 32. Initial experiments indicated that using a learning rate of 1×10^{-4} resulted in faster convergence and less overfitting. The number of epochs of 100 was used because for epochs above 100, the system performance tends to be stable. The batch size of 32 was chosen to balance the training speed with the available computing power. The Adam optimizer was chosen because it combines AdaGrad and RMSprop so that it can adjust the learning rate for each parameter, improving convergence and performance especially on noisy and sparse data. A learning rate scheduler with cosine annealing was employed. Cross-entropy loss was used as the objective function. Evaluation metrics included accuracy, precision, recall, and F1-score to assess classification performance across all classes. Training was conducted on a workstation equipped with an NVIDIA RTX 4070 GPU (12GB VRAM), 12.6 GB RAM, and an Intel Core i9 processor. Average training time per epoch was approximately 90 seconds. The model achieved convergence by the 60th epoch, after which performance saturated.

4. Results and Discussion

In this section, we present a detailed analysis of the experimental results, including training performance, class-specific performance analysis with a confusion matrix, comparative evaluation, cross-validation results, and ablation study. First, we can see the training performance of Vision Transformer and our proposed method respectively on Figure 7 and 8. The effect of utilizing SPT and LSA in the training process can be seen by comparing the training and validation accuracy. We can see that by combining Vision Transformer with SPT and LSA, the model obtained a higher validation accuracy (about 75%) than by using only Vision Transformer (about 60%). The model also achieved the highest validation accuracy faster than by using only Vision Transformer. Because the gap between training and validation accuracy was smaller, the combination of Vision Transformer with SPT and LSA had a better performance in handling overfitting problem.

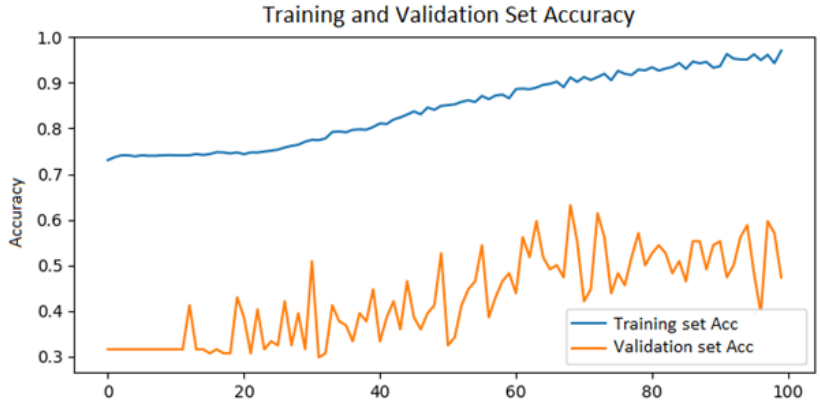

Fig. 7. Training and validation accuracy of Vision Transformer.

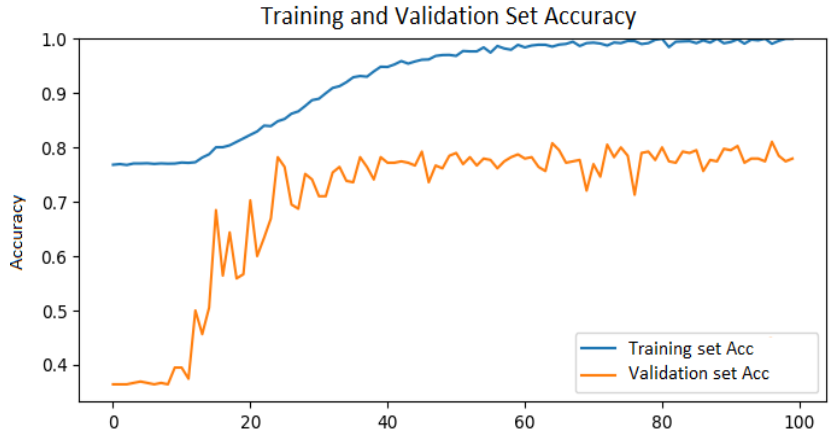

Fig. 8. Training and Validation Accuracy Of Our Proposed Method.

For the testing performance, our proposed ViT-SPT-LSA model was evaluated using four metrics: accuracy, precision, recall, and F1-score, across three cocoa matureness classes: immature, mature, and overmature. The model achieved an overall accuracy of 82.64% and a macro-average F1-score of 0.82.92. The proposed model showed marked improvements in the classification of the mature class, which often exhibits overlapping visual traits with the overmature class. The inclusion of SPT and LSA enabled the model to better generalize to these subtle inter-class differences, outperforming the baseline ViT in both precision and recall. Class-wise performance based on the confusion matrix in Figure 9 shows:

a. Immature: Precision = 0.89, Recall = 0.79, F1-score = 0.84
b. Mature: Precision = 0.74, Recall = 0.89, F1-score = 0.81
c. Overmature: Precision = 0.86, Recall = 0.82, F1-score = 0.84

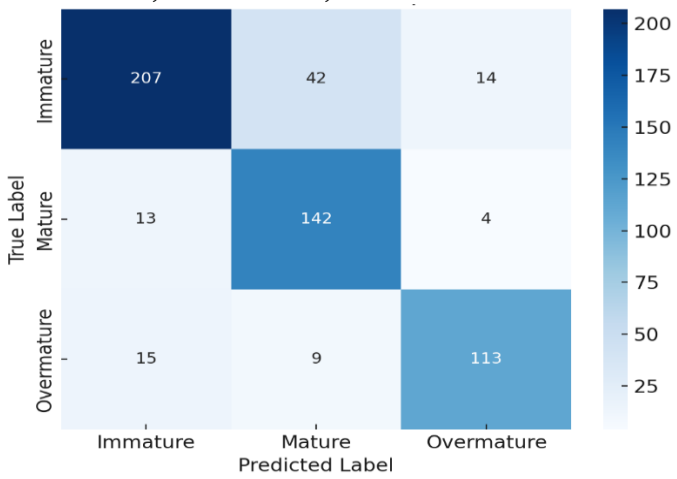

Fig. 9. Confusion matrix of the testing result.

While most misclassifications happen between the immature and mature classes, these findings validate the model's strength in identifying immature and overmature pods. Despite class overlap, the confusion matrix in Figure 9 verifies that 207 of 263 immature pods, 142 of 159 mature pods, and 113 of 137 overmature pods were accurately classified, therefore proving great prediction dependability. Moreover, the precision-recall curves in Figure 10 revealed consistent sensitivity and specificity trade-offs across every class. Suggesting a high degree of class separability, the area under the ROC curve (AUC-ROC) values are 0.83 for immature, 0.80 for mature, and 0.84 for overmature.

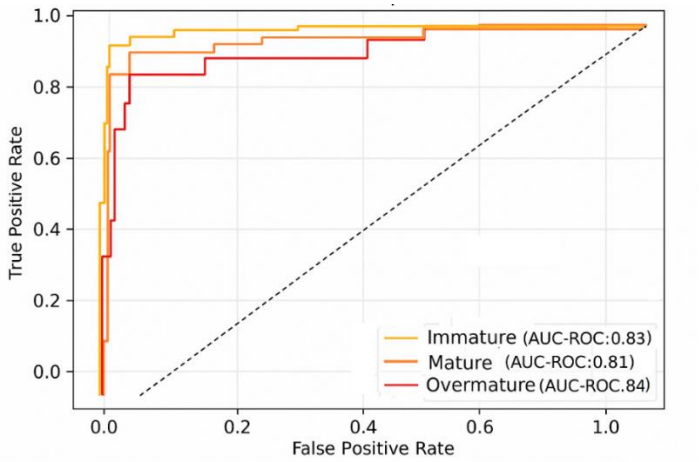

Fig. 10. TPR-FPR curves for each class with AUC scores.

Inference time was measured to assess the model's practical deployment capability. The average inference time per image was 21 milliseconds on an NVIDIA RTX 4070 GPU, demonstrating feasibility for real-time deployment in environments such as smart harvesters or quality control systems. Although training time was intensive, largely due to the complexity of transformer-based architectures and the integration of SPT and LSA, the model compensates with a fast inference time, which is competitive with traditional CNN-based models such as ResNet and EfficientNet in similar agricultural classification tasks. While some lightweight CNNs can achieve inference times below 10 ms, many reported CNN-based approaches operate within the 20–45 ms range depending on hardware and model complexity , while our model achieves a latency of just 21 milliseconds, making it suitable for real-time deployment.

Misclassified examples were predominantly found in the immature–mature and mature–overmature boundaries, where lighting inconsistency and fruit surface shadowing played a role in confusing visual signals. Some fruits displayed multiple matureness indicators on a single pod, leading to ambiguity during labeling and training. These results highlight the persistent challenge of environmental variability in field-collected data. Nevertheless, the use of SPT contributed to improved local texture modeling, while LSA helped focus attention on contextually relevant features. Figure 10 presents several examples of such misclassified images, illustrating the visual ambiguity that challenged the model.

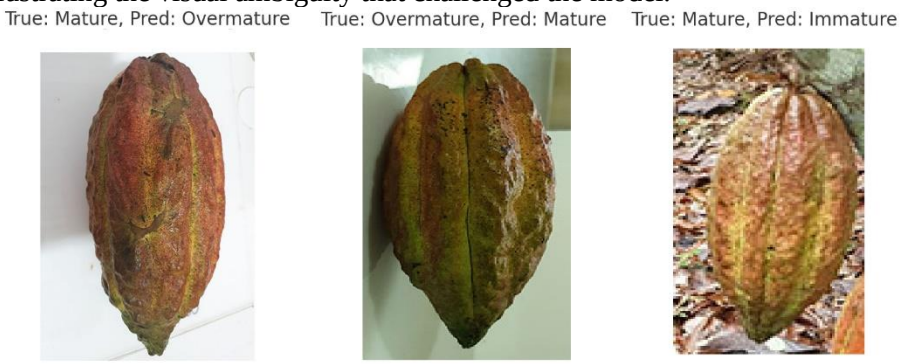

Fig. 11. Examples Of Misclassified Cocoa Fruit Images.

Despite its success, the research faces several limitations. The dataset, while diverse, remains small and may not encompass all variations in cocoa pod appearance. Lighting conditions were not standardized, which may have introduced noise during model training.

Moreover, the class imbalance between mature and non-mature classes could still influence decision boundaries. Addressing these limitations will require the collection of a larger and more balanced dataset, the use of advanced augmentation strategies, or the integration of multimodal data such as hyperspectral or thermal imaging.

We also conducted 10-fold cross validation for ViT and our method to show the improvement provided by using SPT and LSA. The results of 10-fold cross-validation for ViT and the proposed method are shown in Table 2. Our method consistently outperformed ViT, achieving a higher mean accuracy (89.23%) and lower variance (0.24%) compared to ViT (average accuracy 86.5%, variance 0.69%). The reduced variance indicates that the proposed method performs more consistently across different data splits, highlighting its robustness and reliability. This showed the effect of utilizing SPT and LSA on the ViT model performance.

Table 2 - 10-fold Cross validation result.

Fold	ViT Accuracy (%)	Proposed Method Accuracy (%)
Fold 1	85.23	88.41
Fold 2	87.42	89.02
Fold 3	86.87	89.83
Fold 4	88.12	88.74
Fold 5	86.03	89.62
Fold 6	85.94	89.47
Fold 7	87.11	89.25
Fold 8	85.67	89.98
Fold 9	86.21	88.67
Fold 10	86.45	89.33
Mean	86.5	89.23
Variance	0.69	0.24

To further demonstrate the effects of using SPT and LSA, we conducted ablation study. The ablation study evaluated the contributions of SPT and LSA to the system performances. Table 3 summarizes the ablation study results. Adding SPT improved accuracy by 1.48% over the baseline ViT, showing its effectiveness in enriching the training set and improving generalization. Adding LSA improved accuracy by 0.75%, enhancing localized feature extraction. By combining SPT and LSA, we obtained the best performance (82.65%), demonstrating their complementary roles for enhancing the ViT performance in classyfing cocoa fruit ripeness.

Table 3 - Ablation study result.

Model	Accuracy (%)	Macro F1-Score (%)
Model 1 (Baseline ViT)	80.14	80.25
Model 2 (ViT + SPT)	81.62	81.48
Model 3 (ViT + LSA)	80.89	80.93
Model 4 (Proposed Method)	82.65	82.71

We evaluated our proposed method and compared it with other four methods, namely VGG, MobileNet, ResNet and Vision Transformer (ViT). Table 4 summarizes their performance metrics. Our proposed method achieved the highest accuracy (82.65%), outperforming ViT (80.14%) and other models. MobileNet and VGG had the lowest accuracies, indicating limitations in their architectures for this dataset. The proposed method also exhibited the best balance of performances across all classes (82.71%), followed by ViT (80.25%). Furthermore, our method outperformed all other methods across all classes. The most significant improvement was observed for the challenging Mature class, achieving an F1-score of 80.68%, compared to 76.37% for ViT.

Table 4 - Performance comparison.

Method	Accuracy	Macro avg F1-Score	Immature F1-Score	Mature F1-score	Overmature F1-score
VGG	64.76	62.81	72.11	58.48	57.83
MobileNet	72.45	70.68	80.87	64.76	66.41
ResNet	76.39	75.55	81.95	70.17	74.52
ViT	80.14	80.25	81.4	76.37	82.96
Proposed method	82.65	82.71	83.13	80.68	84.33

In summary, our proposed ViT-SPT-LSA model achieved high classification accuracy and robustness, particularly in difficult conditions involving inter-class similarity and lighting variability. Its fast inference time and generalization potential make it well-suited for real-time agricultural deployment and adaptation across related smart farming tasks.

5. Conclusion

This research proposed a cocoa ripeness classification model based on Vision Transformer (ViT), enhanced with Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA). The model achieved a classification accuracy of 82.65% and a macro-average F1-score of 82.71%, outperforming baseline ViT and other CNN-based architectures. These results highlight the effectiveness of combining SPT and LSA in enhancing ViT's generalization on small and variable agricultural datasets. From a theoretical standpoint, the work demonstrates the adaptability of transformer-based architectures for complex agricultural image recognition tasks. Practically, the proposed system offers strong potential for automating cocoa ripeness classification during harvesting or quality control, especially under inconsistent field conditions. Nevertheless, limitations include the relatively small dataset size, susceptibility to lighting variations, and class imbalance, particularly in the mature class. Future work should focus on scaling the dataset, incorporating advanced augmentation or hyperspectral data, and deploying the model on edge computing devices for real-time, in-field applications to fully realize the benefits of smart agriculture technologies.

References

Ala'a, R., & Ibrahim, R. W. (2024). Classification of tomato leaf images for detection of plant disease using conformable polynomials image features. *MethodsX*, *13*, 102844. https://doi.org/10.1016/j.mex.2024.102844

Alimjan, G., Sun, T., Liang, Y., Jumahun, H., & Guan, Y. (2018). A new technique for remote sensing image classification based on combinatorial algorithm of SVM and KNN. *International Journal of Pattern Recognition and Artificial Intelligence*, *32*(2), 1850004. https://doi.org/10.1142/S0218001418590127

Borhani, Y., Khoramdel, J., & Najafi, E. (2022). A deep learning based approach for automated plant disease classification using vision transformer. *Scientific Reports*, *12*(1), 11554. https://doi.org/10.1038/s41598-022-15163-0

Brigato, L., & Iocchi, L. (2021). A close look at deep learning with small data. *2020 25th International Conference on Pattern Recognition (ICPR)*, 2490–2497.

Charco, J. L., Yanza-Montalvan, A., Zumba-Gamboa, J., Alonso-Anguizaca, J., & Basurto-Cruz, E. (2024). ViTSigat: Early Black Sigatoka Detection in Banana Plants Using Vision Transformer. *Conference on Information and Communication Technologies of Ecuador*, 117–130. https://doi.org/10.1007/978-3-031-75431-9_8

Chitta, S., Yandrapalli, V. K., & Sharma, S. (2024). Deep Learning for Precision Agriculture: Evaluating CNNs and Vision Transformers in Rice Disease Classification. *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, 1–6. https://doi.org/10.1109/OTCON60325.2024.10687983

De Silva, M., & Brown, D. (2023). Multispectral plant Disease Detection with Vision transformer–convolutional neural network hybrid approaches. *Sensors*, *23*(20), 8531. https://doi.org/10.3390/s23208531

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv Preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929

El Sakka, M., Mothe, J., & Ivanovici, M. (2024). Images and CNN applications in smart agriculture. *European Journal of Remote Sensing*, *57*(1), 2352386. https://doi.org/10.1080/22797254.2024.2352386

Emmamuel, A., Asim, U., Yu, H., Kim, S., & others. (2022). 3D-CNN method over shifted patch tokenization for MRI-based diagnosis of Alzheimer's disease using segmented hippocampus. *Journal of Multimedia Information System*, *9*(4), 245–252. https://doi.org/10.33851/JMIS.2022.9.4.245

Ergün, E. (2025). High precision banana variety identification using vision transformer based feature extraction and support vector machine. *Scientific Reports*, *15*(1), 10366. https://doi.org/10.1038/s41598-025-95466-0

Eric, O., Gyening, R.-M. O. M., Appiah, O., Takyi, K., & Appiahene, P. (2023). Cocoa beans classification using enhanced image feature extraction techniques and a regularized Artificial Neural Network model. *Engineering Applications of Artificial Intelligence*, *125*, 106736. https://doi.org/10.1016/j.engappai.2023.106736

Essah, R., Anand, D., & Singh, S. (2022). An intelligent cocoa quality testing framework based on deep learning techniques. *Measurement: Sensors*, *24*, 100466. https://doi.org/10.1016/j.measen.2022.100466

Food and Agriculture Organization (FAO). (2023). *Indonesia: Upgrading bulk cocoa into fine cocoa*. https://openknowledge.fao.org/server/api/core/bitstreams/684e2bd3-6b91-48f5-a7cd-4125c5c74cab/content

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 1050–1059. https://dl.acm.org/doi/10.5555/3045390.3045502

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M., & Hu, S.-M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, *8*(3), 331–368. https://link.springer.com/article/10.1007/s41095-022-0271-y

Guo, Q., Qiu, X., Xue, X., & Zhang, Z. (2019). Low-rank and locality constrained self-attention for sequence modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(12), 2213–2222. https://doi.org/10.1109/TASLP.2019.2944078

International Cocoa Organization (ICCO). (2022). *Top 10 cocoa-producers and the issue of child labor in the industry*. https://www.developmentaid.org/news-stream/post/176254/top-10-cocoa-producers

Joshi, B., Bansal, S., & Sharma, C. (2023). Classification of Tomato Leaf Disease using Feature Extraction with KNN Classifier. *2023 Seventh International Conference on Image Information Processing (ICIIP)*, 541–546. https://doi.org/10.1109/ICIIP61524.2023.10537671

Juncai, H., Yaohua, H., Lixia, H., Kangquan, G., & Satake, T. (2015). Classification of ripening stages of bananas based on support vector machine. *International Journal of Agricultural and Biological Engineering*, *8*(6), 99–103. https://doi.org/10.3965/j.ijabe.20150806.1275

Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, *10*, 621. https://doi.org/10.3389/fpls.2019.00621

Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, *11*, 621. https://doi.org/10.3389/fpls.2019.01750

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, *54*(10s), 1–41. https://doi.org/10.1145/3505244

Kharamat, W., Wongsaisuwan, M., & Wattanamongkhol, N. (2020). Durian ripeness classification from the knocking sounds using convolutional neural network. *2020 8th International Electrical Engineering Congress (iEECON)*, 1–4. https://doi.org/10.1109/iEECON48109.2020.229571

Kulkarni, A., Shivananda, A., & Sharma, N. R. (2022). Explainable AI for computer vision. In *Computer Vision Projects with PyTorch: Design and Develop Production-Grade Models* (pp. 325–340). Springer. https://doi.org/10.1007/978-1-4842-8273-1_10

Lee, S. H., Lee, S., & Song, B. C. (2021). Vision transformer for small-size datasets. *arXiv Preprint arXiv:2112.13492*. https://doi.org/10.48550/arXiv.2112.13492

Lin, F., Crawford, S., Guillot, K., Zhang, Y., Chen, Y., Yuan, X., Chen, L., Williams, S., Minvielle, R., Xiao, X., & others. (2023). Mmst-vit: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5774–5784. https://doi.org/10.1109/ICCV51070.2023.00531

Liu, J., & Wang, X. (2021). Plant diseases and pests detection based on deep learning: A review. *Plant Methods*, *17*, 22. https://doi.org/10.1186/s13007-021-00722-9

Lopes, J. F., da Costa, V. G. T., Barbin, D. F., Cruz-Tirado, L. J. P., Baeten, V., & Barbon Junior, S. (2022). Deep computer vision system for cocoa classification. *Multimedia Tools and Applications*, *81*(28), 41059–41077. https://doi.org/10.1007/s11042-022-13097-3

Mishra, A., & Malhotra, M. (2024). A Dual Approach with Grad-CAM and Layer-Wise Relevance Propagation for CNN Models Explainability. *International Conference on Innovation and Emerging Trends in Computing and Information Technologies*, 116–129. https://doi.org/10.1007/978-3-031-80842-5_10

Nahak, P., Pratihar, D. K., & Deb, A. K. (2025). Tomato maturity stage prediction based on vision transformer and deep convolution neural networks. *International Journal of Hybrid Intelligent Systems*, *21*(1), 61–78. https://doi.org/10.3233/HIS-240021

Paneru, B., Paneru, B., & Shah, K. B. (2024). Analysis of Convolutional Neural Network-based Image Classifications: A Multi-Featured Application for Rice Leaf Disease Prediction and Recommendations for Farmers. *arXiv Preprint arXiv:2410.01827*. https://doi.org/10.48550/arXiv.2410.01827

Pothen, Z., & Nuske, S. (2016). Automated assessment and mapping of grape quality through image-based color analysis. *IFAC-PapersOnLine*, *49*(16), 72–78. https://doi.org/10.1016/j.ifacol.2016.10.014

Rad, R. (2024). Vision transformer for multispectral satellite imagery: Advancing landcover classification. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8176–8183. https://doi.org/10.1109/WACV57701.2024.00799

Reedha, R., Dericquebourg, E., Canals, R., & Hafiane, A. (2022). Transformer neural network for weed and crop classification of high resolution UAV images. *Remote Sensing*, *14*(3), 592. https://doi.org/10.3390/rs14030592

Shimazu, R., Leow, C. S., Buayai, P., Makino, K., Mao, X., & Nishizaki, H. (2024). High Quality Color Estimation of Shine Muscat Grape Using Vision Transformer. *2024 International Conference on Cyberworlds (CW)*, 195–202. https://doi.org/10.1109/CW64301.2024.00028

Siregar, B., Pradaning, R., & Hizriadi, A. (2023). Cocoa Ripeness Level Sorting System Using Integrated Computer Vision Technology On Conveyor Belt. *2023 8th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, 1–6. https://doi.org/10.1109/ICEEIE59078.2023.10334634

Suban, I. B., Paramartha, A., Fortwonatus, M., & Santoso, A. J. (2020). Identification the maturity level of carica papaya using the k-nearest neighbor. *Journal of Physics: Conference Series*, *1577*(1), 012028. https://doi.org/10.1088/1742-6596/1577/1/012028

Ulukaya, S., & Deari, S. (2025). A robust vision transformer-based approach for classification of labeled rices in the wild. *Computers and Electronics in Agriculture*, *231*, 109950. https://doi.org/10.1016/j.compag.2025.109950

Yasin, A., & Fatima, R. (2023). On the Image-Based Detection of Tomato and Corn leaves Diseases: An in-depth comparative experiments. *arXiv Preprint arXiv:2312.08659*. https://doi.org/10.48550/arXiv.2312.08659

Yu, C., Wang, J., Chen, Y., & Wu, Z. (2019). Accelerating deep unsupervised domain adaptation with transfer channel pruning. *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2019.8851810

Zhao, J., Berge, T. W., & Geipel, J. (2023). Transformer in UAV Image-Based Weed Mapping. *Remote Sensing*, *15*(21), 5165. https://doi.org/10.3390/rs15215165

Zhou, J., Wang, P., Wang, F., Liu, Q., Li, H., & Jin, R. (2021). Elsa: Enhanced local self-attention for vision transformer. *arXiv Preprint arXiv:2112.12786*. https://doi.org/10.48550/arXiv.2112.12786