Transformer-Based Deep Learning Model for Coffee Bean Classification

Imam Ekowicaksono [1]*, I Wayan Wiprayoga Wisesa [2]*, Vita Fitriani [3]**
* Teknik Informatika, Institut Teknologi Sumatera, Lampung Selatan, Indonesia
** Teknologi Pangan, Institut Teknologi Sumatera, Lampung Selatan, Indonesia
imam.wicaksono@if.itera.ac.id [1], wayan.wisesa@if.itera.ac.id [2], vita.fitriani@tp.itera.ac.id [3]

Article Info

Article history:

Received 2025-07-18
Revised 2025-08-05
Accepted 2025-08-10

Keyword:

Coffee Bean Classification,
Deep Learning,
Transformer,
Vision Transformer,
Swin Transformer.

ABSTRACT

Coffee is one of the most popular beverage commodities consumed worldwide. The process of selecting high-quality coffee beans plays a vital role in ensuring that the resulting coffee has superior taste and aroma. Over the years, various deep learning models based on Convolutional Neural Networks (CNN) have been developed and utilized to classify coffee bean images with impressive accuracy and performance. However, recent advancements in deep learning have introduced novel transformer-based architectures that show great promise for image classification tasks. By incorporating a self-attention module, transformer models excel at generating global context features within images. This ability demonstrate improved and more consistent performance compared to CNN-based models. This study focuses on training and evaluating transformer-based deep learning models specifically for the classification of coffee bean images. Experimental results demonstrate that transformer models, such as the Vision Transformer (ViT) and Swin Transformer, outperform traditional CNN-based models. Swin Transformer model achieves excellent on the coffee bean image classification task, with 95.13% Accuracy and 90.21% F1-Score, while ViT achieves 94.47% Accuracy and 88.93% F1-Score. It indicates their strong capability in accurately identifying and classifying different types of coffee beans. This suggests that transformer-based approaches could be a better alternative for coffee bean image classification tasks in the future.

I. INTRODUCTION

Coffee is one of the most widely consumed plantation commodities worldwide. Indonesia is one of the leading coffee-producing nations worldwide [1]. Coffee consumption has become an integral aspect of the lifestyle of many individuals in Indonesia. The trend of coffee consumption among Indonesians has been on the rise, corresponding to the development of diverse coffee products [2]. Individuals generally consume coffee to enhance their alertness, concentration, and productivity in the workplace [3]. The consumption of coffee, characterized by its superior and distinctive flavor quality, has emerged as a crucial criterion for individuals selecting their preferred coffee products. The flavor quality of coffee is primarily determined by the quality of its beans, specifically green beans. The quality of these coffee beans can be assessed based on their shape, size, color, and the presence of defective beans, known as coffee defects [4]. The quality of coffee beans is typically indicated by a bluish-green hue, standard shape and size, uniformity in bean size, and absence of defects. Various factors can influence the quality of these beans, including the specific coffee variety, environmental conditions of the cultivation area, and methods employed during harvesting and post-harvest processing [5].

The coffee industry is committed to minimizing defects in coffee beans, as these imperfections can lead to undesirable flavors and aromas, commonly known as off-flavors in roasted coffee. Typical defects in coffee beans include broken, damaged, hollow, black, and immature beans [6]. Additionally, coffee beans with irregular shapes, such as the peaberry and longberry, are frequently encountered. According to Yilma and Kufa [7], peaberry coffee beans are characterized by a round, somewhat oval shape with variations in size, with some beans larger and others smaller than standard coffee beans. Additionally, they tend to exhibit a greater density. In contrast, longberry coffee beans are distinguished by their elongated shape and greater length than regular coffee beans.

Peaberry and longberry coffee beans are not categorized as defects based on their flavor parameters, as they possess favorable flavor profiles. Notably, peaberry coffee beans command a higher market value because of their rarity [8]. The presence of coffee beans with unusual sizes and shapes can lead to inconsistencies among the beans. This irregularity in size results in uneven roasting, which subsequently diminishes flavor quality. Therefore, it is essential to classify coffee beans into categories such as normal beans, peaberry, longberry, and defects to ensure their uniformity.

Several studies have been conducted to classify coffee bean images based on their quality [9], [10], [11]. Febriana et al. [12] implemented ResNet-18 and MobileNetV2 models to classify Arabica coffee beans. This study introduced the USK-Coffee Dataset, which includes Arabica Longberry, Peaberry, Premium, and Defective coffee beans. The dataset comprised 8,000 images, with each class (Longberry, Peaberry, Premium, and Defect) containing 2,000 images. Furthermore, the dataset was divided into 4,800 training images, 1,600 validation images, and 1,600 test images. The models were trained on the dataset for 25 epochs, with a batch size of 4. Overall, the study achieved a training accuracy of 98% and test accuracy of 81% using the ResNet-18 model. In contrast, the MobileNetV2 model yielded accuracies of 85% and 81% on the training and test data, respectively. This study provides a public dataset that serves as a benchmark for classifying Arabica coffee beans. A limitation of this study is that the implemented CNN models exhibited overfitting on the training data, resulting in decreased accuracy in the test data.

Santoso et al. [13] employed ResNet-101 for the classification of coffee-bean images. This study leveraged the grayscale-level features of coffee bean images to enhance the classification performance of the model. The dataset used in this study comprised 900 images of Arabica, Liberica, and Robusta coffee beans. Each category, Arabica, Liberica, and Robusta, included 300 distinct RGB images. Subsequently, the preprocessing involved extracting the mean, standard deviation, skewness, energy, entropy, and smoothness values from the pixel value distribution in the grayscale images. The study achieved an accuracy of 99% during the training phase and 100% during the testing phase. Although the dataset remains experimental, it presents opportunities for further exploration using transfer learning strategies.

Jiao et al. [14] developed Swin-HSSAM, a deep learning model for the classification of coffee beans, utilizing transformers. Swin-HSSAM incorporates transformer blocks from the Swin Transformer [15] to extract features and employs a Selective Attention Module (SAM) [16] to enhance the model's capacity to discriminate features from each transformer block. This study utilized a dataset comprising coffee bean images, totaling 10,378 data points, categorized into first-, second-, and third-grade and defective beans. The Swin-HSSAM model demonstrated superior performance compared to classification models such as AlexNet [17], ResNet-50 [18], VGG16 [19], MobileNetV2 [20], and Vision Transformer [21], achieving an average accuracy of 96%.

One of the limitations in the research on coffee bean classification using convolutional neural networks (CNNs) is the misclassification of peaberry beans as premium beans [12]. This issue may stem from the limitations of CNN models in capturing global spatial relationships within an image. In the context of peaberry and premium coffee beans, which share similarities in color and shape, CNN models such as ResNet18 and MobileNetV2 may face challenges in distinguishing between them, as they primarily focus on local features. To address this limitation, future research could explore the attention mechanisms or transformer-based architectures over CNN models. These approaches could enhance the model's ability to capture global contextual information and potentially improve the differentiation between peaberry and premium beans.

Transformer-based deep learning models demonstrate superior performance compared to Convolutional Neural Networks (CNN) [14], [22], [23] for the classification of coffee bean images. Specifically, Vision Transformer (ViT) and Swin Transformer models were employed to classify Arabica coffee bean images within the USK-Coffee dataset [12]. Unlike CNN-based models, which primarily extract local features, ViT and Swin Transformer models can generate global context features in images using a self-attention module. This capability results in models that exhibit enhanced and more robust performances than their CNN-based counterparts.

This study advances the field of coffee bean classification methods, with a particular focus on Arabica coffee beans. By investigating transformer-based classification models, this study offers valuable insights into the comparative analysis of transformer-based and CNN-based models for coffee bean classification tasks. This comparison yielded a comprehensive understanding of the respective strengths and limitations of these models in the context of coffee bean classification. Such analyses aid researchers and industry professionals in making informed decisions when selecting appropriate models for similar tasks. Furthermore, the study's emphasis on Arabica coffee beans enhances our understanding of specific classification challenges and opportunities within this significant coffee variety, potentially leading to improved quality control and grading processes in the coffee industry.

II. Methods

The research was organized into five primary stages: a literature review, the collection and design of the model and dataset, experimentation, evaluation and analysis, and the formulation of conclusions and future directions. These stages are illustrated in Figure 1. The literature review seeks to identify and analyze research opportunities in the domain of deep learning for coffee bean image classification. The stage involving the collection and design of the model and dataset was conducted to facilitate the comparison and development of models using the benchmark dataset. The experimentation phase is crucial for achieving optimal research outcomes, which are subsequently evaluated and analyzed in this study. The final stage encompasses drawing conclusions and analyzing potential avenues for further research.

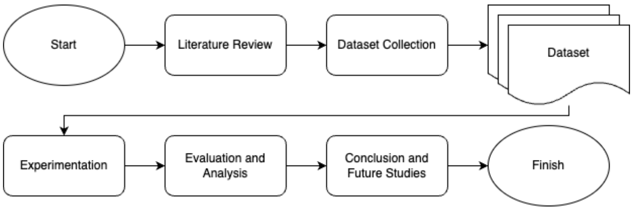


Figure 1. Research Stages

A. *Vision Transformer (ViT)*

The Vision Transformer (ViT) [21] represents a deep learning architecture specifically designed for digital image classification, utilizing the principles of transformers. The architecture of the ViT is shown in Figure 2. This model draws inspiration from the Transformer model [24], which is prominent in the domain of Natural Language Processing (NLP) for tasks such as machine translation. ViT operates by segmenting images into smaller units known as patches, which are then processed as sequences within the transformer encoder, as shown in Figure 3. A pivotal component of the transformer encoder is the Multi-Head Self-Attention (MHSA) module. This module is integral to transformer-based models because it facilitates the extraction of the global context from the data, thereby enhancing the model accuracy.

Suppose $x \in \mathbb{R}^{H\times W\times C}$ is an RGB image with a size of $H \times W$ pixels. Next, x is divided into smaller patches $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$ of size P×P, where N is the number of patches, P is the patch size, and C is the number of channels in the image (for an RGB image, C=3). Each patch, which is in the form of a 2D matrix, is then subjected to a flattening operation that transforms the patch size into a 1D vector (often referred to as the patch embedding). This operation is performed using a linear projection module defined in Equation 1, where $\boldsymbol{z}_0$ is the patch embedding and D is the dimensionality of the latent vector constant.

$$\boldsymbol{z}_0 = \left[x_{class};\ \boldsymbol{x}_p^1\boldsymbol{E};\ \boldsymbol{x}_p^2\boldsymbol{E};\ \cdots\ ;\ \boldsymbol{x}_p^1\boldsymbol{E}\right] + \boldsymbol{E}_{pos} \tag{1}$$

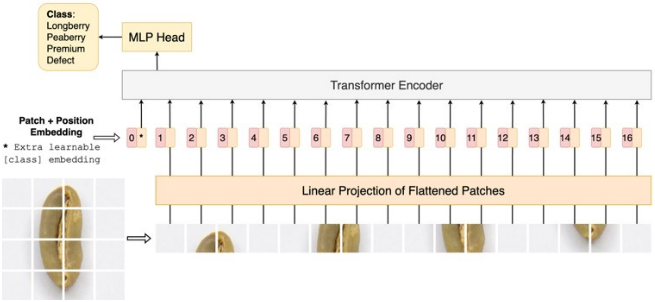

Class:
Longberry
Peaberry
Premium
Defect
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Linear Projection of Flattened Patches

Figure 2. Vision Transformer (ViT)

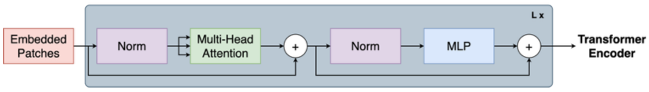


Figure 3. Vision Transformer (ViT) Encoder

The Vision Transformer (ViT) model comprises multiple transformer encoder layers that determine the model scale. Each Transformer encoder layer alternates operations between the Multi-Head Self-Attention (MHSA) module and the Multi-Layer Perceptron (MLP) block. Layer normalization (LN) and residual connections are interposed between the MHSA and MLP, facilitating deep model training. Within the Transformer encoder, the MLP block consists of two layers of linear projection modules that utilize the GELU activation function [25].

$$\boldsymbol{z}'_l = MHSA\big(LN(\boldsymbol{z}_{l-1})\big) + \boldsymbol{z}_{l-1};\ l = 1,\cdots,L \tag{2}$$

$$\boldsymbol{z}_l = MLP\big(LN(\boldsymbol{z}'_l)\big) + \boldsymbol{z}'_l;\ l = 1,\cdots,L \tag{3}$$

$$\boldsymbol{y} = LN(\boldsymbol{z}_l^0) \tag{4}$$

B. *Vision Transformer (ViT)*

Multi-Head Self-Attention (Equation 8) is a combination of several self-attention heads. ViT uses a self-attention module [24] to obtain the global context from the input image in the form of patch embeddings. Each head in the self-attention module receives an input in the form of a sequence $\boldsymbol{z} \in \mathbb{R}^{N\times D}$ and then produces attention scores $SA(\boldsymbol{z})$ generated from the weight matrix A_{ij} with the value vector $\boldsymbol{v}$ (Equation 7). The matrix is obtained from the match between the queries q^i and keys k^j. The query $\boldsymbol{q}$, key $\boldsymbol{k}$, and value $\boldsymbol{v}$ vectors are obtained by applying a linear projection operation to the vector with the matrix (see Equation 5).

$$[\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}] = \boldsymbol{z}\boldsymbol{U}_{qkv} \tag{5}$$

$$A = softmax\left(\frac{qk^T}{\sqrt{D_h}}\right) \tag{6}$$

$$SA(\boldsymbol{z}) = A\boldsymbol{v} \tag{7}$$

$$MHSA(\boldsymbol{z}) = [SA_1(z);\ SA_2(z);\ \cdots;\ SA_k(z)]\boldsymbol{U}_{mhsa} \tag{8}$$

C. *Swin Transformer (Swin)*

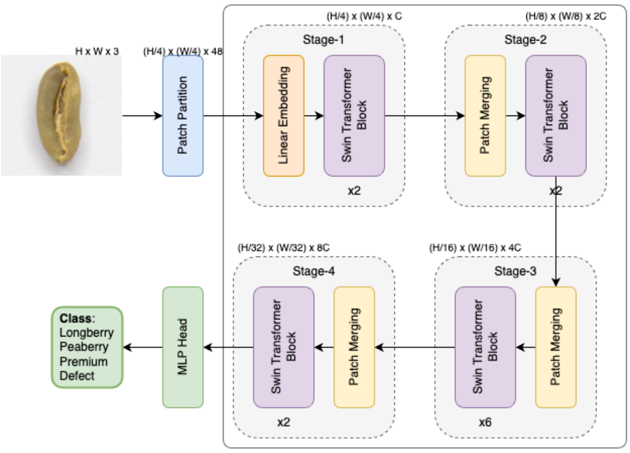


Figure 4. Swin Transformer (Swin)

Swin Transformer (Figure 4), or Swin for short, is a transformer-based deep learning model for classification. Unlike ViT, Swin divides an image into small patches in the form of sequences that are gradually processed into smaller sequences hierarchically as the output of the Swin Transformer Block (Figure 5). Swin Transformer Block making the Swin model more efficient compared to ViT.

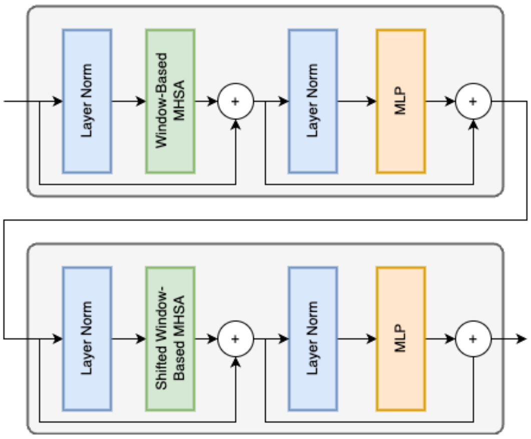


Figure 5. Swin Transformer Block

An integral component of the Swin model is the shifting window operation (Figure 6). This process enhances the model efficiency by eliminating the need to process the entire image simultaneously. Furthermore, the shifting window scheme in Swin is executed by altering positions, thereby facilitating the connection of adjacent image segments. Consequently, the global information inherent in the image was effectively captured. Through this shifting window concept, the Swin model acquires information regarding the relationships between both proximate and more distant image segments.

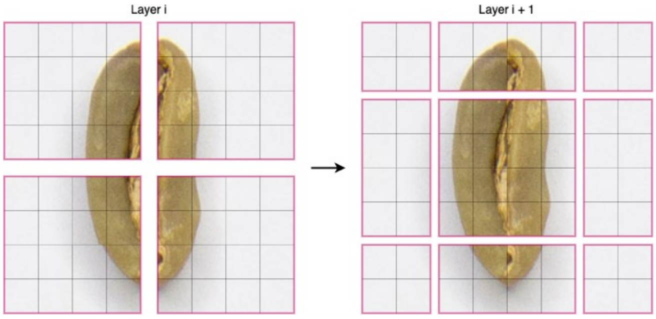


Figure 6. Illustration of shifted windows scheme

The Swin Transformer offers significant advantages in terms of speed and efficiency, particularly when processing large images. This model adeptly comprehends both local (small-scale) and global (large-scale) information from images through its window-and shifting window mechanisms. Consequently, it demonstrates high accuracy and is frequently employed in various image-processing tasks, including classification and segmentation. Furthermore, the model exhibits flexibility, as it can process images of varying sizes without necessitating complex special configuration.

D. *Model Evaluation*

The model performance was evaluated using a confusion matrix. The confusion matrix (Table I) comprises True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. TP denotes the model's accurate prediction of the actual positive class values, whereas TN signifies the model's accuracy in predicting the actual negative class values. FP reflects the model's error in predicting negative classes as positive, whereas FN indicates the model's error in predicting positive classes as negative.

The TP, TN, FP, and FN values were subsequently employed to calculate precision (Equation 9), recall (Equation 10), F1-score (Equation 11), and accuracy (Equation 12). Precision evaluates the model's accuracy in predicting positive classes out of all predictions made, whereas recall (sensitivity) assesses the model's effectiveness in predicting positive classes out of all actual positive classes. Accuracy represents the model's prediction rate across the entire dataset, and the F1-score denotes the harmonic mean between precision and recall [13].

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

$$F_1score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{11}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

TABLE I
CONFUSION MATRIX AS EVALUATION METRICS

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

III. RESULT AND DISCUSSION

A. *Experimental Setup*

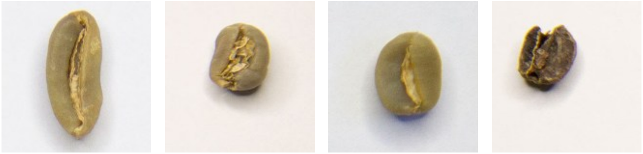
The experiments and model evaluations were performed fairly using the PyTorch framework on an NVIDIA GeForce RTX 4080 Super. The USK-Coffee dataset [12] was used to train the ViT and Swin models. The ViT and Swin models were trained through a fine-tuning scheme for 20 epochs, employing a learning rate of 1e-5, a batch size of 32, and both Adam and SGD [26] optimizers. The model with the lowest validation loss was selected for evaluation of the test data.

B. *Datasets*

This study used the USK-Coffee Dataset as benchmark data for the training and evaluation processes of all models, Swin and ViT. The USK-Coffee Dataset consists of 8,000 images of Arabica coffee beans belonging to the Longberry, Peaberry, Premium, and Defect classes. Table II describes the dataset distribution for each class in the training, validation, and test data. Each class consisted of 2,000 RGB color images, with 1,200 images for the training data, 400 images for the validation data, and 400 images for the test data. Figure 7 shows example images of each class.

TABLE II
NUMBER OF TRAINING, VALIDATION, AND TEST DATA WITHIN THE USK-COFFEE DATASET [12]

No.	Class	Train	Validation	Test	Total
1.	Longberry	1200	400	400	2000
2.	Peaberry	1200	400	400	2000
3.	Premium	1200	400	400	2000
4.	Defect	1200	400	400	2000
	Total	4800	1600	1600	8000



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

Figure 7. Illustration of the USK-Coffee Dataset: (a) Longberry, (b) Peaberry, (c) Premium, and (d) Defect

The training process adopts a comprehensive strategy for data augmentation, employing randaug [27] as the principal method to apply randomly selected image transformations. This approach enhances dataset diversity and improves model generalization. Subsequent to randaug, images are subjected to center cropping to a uniform size of 224 × 224 pixels, ensuring consistency and compatibility with prevalent convolutional neural network architectures. Horizontal flipping is incorporated as an additional augmentation technique, effectively doubling the dataset size and introducing further variability. Finally, each image is normalized using the ImageNet standard, with mean values of [0.485, 0.456, 0.406] and standard deviation values of [0.229, 0.224, 0.225] for the RGB channels, thereby scaling pixel values to a consistent range. This combination of augmentation and preprocessing techniques aims to create a robust and diverse dataset that enhances the model's learning and generalization capabilities.

C. *Result*

Table III presents a comparative analysis of the complexities associated with the ViT and Swin models. The metrics employed for assessing complexity include the number of parameters utilized in the model, model size in megabytes (MB), floating-point operations per second (FLOPs), and multiply accumulate operations (MACs). The Vision Transformer (ViT) necessitates approximately 17.56 GFLOPs and 16.85 GMACs for a single inference on a standard input, indicating the substantial number of floating-point and multiply accumulate operations involved. This significant computational requirement renders ViT relatively resource-intensive, particularly when applied to high-resolution images or devices with power constraints. Conversely, the Swin Transformer exhibits linear complexity concerning image size, thereby enhancing its efficiency compared to ViT, with approximately 15.43 GFLOPs and 15.44 GMACs per inference.

TABLE III
COMPUTATIONAL COMPLEXITY COMPARISON

Model	Params	Model Size (MB)	GFLOPs	GMACs
ViT	85,801,732	327.30	17.56	16.85
Swin	86,747,324	331.35	15.43	15.44
ResNet18[12]	11,689,512	44.7	1.81	NA
MobileNetV2[12]	3,504,872	13.6	0.30	NA

ViT demonstrates remarkable accuracy and effectively captures global features, it demands high computational resources and a large volume of training data. Swin achieves a commendable balance by attaining high accuracy with enhanced efficiency. ResNet18 has compact model size; however, it encounters challenges in identifying subtle differences in coffee beans with similar shapes and colors. MobileNetV2 is highly efficient, but its performance is insufficient that demands meticulous attention to detail.

The efficacy of deep learning models for image classification utilizing transformers, specifically ViT and Swin, is elucidated in the experimental results. Both models were trained using the USK-Coffee dataset. Initially, the models are trained with the training data, and the loss values during training are projected. Subsequently, the validation loss was computed, and the accuracy value for each epoch, based on the validation data, was projected. The model with the highest validation accuracy was evaluated using the test data, and the metrics for TP, TN, FP, FN, Precision, Recall, F1-score, and Accuracy were projected. The experimental results were then analyzed and compared with those of state-of-the-art research [12].

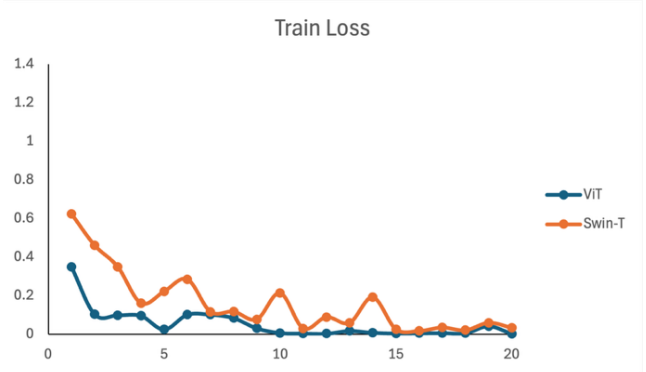


Figure 8. Train phase loss history chart

Figure 8 illustrates a comparative analysis of the performance of the Vision Transformer (ViT) and Swin

Transformer models in classifying images of Arabica coffee beans using the USK-Coffee training dataset. During the initial five epochs, the Swin model exhibits higher loss values relative to the ViT model, suggesting that the Swin model necessitates a greater number of iterations to effectively discern patterns within the dataset. As the number of epochs increased, the training loss performance of the Swin model demonstrated an improved decline, albeit with persistent fluctuations in the loss values. Conversely, the ViT model exhibited a more stable and consistent loss performance, indicating its superior capability to learn the patterns and features of the data.

The observed fluctuations in the loss values during the initial five epochs of the Swin model training phase suggest that the training process requires further adaptation. This variability may be attributed to the complexity of the data or the requirement for adjustments in hyperparameter settings. Nevertheless, as the number of epochs increased, the loss values during the training phase of both the Swin and ViT models tended to stabilize and converge, thereby enhancing the generalization capabilities of the models.

The second stage entails computing the loss of the validation dataset. Figure 9 illustrates the loss history graph for the validation phase of the Swin and ViT models, respectively. Both models exhibited rapid convergence of the validation loss values. Initially, the Swin model yielded a notably high validation loss value, indicating that, during the early phase of training, Swin encountered challenges in classifying the validation data. However, as the number of epochs increased, Swin adapted to the validation data and demonstrated improved classification performance. In contrast, the validation loss of the ViT model tended to remain stagnant at a higher level. Overall, Swin exhibits superior performance compared to ViT as it consistently reduces the validation loss throughout the training process.

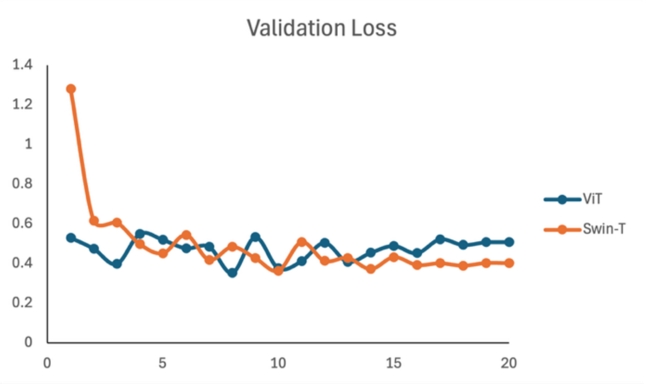


Figure 9. Validation phase loss history chart

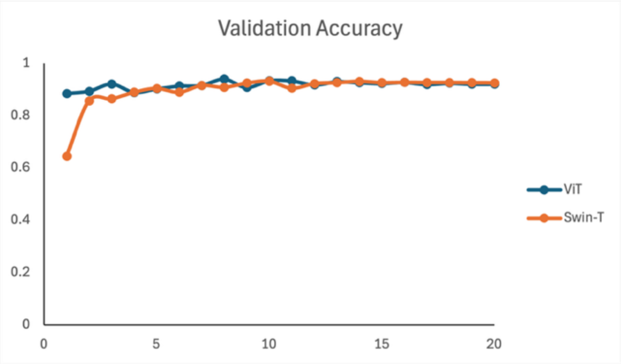


Figure 10. Validation accuracy history chart

Subsequently, the accuracy metrics during the validation phase were computed for each model. Figure 10 illustrates the accuracies of the ViT and Swin models on the validation dataset. Overall, the ViT model exhibited consistently high accuracy throughout the training process. In contrast, the Swin model demonstrated a marked improvement in accuracy from the onset of training. Notably, the Swin model’s validation accuracy surpassed that of the ViT model from epoch 15 onwards, maintaining stability thereafter. Both models achieved accuracy levels exceeding 90%, indicating their efficacy in classifying Arabica coffee bean images.

TABLE IV
CONFUSION MATRIX OF VIT ON USK-COFFEE DATASET

Class	Peaberry	Premium	Longberry	Defect
Peaberry	380	15	2	3
Premium	27	362	3	8
Longberry	22	13	363	2
Defect	29	36	17	318

TABLE V
CONFUSION MATRIX OF SWIN ON USK-COFFEE DATASET

Class	Peaberry	Premium	Longberry	Defect
Peaberry	374	21	0	5
Premium	10	380	3	7
Longberry	13	18	366	3
Defect	18	32	26	324

Tables IV and V display the outcomes of image classification for Arabica coffee beans using the USK-Coffee dataset. The values on the main diagonal denote the class values accurately predicted by the model. Conversely, the values in the upper and lower triangles represent those that were incorrectly predicted by the model. Overall, the diagonal values in the confusion matrix of the Swin model were higher than those in the ViT model, suggesting that the Swin model demonstrated superior classification accuracy compared to the ViT model.

In ViT model, the most prevalent error occurred within the Defect class, which was frequently misidentified as Peaberry (29), Premium (36), and Longberry (17). This indicates the model's challenge in accurately discerning the unique characteristics of this class. Furthermore, there were significant errors with Peaberry being predicted as Premium

(15) and Premium being misclassified as Peaberry (27). In Swin model, Defect class demonstrate the highest rate of misclassification, frequently being classified as Premium (32) and Longberry (26), indicating that these categories possess similar characteristics. Peaberry is still often misidentified as Premium (21), while Longberry is commonly misclassified as Peaberry (13) and Premium (18). These inaccuracies underscore that the model's representation of visual features, particularly shape and color, remains insufficiently distinct among the classes.

In Table VI, the Swin model demonstrates superior performance compared to CNN-based model [12], ResNet18, MobileNetV2 and ViT. Swin achieved a precision of 90.65%, recall of 90.25%, and F1-score of 90.21%. This suggests that the Swin model is the most effective in classifying Arabica coffee bean images, exhibiting balanced Precision and Recall values.

TABLE VI
MODEL PERFORMANCE COMPARISON ON USK-COFFEE DATASET

Model	Precision	Recall	F1-score
ViT	89.58%	88.94%	88.93%
Swin	**90.65%**	**90.25%**	**90.21%**
ResNet18 [12]	84.14%	81.12%	82.60%
MobileNetV2 [12]	81.42%	81.31%	81.36%

The model's high prediction accuracy (precision) and capability to identify positive classes (recall) can be attributed to its hierarchical feature extraction and window-based attention mechanisms. These features enhance the model's ability to effectively capture spatial and contextual information in images.

TABLE VII
ACCURACY OF EACH CLASS ON USK-COFFEE DATASET

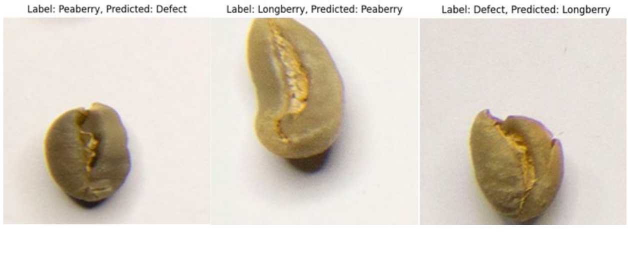
Class	ViT	Swin	ResNet18 [12]	MobileNet V2[12]
Peaberry	93.88%	**95.81%**	69.50%	81.50%
Premium	93.63%	**94.31%**	93.25%	76.25%
Longberry	**96.31%**	96.06%	80.50%	86.75%
Defect	94.06%	**94.31%**	81.25%	80.75%
Mean Acc	94.47%	95.13%	81.13%	81.31%

Transformer-based models, specifically ViT and Swin, generally achieve superior and more consistent accuracy across all coffee classes than CNN-based models, such as ResNet18 and MobileNetV2. Table VII presents a detailed comparison of the model accuracies for each class. The Swin model exhibits the highest accuracy in the Peaberry class (95.81%) and the Premium class (94.31%), alongside highly competitive performance in the Longberry (96.06%) and Defect (94.31%) classes. Similarly, the ViT model demonstrated exceptional performance, achieving the highest accuracy in the Longberry class (96.31%) and accuracy levels comparable to Swin in the other classes, with 93.88% in Peaberry, 93.63% in Premium, and 94.06% in Defects. Both transformer models consistently surpassed the other models across all classes, maintaining accuracies above 93%, which underscores their proficiency in capturing the global and complex features inherent in coffee bean images.

IV. DISCUSSION

This study investigated the application of transformer-based deep learning models for classifying Arabica coffee bean images. The primary models employed in this study are the Vision Transformer (ViT) and Swin Transformer, whose performances are compared with convolutional neural network (CNN) models such as ResNet18 and MobileNetV2. The dataset used was USK-Coffee, which comprised 8,000 images of coffee beans categorized into four groups: Longberry, Peaberry, Premium, and Defect. Both the ViT and Swin models were trained for 20 epochs using a fine-tuning approach with adjusted parameters. The primary objective of this study was to assess the efficacy of transformer models in accurately classifying coffee bean images.

The training outcomes indicated that the ViT maintained a stable performance throughout the training phase. However, it exhibited suboptimal results during validation. In contrast, the Swin Transformer initially experienced fluctuations at the onset of training but demonstrated a marked improvement as the number of epochs increased. Analysis of the validation loss and accuracy graphs revealed that Swin consistently enhanced its performance, eventually surpassing ViT from the 15th epoch onward. Both models achieved validation accuracies exceeding 90%, underscoring the efficacy of the transformer approach in processing the complex visual features of coffee bean images. The superior performance of Swin is attributed to its hierarchical architecture and shifted window mechanism, which enhance spatial information extraction.


Label: Peaberry, Predicted: Defect
Label: Longberry, Predicted: Peaberry
Label: Defect, Predicted: Longberry

(a) (b) (c)

Figure 11. *Example of model prediction errors with class labels.*

Figure 11 presents several instances of prediction errors encountered by the model in classifying coffee bean types. The initial error, depicted in Figure 11(a), involved the model erroneously identifying a peaberry as a defect. This misclassification is likely attributable to the non-spherical shape of the peaberry, which led the model to interpret it as a defective bean. The subsequent error, illustrated in Figure 11(b), involved the model incorrectly categorizing a longberry as a peaberry. Finally, Figure 11(c) demonstrates a scenario in which the model mistakenly recognizes a defective bean as a longberry.

The identified errors highlight the complexities of developing an accurate model for classifying coffee beans.

The diverse shapes and visual characteristics unique to each type of coffee bean can create ambiguity in the classification process. For example, a peaberry that is not perfectly round or a defect resembling a longberry might confuse the model. This underscores the need to augment the training dataset with a wide variety of bean shapes and conditions and possibly incorporate additional features beyond visual appearance to improve the model's accuracy in distinguishing different types of coffee beans.

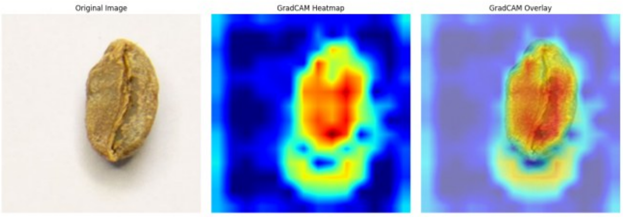


Figure 12. ViT *Grad-CAM visualization of defect bean sample.*

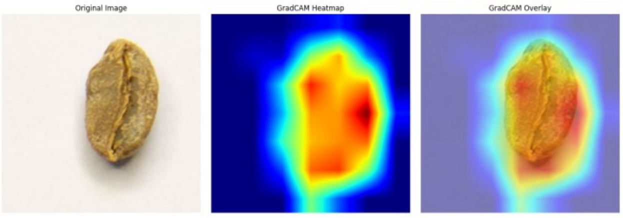


Figure 13. Swin *Grad-CAM visualization of defect bean sample.*

Figures 12 and 13 show how the Vision Transformer (ViT) and Swin Transformer models focus on coffee beans with defects using Grad-CAM[28]. In ViT, the heatmap mainly focuses on the center of the bean, spreading out but remaining within the outline of the bean. This means that ViT captures important features, but its focus is slightly spread out and not very precise. This occurs because of the ViT's design, which uses patch embedding and global self-attention, making it good at seeing the overall picture but not fine details. ViT provides a broad view of important areas, but the edges are not clear.

In contrast, the Swin Transformer (Figure 13) showed a more precise focus, closely matching the shape of the coffee bean. It highlights areas with defects, such as cracks or rough textures. This shows that Swin, with its hierarchical structure and shifted window mechanism, is good at capturing local details while maintaining the overall context. Swin helps users or researchers to determine which parts of an image affect the model's decisions. This is a significant advantage of Swin over ViT for classifying small objects with complex structures, such as coffee beans.

However, both models have their limitations. ViT provides a strong global view, but its focus is less precise, making detailed analysis difficult, especially when defects resemble the background or other beans. Swin focuses more accurately but is more complex and requires proper window settings to avoid losing important information. For coffee beans with many visual differences, Swin's mix of local and global features is helpful, but its results still require manual checking across samples. Therefore, the choice of model depends on whether precise spatial details (Swin) or a general global view (ViT) is required.

V. Conclusion

This study conducted a comparative analysis of transformer-and convolutional neural network (CNN)-based models for the classification of Arabica coffee beans. The transformer-based model, specifically the Swin Transformer, demonstrated superior performance compared to the CNN-based model across various performance metrics. The Swin Transformer achieved a precision of 90.56%, 90.25%, an F1-score of 90.21%, and an average accuracy of 95.13%, respectively. These findings underscore the potential of transformer architectures for image classification tasks, particularly in the realm of agricultural product classification. The enhanced performance of the Swin Transformer suggests that its capability to capture both local and global features in images is particularly advantageous for distinguishing between different types and qualities of Arabica coffee beans. This study contributes to the expanding body of evidence supporting the efficacy of transformer-based models in computer vision tasks, with a specific emphasis on the classification of coffee beans. However, transformer-based image classification models are constrained by their reliance on extensive, high-quality training datasets. In the absence of a dataset that is both sufficiently large and representative, these models may fail to yield optimal results and are prone to overfitting.

Future research on the classification of coffee beans could benefit from incorporating a wider range of coffee bean types, including Arabica, Robusta, and Liberica. This approach enhances the generalizability of the classification model, allowing for more accurate identification and categorization of various coffee bean types across diverse samples. By integrating multiple coffee bean varieties into the study, researchers could gain valuable insights into the unique characteristics of each type of coffee bean. This expanded dataset would enable the identification of specific visual, textural, or chemical markers that are distinctive to Arabica, Robusta, and Liberica beans. Such a comprehensive analysis could not only improve the accuracy of automated classification systems but also deepen the understanding of coffee bean attributes, potentially benefiting various sectors of the coffee industry, from quality control in production to consumer education and product innovation.

This research can be extended within the Internet of Things (IoT) domain to develop edge devices capable of classifying or identifying defective coffee beans. Such a device could be specifically designed for direct application by coffee farmers, thereby assisting them in improving the quality of their coffee production. By utilizing this device, farmers would be able to efficiently distinguish and segregate premium coffee beans from defective ones. This advancement would not only streamline the production process but also enhance the market value of the coffee produced by the farmers.

REFERENCES

[1] Y. Yulianti, N. Andarwulan, D. R. Adawiyah, D. Herawati, and D. Indrasti, 'Physicochemical characteristics and bioactive compound profiles of Arabica Kalosi Enrekang with different postharvest processing', *Food Sci. Technol*, vol. 42, 2022, doi: 10.1590/fst.67622.

[2] D. Herawati *et al.*, 'Impact of bean origin and brewing methods on bioactive compounds, bioactivities, nutrition, and sensory perception in coffee brews: An Indonesian coffee gastronomy study', *International Journal of Gastronomy and Food Science*, vol. 35, p. 100892, Mar. 2024, doi: 10.1016/j.ijgfs.2024.100892.

[3] R. M. Van Dam, F. B. Hu, and W. C. Willett, 'Coffee, Caffeine, and Health', *N Engl J Med*, vol. 383, no. 4, pp. 369–378, Jul. 2020, doi: 10.1056/NEJMra1816604.

[4] W. B. Sunarharum, S. S. Yuwono, N. B. S. W. Pangestu, and H. Nadhiroh, 'Physical and sensory quality of Java Arabica green coffee beans', *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 131, p. 012018, Mar. 2018, doi: 10.1088/1755-1315/131/1/012018.

[5] R. A. Fadri, K. Kesuma Sayuti, N. Nazir, and I. Suliansyah, 'Evaluation of the Value of the Defective and Taste of Arabica Coffee (Coffea Arabica L) West Sumatera', *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 819, no. 1, p. 012004, Jul. 2021, doi: 10.1088/1755-1315/819/1/012004.

[6] E. N. Koeffer, *Progress in Food Chemistry*. New York: Nova Science Publishers, Incorporated, 2008.

[7] A. Yilma and T. Kufa, 'Coffee Peaberry as A Potential Seed Source for Production', *IJRSSET*, vol. 7, no. 9, pp. 30–35, 2020.

[8] H. L. Gope and H. Fukai, 'Peaberry and normal coffee bean classification using CNN, SVM, and KNN: Their implementation in and the limitations of Raspberry Pi 3', *AIMSAGRI*, vol. 7, no. 1, pp. 149–167, 2022, doi: 10.3934/agrfood.2022010.

[9] C. C. Enriquez, J. Marcelo, D. R. Verula, and N. J. Casildo, 'Leveraging deep learning for coffee bean grading: A comparative analysis of convolutional neural network models'.

[10] T. A. Heryanto and I. G. B. B. Nugraha, 'Classification of Coffee Beans Defect Using Mask Region-based Convolutional Neural Network', in *2022 International Conference on Information Technology Systems and Innovation (ICITSI)*, Bandung, Indonesia: IEEE, Nov. 2022, pp. 333–339. doi: 10.1109/ICITSI56531.2022.9970890.

[11] V. A. M. Luis, M. V. T. Quinones, and A. N. Yumang, 'Classification of Defects in Robusta Green Coffee Beans Using YOLO', in *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, Kota Kinabalu, Malaysia: IEEE, Sep. 2022, pp. 1–6. doi: 10.1109/IICAIET55139.2022.9936831.

[12] A. Febriana, K. Muchtar, R. Dawood, and C.-Y. Lin, 'USK-COFFEE Dataset: A Multi-Class Green Arabica Coffee Bean Dataset for Deep Learning', in *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, Malang, Indonesia: IEEE, Jun. 2022, pp. 469–473. doi: 10.1109/CyberneticsCom55287.2022.9865489.

[13] B. R. Santoso, C. A. Sari, and E. H. Rachmawanto, 'Coffee Beans Classification Using Convolutional Neural Networks Based On Extraction Value Analysis In Grayscale Color Space', *JAIC*, vol. 9, no. 1, pp. 31–37, Jan. 2025, doi: 10.30871/jaic.v9i1.8916.

[14] Y. Jiao *et al.*, 'Swin-HSSAM: A green coffee bean grading method by Swin transformer', *PLoS One*, vol. 20, no. 5, p. e0322198, May 2025, doi: 10.1371/journal.pone.0322198.

[15] Z. Liu *et al.*, 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows', *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9992–10002, 2021, doi: 10.1109/ICCV48922.2021.00986.

[16] F. Chollet, 'Xception: Deep learning with depthwise separable convolutions', *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1800–1807, 2017, doi: 10.1109/CVPR.2017.195.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet Classification with Deep Convolutional Neural Networks', in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012.

[18] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[19] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, 'MobileNetV2: Inverted Residuals and Linear Bottlenecks', Mar. 21, 2019, *arXiv*: arXiv:1801.04381. doi: 10.48550/arXiv.1801.04381.

[21] A. Dosovitskiy *et al.*, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', *International Conference on Learning Representations*, Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[22] R. Karthik, R. Aswin, K. S. Geetha, and K. Suganthi, 'An Explainable Deep Learning Network With Transformer and Custom CNN for Bean Leaf Disease Classification', *IEEE Access*, vol. 13, pp. 38562–38573, 2025, doi: 10.1109/ACCESS.2025.3546017.

[23] R. Selvanarayanan, S. R, G. T, and K. L, 'Hybrid Vision Transformer and CNN for Detection of Overripe Coffee Berry Disease (OCBD) in Coffee Plantation', in *2024 International Conference on Emerging Research in Computational Science (ICERCS)*, Coimbatore, India: IEEE, Dec. 2024, pp. 1–7. doi: 10.1109/ICERCS63125.2024.10895612.

[24] A. Vaswani *et al.*, 'Attention Is All You Need', 2017, *arXiv*: arXiv:1706.03762. Accessed: Oct. 29, 2024. [Online]. Available: http://arxiv.org/abs/1706.03762

[25] D. Hendrycks and K. Gimpel, 'Gaussian Error Linear Units (GELUs)', Jun. 06, 2023, *arXiv*: arXiv:1606.08415. doi: 10.48550/arXiv.1606.08415.

[26] I. Loshchilov and F. Hutter, 'SGDR: Stochastic Gradient Descent with Warm Restarts', May 03, 2017, *arXiv*: arXiv:1608.03983. doi: 10.48550/arXiv.1608.03983.

[27] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, 'RandAugment: Practical automated data augmentation with a reduced search space', Nov. 13, 2019, *arXiv*: arXiv:1909.13719. Accessed: Aug. 30, 2024. [Online]. Available: http://arxiv.org/abs/1909.13719

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization', *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.