International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I6P128

Volume 72 Issue 6, 304-314, June 2024

Original Article

Image Classification of Green Arabica Coffee Using Transformer-Based Architecture

Maharani Nurul Izza[1], Gede Putra Kusuma[2]

[1,2]*BINUS Graduate Program - Master of Computer Science, Bina Nusantara University*
Jakarta, Indonesia.

[1]*Corresponding Author : maharani.izza@binus.ac.id*

Received: 25 March 2024 Revised: 17 May 2024 Accepted: 06 June 2024 Published: 29 June 2024

Abstract *- Indonesia is the second-largest producer of coffee, yet many still rely on manual classification methods for bean sorting. This manual approach is prone to human error and can lead to significant economic disadvantages. While numerous studies have aimed to classify coffee beans, none have utilized a compact dataset of Indonesian Arabica coffee beans employing a transformer-based architecture solution. Therefore, this paper proposes an evaluation and optimization of a deep learning model for classifying green Arabica coffee using transformer-based architectures. The transformer models utilized include Vision Transformer (ViT), Data Efficient in Transformer (DeiT), and Swin Transformer. Our research demonstrates the highest accuracy on the test data, with 84.75% using Swin Transformer, 82.25% using ViT, and 81.12% using DeiT. These accuracies represent improvements over previous baseline studies.*

Keywords *- Deep learning, Data efficient in transformer, Swin transformer, Transformer model, Vision transformer.*

1. Introduction

Coffee is a highly promising sector in Indonesia, with plantation areas spanning over 1.25 hectares. According to the International Coffee Organization (2020), Indonesia ranks second in coffee production in Asia and the Pacific, following Vietnam [1]. However, a recent survey conducted at KNT Coffee in Banda Aceh, one of Indonesia's largest coffee plantations, revealed that the manual separation process of coffee beans by human workers can lead to errors and lengthy processing times [2]. Manual coffee bean classification not only impedes domestic and international trade processes but also poses challenges for workers, such as fatigue and stress, potentially impacting the quality of coffee beans. Therefore, automating the image classification of coffee beans is imperative, especially in Indonesia.

Indonesia boasts several varieties of coffee beans, including Arabica coffee, which is categorized into four classes: Peaberry, Longberry, Premium, and Defect. Previous studies have utilized a variety of models to classify coffee beans, including Convolutional Neural Networks (CNN), k-Nearest Neighbors (KNN), ResNet-152, VGG16, Artificial Neural Networks (ANN), Support Vector Machines (SVM), MobileNetV2, and ResNet-18 [2]–[7]. However, the highest accuracy achieved in multiclass coffee bean image classification remains at 81.31%. With advancements in deep learning methods for image classification, there is potential to improve accuracy using alternative methods.

The latest breakthrough in computer vision is the development of transformer-based models. The Vision Transformer (ViT), first proposed by Dosovitskiy et al. in 2021, utilizes a self-attention mechanism to understand relationships between different parts of an image by assigning importance values to patches and focusing on relevant information [8]. Vision Transformer offers a competitive advantage due to its ability to train on large datasets with fewer computational resources. In the research conducted by Dosovitskiy et al., ViT achieved an accuracy of 88.55% on the ImageNet dataset, 90.72% on ImageNet-ReaL, 94.55% on CIFAR-100, and 77.63% on VTAB. This study concluded that ViT is capable of delivering excellent performance, even rivaling CNN architectures on several datasets (ImageNet, CIFAR-100, VTAB, etc.)

Another notable transformer model is the Data-efficient Image Transformer (DeiT). In the research conducted by Touvron et al., DeiT achieved competitive results compared to ViT while using fewer images. In this study, DeiT was tested using the ImageNet-1K dataset and achieved an accuracy of 84.4% [9], surpassing the accuracy of ViT, which reached 77.9% on the same dataset. In 2021, Liu et al. proposed the swin transformer model, a modification of the vision transformer. This model employs a hierarchical design that enables more efficient and scalable image processing compared to previous transformer-based models [10]. The innovation in the Swin transformer lies in its introduction of computational efficiency and performance improvement through the shifted window concept, which enhances the model's ability to capture spatial relationships within images. In their research, Liu et al. tested the swin transformer using the ImageNet-1K dataset, achieving an accuracy of 87.3%, the highest reported accuracy for this dataset.

However, transformers have not yet been applied to coffee bean classification in previous research. The application of transformer models in image classification has recently garnered significant attention, with models like Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), and Swin Transformer at the forefront. Despite their promising results, there remains a gap in the literature regarding their application to specific domains, such as the classification of coffee bean images.

This research aims to address this gap by evaluating the performance of these advanced transformer models on the USK-Coffee dataset, a task that has not been sufficiently explored in previous studies. Additionally, Bayesian optimization is utilized for hyperparameter tuning, which is expected to be more efficient and yield models with higher accuracy compared to traditional methods like random search and grid search. Bayesian optimization provides an effective methodology for optimizing expensive black-box functions [11].

The models are further optimized using on-the-fly augmentation techniques, including random horizontal and vertical flips, random contrast, and random rotation. This research is thus expected to identify the best methods for employing transformer-based architectures to classify coffee beans and achieve optimal performance. This paper is structured as follows: Section 2 presents a review of previous studies, Section 3 outlines the methodology, Section 4 presents the results; finally, Section 5 summarizes the analysis, concludes, and provides recommendations for future research.

2. Related Works

Several models have been utilized in previous research for the classification of beans and fruits, including CNN, KNN, ResNet-152, VGG16, ANN, SVM, MobileNetV2, and ResNet-18. This section provides an overview of various previous studies that serve as references for this research. Arboleda et al. employed an Artificial Neural Network (ANN) model to classify 255 images of coffee beans into three classes, achieving an accuracy of 96.66% [12]. Similarly, Amanina and Saraswati researched coffee bean classification using the k-Nearest Neighbors (KNN) algorithm, achieving an 80% accuracy on 300 coffee beans divided into six classes [4]. Jumarlis et al. utilized KNN to classify 68 coffee beans based on their quality level, achieving a 90% accuracy rate [13]. Adiwijaya et al. also employed KNN, achieving an 83% accuracy using 90 coffee beans [14].

In 2020, Adhitya et al. classified seven classes of cacao beans using Convolutional Neural Networks (CNN) and the Gray Level Co-Occurrence Matrix (GLCM) for feature extraction, employing Support Vector Machine (SVM) and XGBoost as classifiers through the Waikato Environment for Knowledge Analysis (WEKA) software. The results favored CNN over GLCM, with CNN achieving a 65.08% accuracy using SVM as the classifier [3].

Gope and Fukai (2020) utilized a CNN model to classify binary coffee beans, comparing its performance with SVM and employing Principal Component Analysis (PCA) for feature extraction. The research showed CNN achieving the highest accuracy at 97% [15]. In a subsequent study, Gope and Fukai (2022) enhanced the CNN model, achieving an increased accuracy of 98.17% [7].

Recent advancements in Convolutional Neural Network (CNN) models have been extensively explored, with notable research focusing on ResNet, VGG, and MobileNet architectures. Janandi and Cenggoro (2020) conducted a comparative study on the performance of ResNet-152 and VGG16 in classifying 160 coffee beans into three categories, finding that ResNet-152 achieved the highest accuracy at 73.3% [5]. Building on this, Febriana et al. employed MobileNetV2 and ResNet-18 models to classify 8000 Arabica coffee beans into four categories, with MobileNetV2 attaining the highest accuracy of 81.31% [2].

Despite the variety of models utilized in previous research, the highest accuracy has often been achieved with a relatively low volume of data, potentially indicating overfitting. MobileNetV2 has demonstrated the highest accuracy with compact data in the multiclass classification of coffee beans at 81.31%.

However, there has been a lack of research exploring the classification of coffee beans using more modern models such as transformers. With advancements in deep learning methods for image classification, there is potential to improve the accuracy of coffee bean classification using alternative methods.

Therefore, this research aims to test the latest computer vision transformer-based architectures, such as ViT, DeiT, and Swin Transformer models, for coffee bean classification using the USK-Coffee dataset. Bayesian optimization is employed to enhance and optimize model performance. Additionally, augmentation on the fly, including random horizontal and vertical flips, random contrast, and random rotation, are implemented to enrich the dataset and enhance model performance. The selection of the USK-Coffee dataset is based on its comprehensiveness, comprising 8,000 images of Arabica coffee beans from Aceh and its recent release in 2022, making it the latest dataset available for Indonesian coffee beans.

3. Methodology

3.1. Overview

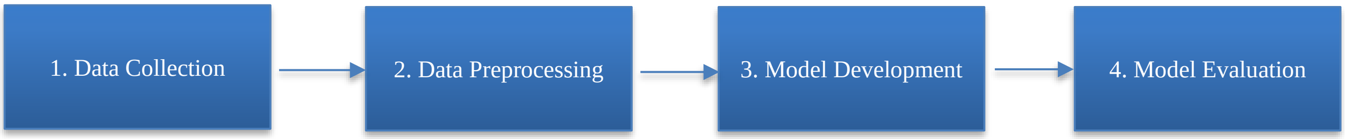


Fig. 1 Proposed method of coffee bean image classification

The proposed method comprises four steps, as illustrated in Figure 1 above. The first step involves data collection, where coffee bean images are gathered from the dataset and stored in the researchers' database. Subsequently, in the data preprocessing step, the images are cropped and resized to facilitate easier learning by the model. Following this, three transformer-based architectures, including ViT, DeiT, and Swin Transformers, are developed for feature extraction, coupled with trainable linear layers for classification. The final step involves model evaluation, where the data is divided into training, validation, and testing sets. Hyperparameter tuning is performed using Bayesian optimization, and the best model for each type is saved. Finally, the best-performing models are tested and compared to select the most optimized model.

3.2. Dataset

In the data collection step, the USK-Coffee dataset is obtained, containing 8000 images of coffee beans from Aceh, Indonesia. The USK-Coffee dataset comprises four classes: premium, longberry, peaberry, and defect, as depicted in Figure 2. A round, large shape and bluish-green color characterize premium beans. Longberry beans are elongated and large. Peaberry beans are small, round, and dense. Defect beans exhibit damaged or incomplete appearances, often appearing hollow due to uneven sizing. Each image in the USK-Coffee dataset measures 3x256x256 pixels. Published in 2022 by Febriana et al., the USK-Coffee dataset is relatively new and offers comprehensive data pertaining to Arabica coffee beans in Indonesia [2].

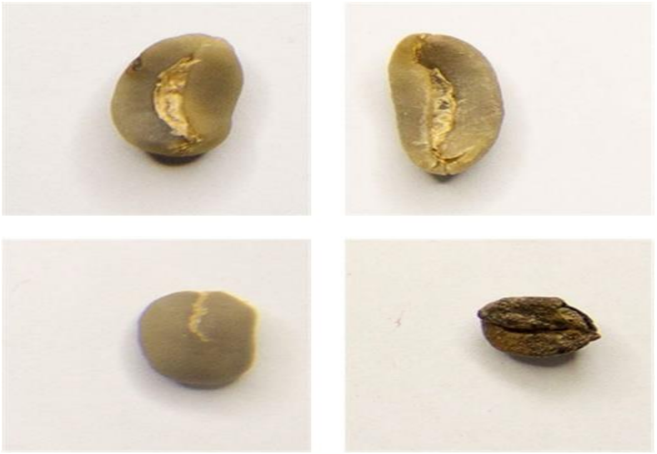


Fig. 2 Samples of the USK-Coffee dataset (premium, longberry, peaberry, defect)

3.3. Data Preprocessing

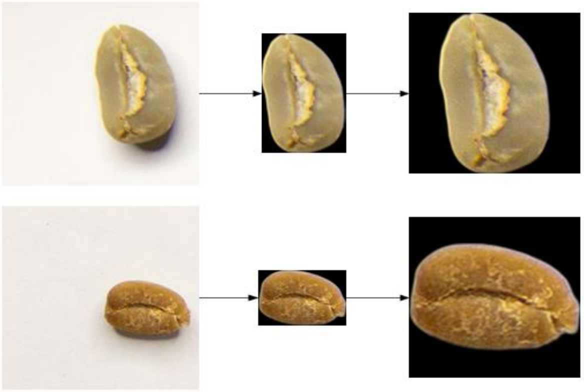


Fig. 3 Crop and enlarge the coffee bean to a Size of 3x224x224

Figure 3 illustrates the preprocessing steps. During the data preprocessing phase, the researchers crop the background of the coffee and enlarge its image. Additionally, the researchers also resize the original image from 3x256x256 to the standard size of 3x224x224, suitable for the transformer-based architecture models. The primary objective of this step is to ensure that the models focus on the coffee object accurately.

3.4. Model Development

In the following step, the researchers develop models by comparing three deep learning architectures for feature extraction to classify green coffee beans: Vision Transformer (ViT), Data Efficiency in Transformer (DeiT), and Swin transformer. Below is an explanation of the ViT, DeiT and Swin transformers used in this research.

3.4.1. ViT Model

The first developed model utilizes a Vision Transformer (ViT). ViT is a model that represents images as a set of vectors or "tokens" within a transformer network [8]. The vision transformer was initially proposed by Dosovitskiy et al. in 2021, marking a significant innovation in applying transformers to image data, whereas transformers had previously been predominantly used for Natural Language Processing (NLP). ViT employs the concept of self-attention, enabling the model to understand relationships between different parts of an image by assigning importance values to patches and focusing on relevant information. Figure 4 illustrates how the Vision Transformer (ViT) model operates. Initially, the input image is divided into fixed-size patches.

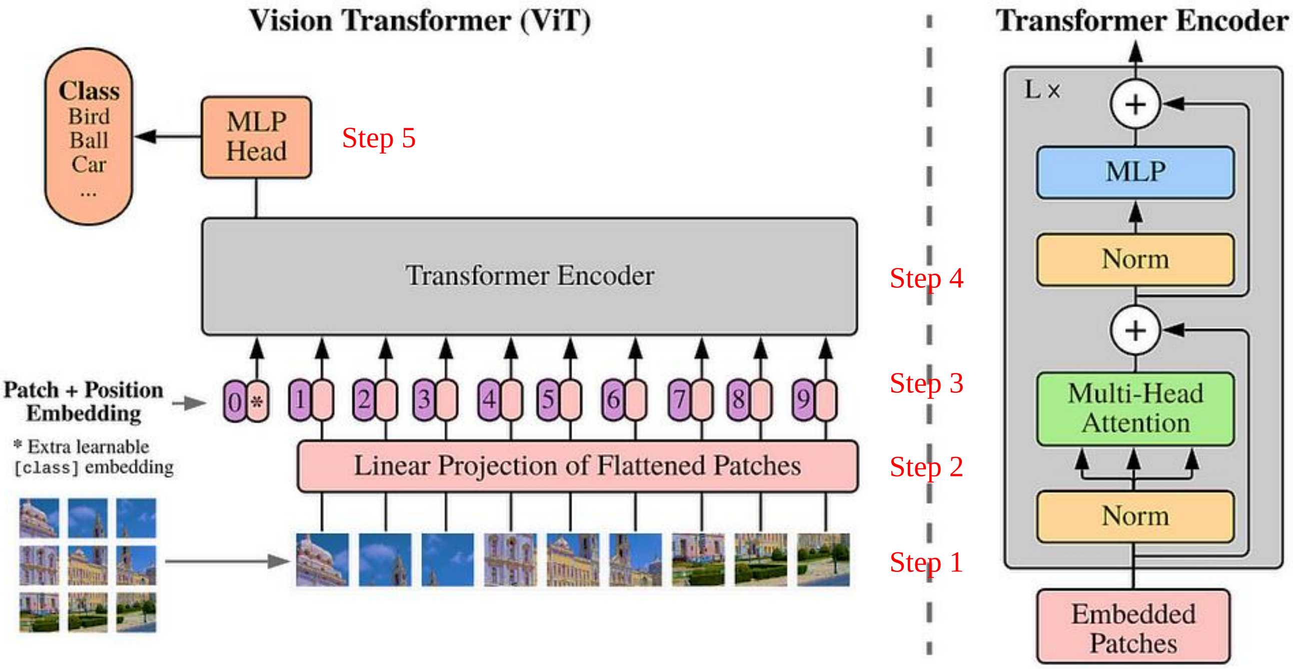


Fig. 4 Model ViT[8]

$$Number\ of\ Patches\ N = \left(\frac{H}{P}\right) \times \left(\frac{W}{P}\right) \tag{1}$$

Next, the patches are flattened, converting the image from 2D to 1D patches and linearly embedded using a fully connected layer.

$$Patch\ vector\ size = P \times P \times C \tag{2}$$

In the third step, positional embeddings are added to the patches to retain the positional information of the image. A special classification token (CLS) is added to the positional embeddings. These patch vectors are then passed as input to the transformer encoder.

The transformer encoder processes these inputs through multiple layers of self-attention and MLP blocks to output a final sequence, which is used for classification tasks. The classification layer consists of an MLP with one hidden layer during pre-training and a single linear layer for fine-tuning.

3.4.2. DeiT Model

The second model developed uses DeiT. DeiT employs mechanisms such as attention pooling to reduce the complexity of calculations in recognizing patterns and image representation, and it utilizes knowledge distillation, transferring knowledge from a larger model to a lighter one. Based on baseline studies, DeiT models employ convolution networks as the teacher model, and this architecture achieves competitive results compared to ViT by using a smaller number of images [9].

Figure 5 illustrates the operation of the Data-efficient Image Transformer (DeiT) model. Initially, the input image is divided into fixed-size patches.

In the DeiT model, the image is processed similarly to the Vision Transformer (ViT) model by dividing it into 16x16 patches, which are then transformed into image vectors or patch tokens. Next, positional encoding is added to the patch embeddings to maintain the spatial order information of each patch.

DEiT adds a special class token to the sequence of patch embeddings. This token serves as the final representation used for image class prediction after passing through all encoder layers. The transformer encoder processes these inputs through multiple layers of self-attention and MLP blocks to output a final sequence.

One innovation in DeiT is the addition of a distillation token designed to receive extra information from a teacher model during training. In Touvron et al.'s research, the teacher model used is a Convolutional Neural Network (CNN), which has advantages over the transformer's inductive bias and is simpler, allowing it to learn more easily with less data.

After passing through all encoder layers, the class token and distillation token predict the class. The class token predicts the main image class, while the distillation token helps improve the prediction with additional information from the teacher model.

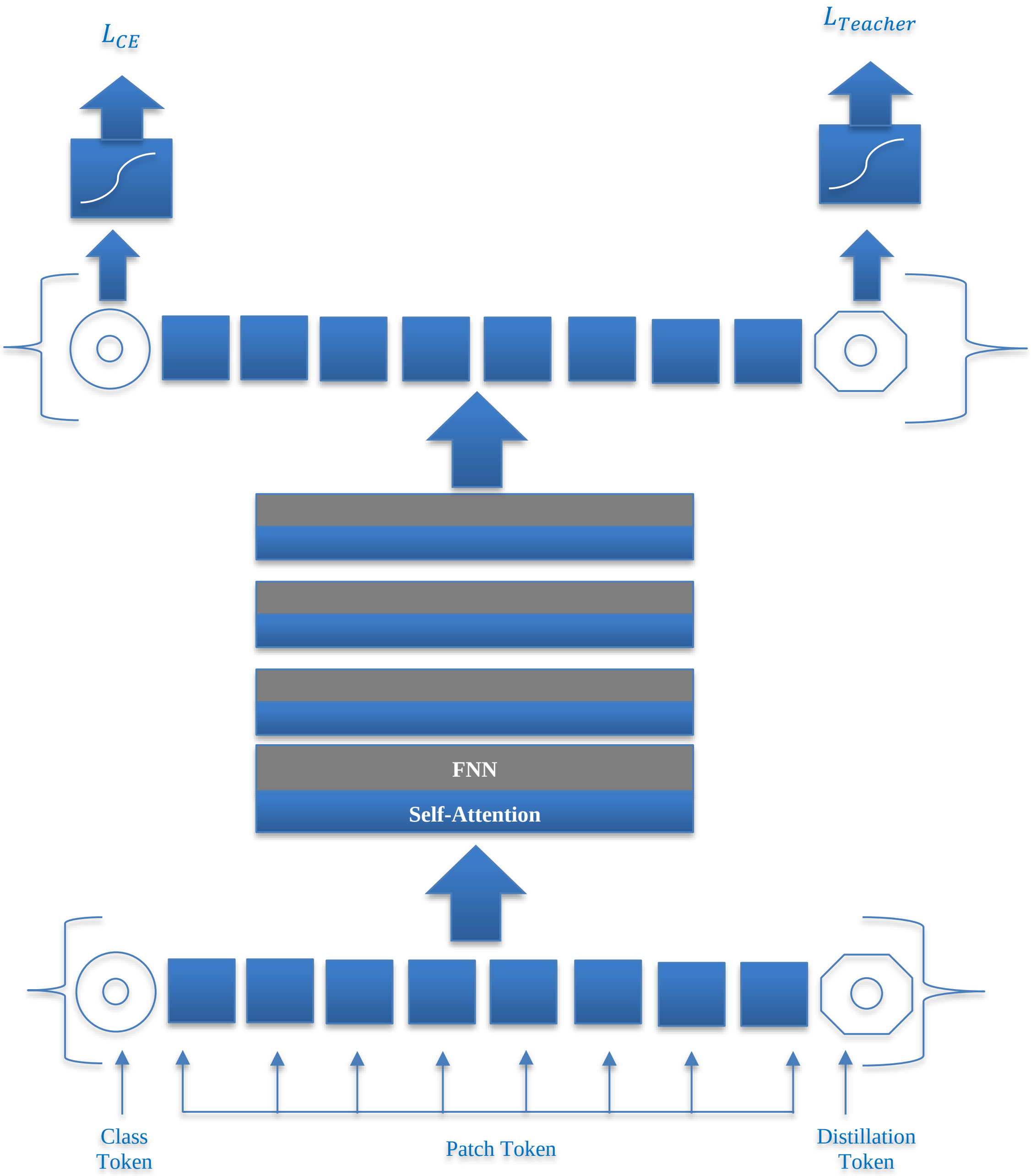


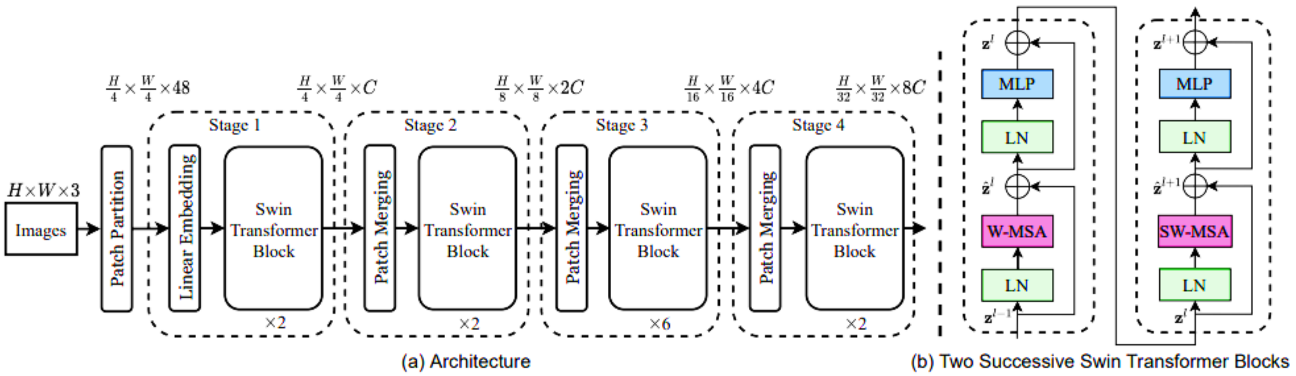
Fig. 5 Model DeiT [9]

3.4.3. Swin Transformer

The last model developed is the Swin Transformer model. The Swin transformer, proposed by Liu et al., represents a development of the Vision transformer using a shifted window [10].

Swin transformer partitions the input image into non-overlapping patches, contrasting with some previous transformer-based approaches that used overlapping patches. The swin transformer model processes each patch through a series of transformer layers. This hierarchical approach helps capture both local and global information in an image.

Figure 6 illustrates how the swin transformer model functions. The Swin transformer processes the input image by first dividing it into small fixed-size patches, such as 4x4 pixels, which are then flattened and projected into a higher dimension. These patches are processed in swin transformer blocks, utilizing Window-based Multi-Head Self-Attention (W-MSA) and Shifted Window-based Multi-Head Self-Attention (SW-MSA) to generate feature representations. Similar to ViT and DeiT, the swin transformer adopts only the encoder part of the transformer (without the decoder), as it requires only one-way processing from the image to the class label.



(a) Architecture

(b) Two Successive Swin Transformer Blocks

Fig. 6 Model swin transformer [10]

Each Swin transformer block includes residual connections and layer normalization for improved stability and convergence. Patch merging layers combine information from adjacent patches, reducing spatial resolution and increasing channel dimensions.

This process is repeated multiple times, with the final output used for tasks such as image classification, object detection, or image segmentation. The innovation of the Swin transformer lies in its computational efficiency and enhanced performance by using shifted windows to capture better spatial relationships in images. In this research, the model is applied to classify coffee bean images.

With its hierarchical and progressive approach, the Swin transformer can focus on important information within each window without having to process the entire image at once. This reduces the need for memory and computation during processing. This approach also improves efficiency in image processing, especially for high-resolution images that require greater computational resources.

3.4.4. Architecture Developed

Figure 7 depicts the utilization of pretrained models in this research for feature extraction. Specifically, from models ViT and DeiT, the researchers utilize the last output block, which corresponds to layer number 149, and send it to a modified classifier comprising sequential trainable layers.

For the Swin transformer, the researchers employ the last block, identified as layer number 170, for input to the modified classifier. Each model's classifier incorporates a linear layer, ReLU activation function, and dropout layer at the model's head.

Throughout the training phase, researchers implement data augmentation to enhance the dataset dynamically. The data augmentation techniques employed include random horizontal and vertical flips, random contrast, and random rotation.

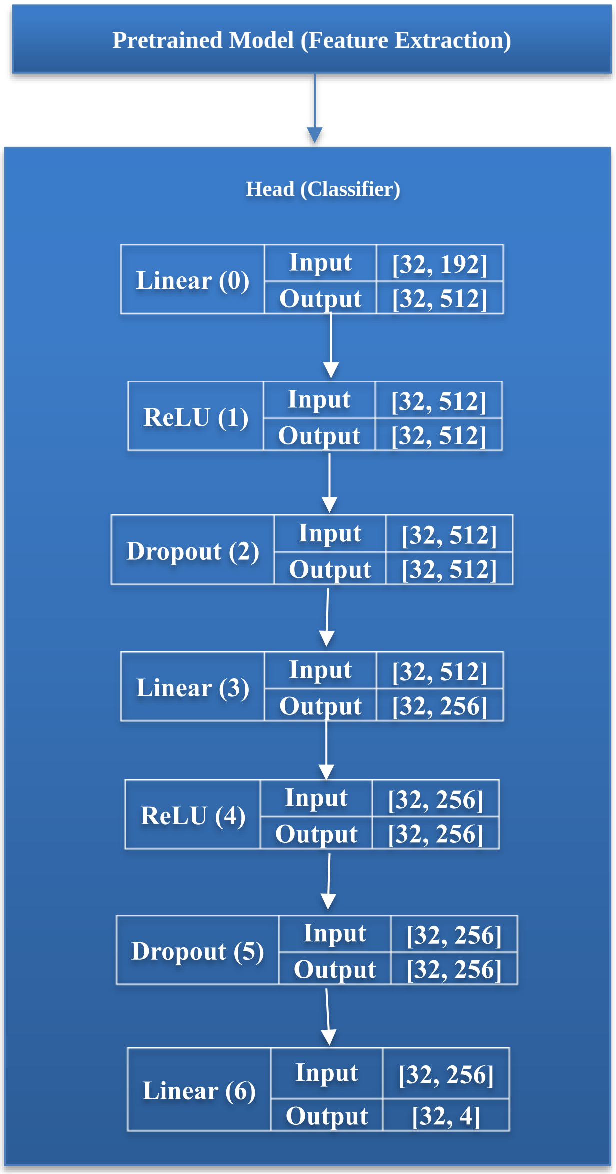


Fig. 7 Model architecture developed

3.5. Model Evaluation

Once the models have been developed, the USK-Coffee dataset will be fed into them. The USK-Coffee dataset is partitioned into training, validation, and test sets, with their distribution established by the dataset creators [2]. The allocation of these datasets is presented in Table 1 below. The training and validation sets will be employed during the training and tuning phases for each transformer-based model architecture. In the training process, the model has a fixed parameter setup, as shown in Table 2. A step size of 3 and a gamma of 0.97 are used for scheduling the learning rate. The step size and gamma together define a step-wise decay schedule for the learning rate. These parameters control how often the learning rate is adjusted and how much it is decreased with the AdamW optimizer. This research also implements early stopping with patience of 5, a maximum of 100 epochs, and a batch size of 32 to obtain the best model for each transformer-based architecture type.

During the training, Bayesian hyperparameter tuning is implemented to optimize the models. The parameters tuned include learning rate, weight decay rate, and dropout layer rate, as shown in Table 3. The objective of tuning is to maximize validation accuracy. In order to utilize resources efficiently, the trial is set to run 100 times within a maximum of 7200 seconds. After the training and tuning processes are complete, the best model is saved to storage. The researchers then use the test dataset to measure the performance of the best model from each transformer-based architecture. The performance results in a confusion matrix, which is tabulated using an accuracy score and average F1 score. Finally, the three models are compared to select the best-optimized model.

Table 1. Distribution of USK-Coffee dataset

Class	Training	Validation	Test
Peaberry	1200	400	400
Longberry	1200	400	400
Premium	1200	400	400
Defect	1200	400	400
Total	4800	1600	1600
Total Data	8000		

Table 2. Parameter setup used in this research

Parameter	Number
Batch Size	32
Epoch	100
Patience	5
Gamma	0.97
Step Size	3

Table 3. Parameter tuning used in this research

Parameter	Range
Learning Rate	0.0001-0.1
Weight Decay Rate	0.1-0.5
Drop Out Rate	0.001-0.1

4. Results and Discussion

4.1. Training and Validation Results

4.1.1. Training and Validation Result for ViT Model

During the training and Bayesian hyperparameter tuning process, the researchers obtained parameter settings for ViT, as shown in Table 4. Using these parameters, the best model was attained with maximum validation accuracy. Figures 8 and 9 below show the performance of the model using training and validation data.

Figure 8 depicts the loss graph of the ViT model utilizing both training and validation data. The graph demonstrates a consistent downward trend in training data loss, commencing at approximately 100% and steadily decreasing to 70%. Similarly, the validation data loss exhibits a declining trend after the 5th epoch, diminishing from around 90% to 70%. Notably, the ViT model achieves a loss of 74.13% on the training data and 71.35% on the validation data. Subsequently, in Figure 9, the accuracy graph of the ViT model employing both training and validation data is presented. The graph illustrates commendable performance, characterized by an escalating accuracy with increasing epochs. In the training data, accuracy demonstrates a stable and consistent ascent, ranging from 70% to 80%, indicating progressive learning from the training data.

Moreover, accuracy in the validation data shows an upward trajectory, peaking at epoch 12 with an accuracy of 80%. The ViT model attains its peak accuracy at 80.13% on the training data and 85.81% on the validation data. This trend underscores the model's capacity to generalize effectively, evident in the rising accuracy across both training and validation datasets.

Table 4. Parameter setting for VIT model

Parameter	Number
Learning Rate	0.0024
Weight Decay Rate	0.0038
Drop Out Rate	0.24

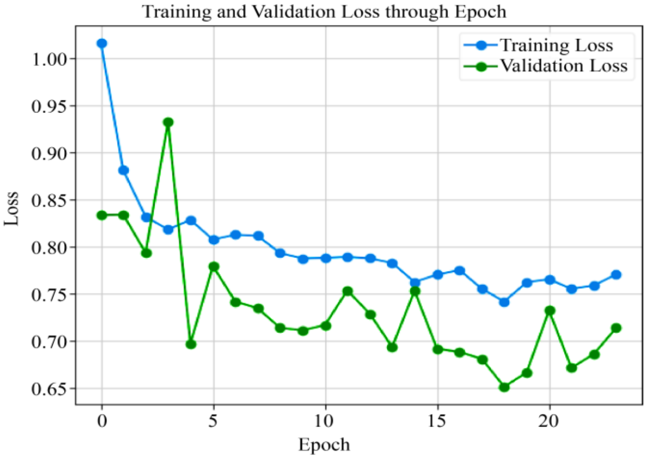


Fig. 8 Training and validation loss for ViT

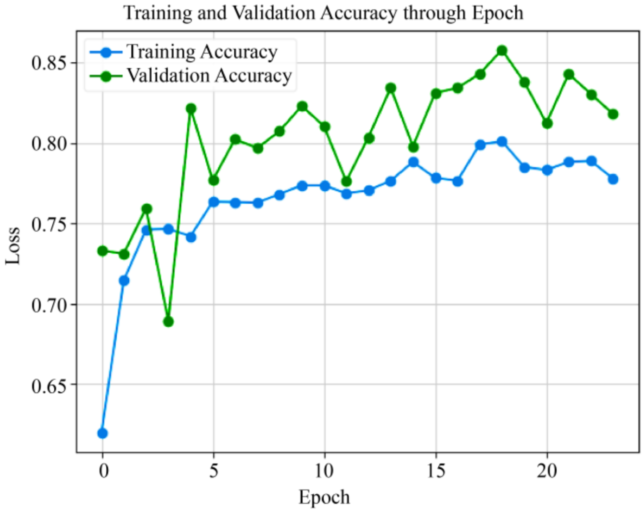


Fig. 9 Training and validation accuracy for ViT

4.1.2. Training and Validation Result for DeiT Model

Next, the evaluation of the DeiT model is conducted. The parameters retrieved from Bayesian hyperparameter tuning are presented in Table 5. Figures 10 and 11 below present the performance of the DeiT model utilizing training and validation data.

Figure 10 exhibits the loss graph of the DeiT model utilizing training and validation data. The graph illustrates a significant reduction in loss on the training data, declining from 100% to 85% and subsequently stabilizing around 75%, indicative of effective learning from the training data.

Conversely, the validation data loss displays a similar initial decrease but shows fluctuations, experiencing an increase in loss at epoch 8 before declining again until reaching approximately 70%. The DeiT model achieves a loss value of 74.45% on the training data and 67.31% on the validation data.

In Figure 11, the accuracy graph for both training and validation data using the DeiT model is depicted. The graph demonstrates a substantial increase in accuracy on the training data, starting from around 65% and stabilizing around 80%, signifying enhanced efficacy in learning from the training data.

Similarly, accuracy on the validation data exhibits an ascending trend, albeit with fluctuations, culminating in a peak above 82%. The model achieves its highest accuracy values at 79.42% on the training data and 83.69% on the validation data.

Table 5. Parameter setting for DeiT model

Parameter	Number
Learning Rate	0.0018
Weight Decay Rate	0.0012
Drop Out Rate	0.113

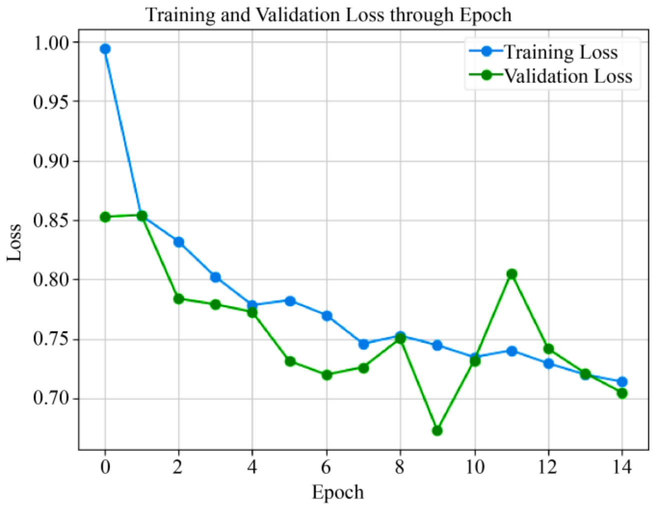


Fig. 10 Training and validation loss for DeiT

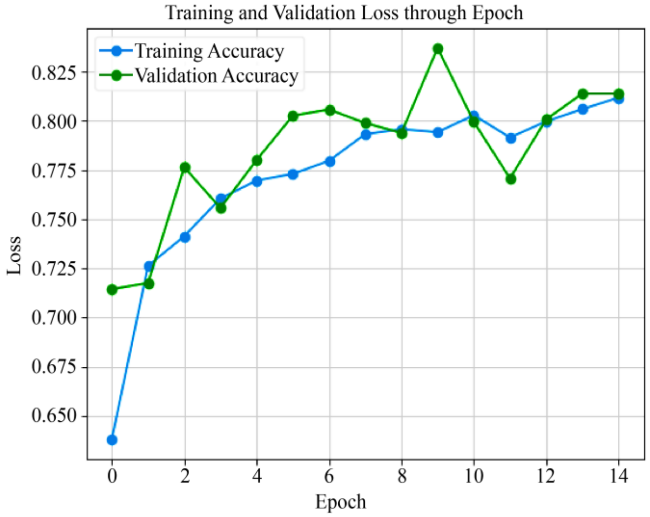


Fig. 11 Training and validation accuracy for DeiT

4.1.3. Training and Validation Result for Swin Transformer

Lastly, the Swin transformer models were evaluated. Through a Bayesian hyperparameter tuning process, parameter settings for the Swin Transformer were obtained, as depicted in Table 6. Utilizing these parameters, the optimal model with maximum accuracy was achieved. Figures 13 and 14 illustrate the performance of the model using training and validation data.

Figure 12 depicts the loss graph of the Swin Transformer model on the training and validation data presented. On the training data, the loss value decreases rapidly from an initial value of 100% to less than 80%, steadily decreasing further to nearly 70%. This trend suggests effective learning from the training data. Meanwhile, on the validation data, the loss value exhibits notable fluctuations, initiating at 85% and gradually decreasing to below 65%. This model achieves its lowest loss value at 72.22% on the training data and 63.96% on the validation data.

Table 6. Parameter setting for SWIN Transformer

Parameter	Number
Learning Rate	0.0032
Weight Decay Rate	0.07
Drop Out Rate	0.113

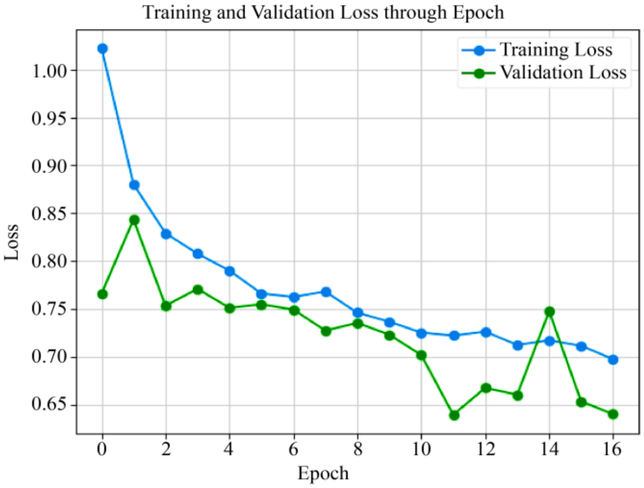


Fig. 12 Training and validation loss for Swin Transformer

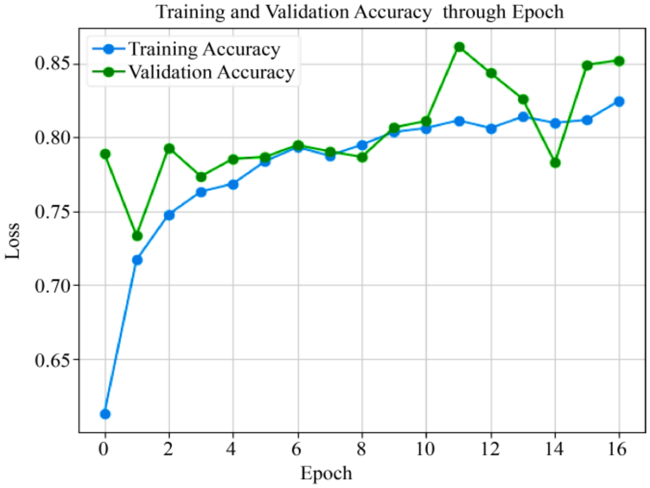


Fig. 13 Traning and validation accuracy for Swin Transformer.

Figure 13 illustrates the accuracy graph of the Swin Transformer model utilizing training and validation data. On the training data, the accuracy value experiences a sharp rise from around 65% to over 75% in the early epochs, gradually increasing to above 80%. This trend signifies an effective learning process from the model on the training data. Conversely, on the validation data, the accuracy value demonstrates fluctuations, albeit generally trending upwards, surpassing 85%. The Swin transformer model achieves its highest accuracy values at 81.17% on the training data and 86.19% on the validation data. These accuracy values exceed those of the previous two models, indicating superior performance in classifying green Arabica coffee beans. This highlights the Swin transformer model's optimal performance, particularly when dealing with limited datasets.

4.1.4. Summary of Training and Validation Results

The performance results of the training and validation models from previous research are presented in Table 7 and compared with the results from this research, shown in Table 8. As depicted, the model ResNet-18, as per prior research, still achieves the highest validation and training accuracy. Following closely is the Swin Transformer, with a validation accuracy of 86.16%, a validation loss of 63.96%, a training accuracy of 81.17%, and a training loss of 72.22%.

Notably, the Swin Transformer also receives the lowest validation loss. Its architecture enables the model to capture both global and local contexts of the coffee bean image, resulting in superior performance compared to other transformer models. Subsequently, the ViT model secures the third-best performance, boasting an 85.81% validation accuracy, a 71.35% validation loss, an 80.13% training accuracy, and a 74.13% training loss.

The higher accuracy of ViT over DeiT suggests that accuracy relies more heavily on the transformer architecture, considering DeiT combines transformer and CNN. DeiT, with a validation accuracy of 83.69%, a validation loss of 67.31%, a training accuracy of 79.42%, and a training loss of 74.45%, may prioritize efficiency in using transformers with less data. However, compared to MobileNetV2 from previous research, DeiT demonstrates greater accuracy in predicting coffee bean multiclass image classification.

Table 7. Summary of training and validation results from previous research

Model Name	Train Loss	Train Acc.	Val Loss	Val Acc.
Mobile NetV2[1]	Not Reported	85.88%	Not Reported	81.18%
ResNet -18[1]	Not Reported	98.81%	Not Reported	95.06%

[1] Based on reported results in [2]

Table 8. Summary of training and validation results from this research

Model Name	Train Loss	Train Acc.	Val Loss	Val Acc.
ViT	74.13%	80.13%	71.35%	85.81%
DeiT	74.45%	79.42%	67.31%	83.69%
Swin Transformer	72.22%	81.17%	63.96%	86.19%

4.2. Testing Results

Table 9. Testing results from previous research

Model Name	Accuracy	Average F1 – Score
MobileNetV2[1]	81.31%	Unreported
ResNet-18[1]	81.13%	Unreported

[1] Based on reported results in [2]

Table 10. Testing results from this research

Model Name	Accuracy	Average F1 – Score
ViT	82.25%	82.02%
DeiT	81.12%	81.02%
Swin Transformer	84.75%	84.80%

Table 9 presents accuracy and F1-score results on the testing data from previous research, juxtaposed with the findings from this research displayed in Table 10.

From the data displayed, the Swin transformer achieved the highest accuracy at 84.75% with an average F1-score of 84.80%. This performance notably surpasses that of previous studies utilizing convolutional architecture (CNN), such as MobileNetV2, with an accuracy score of 81.31% and an F1 score of 81.03%.

Despite the Swin transformer's validation score being lower than ResNet-18, it demonstrates superior performance in testing results. Following this, the ViT also demonstrates higher accuracy with a score of 82.25% and an average F1 score of 82.02%.

Conversely, the DeiT, a combination of transformer and CNN, obtains the lowest score with an accuracy of 81.12% and an average F1-score of 81.02%. These findings underscore the superior accuracy of transformer-based architectures in predicting coffee bean multiclass image classification.

5. Conclusion and Future Works

Based on this research, the Swin Transformer outperforms other models in coffee bean multiclass image classification, achieving an accuracy score of 84.75%. This score surpasses previous studies which utilized MobileNetV2, a CNN-based architecture and attained an accuracy score of 81.1%. Additionally, the ViT achieves a higher score in data testing compared to previous studies, with an accuracy score of 82.25%. In contrast, the DeiT obtains the lowest score, with an accuracy of 81.12%.

The research demonstrates that transformer-based architectures are more accurate in predicting coffee bean multiclass image classification. The preprocessing process, which involves cropping and enlarging the size of coffee beans, facilitates the ability of the transformer-based architecture to focus and learn the data effectively. Bayesian hyperparameter tuning also contributes to optimal learning by retrieving parameters such as learning rate, dropout rate, and weight decay rate, which produce the best model. These optimal rates prevent overfitting and enable accurate learning.

This research is limited to the development, evaluation, and optimization of deep learning solutions that yield better performance. Therefore, it is recommended for future research to implement the model into tools that enable people to classify Arabica green bean coffee.

Acknowledgments

The authors would also like to extend gratitude to Febriana et al. for allowing and providing access to the USK-Coffee dataset.

References

[1] Coffee Report and Outlook April 2023 - ICO, International Coffee Organization, pp. 1-39, 2023. [Online]. Available: https://icocoffee.org/documents/cy2022-23/Coffee_Report_and_Outlook_April_2023_-_ICO.pdf

[2] Alifya Febriana et al., "USK-COFFEE Dataset: A Multi-Class Green Arabica Coffee Bean Dataset for Deep Learning," *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, Malang, Indonesia, pp. 469-4736, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Yudhi Adhitya et al., "Feature Extraction for Cocoa Bean Digital Image Classification Prediction for Smart Farming Application," *Agronomy*, vol. 10, no. 11, pp. 1-16, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Nurun Najmi Amanin, and Galuh Wilujeng Saraswati, "Classification of Arabica Coffee Green Beans Using Digital Image Processing Using the K-Nearest Neighbor Method," *Journal of Applied Intelligent System*, vol. 7, no. 2, pp. 111-119, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Robby Janandi, and Tjeng Wawan Cenggoro, "An Implementation of Convolutional Neural Network for Coffee Beans Quality Classification in a Mobile Information System," *2020 International Conference on Information Management and Technology (ICIMTech)*, Bandung, Indonesia, pp. 218-222, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Igor R. Fermo et al., "Development of A Low-Cost Digital Image Processing System for Oranges Selection Using Hopfield Networks," *Food and Bioproducts Processing*, vol. 125, pp. 181-192, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] 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," *AIMS Agriculture and Food*, vol. 7, no. 1, pp. 149-167, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Alexey Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv*, pp. 1-22, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Hugo Touvron et al., "Training Data-Efficient Image Transformers & Distillation through Attention," *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 10347-10357, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ze Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 9992-10002, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Xilu Wang et al., "Recent Advances in Bayesian Optimization," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1-36, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Edwin R. Arboleda, Arnel C. Fajardo, and Ruji P. Medina, "Classification of Coffee Bean Species using Image Processing, Artificial Neural Network and K Nearest Neighbors," *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, Bangkok, Thailand, pp. 1-5, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] Mila Jumarlis, Mirfan Mirfan, and Abdul Rachman Manga, "Classification of Coffee Bean Defects Using Gray-Level Co-Occurrence Matrix and K-Nearest Neighbor," *ILKOM Jurnal Ilmiah*, vol. 14, no. 1, pp. 1-9, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Nelly Oktavia Adiwijaya et al., "The Quality of Coffee Bean Classification System Based on Color by Using K-Nearest Neighbor Method," *Journal of Physics: Conference Series*, vol. 2157, pp. 1-8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Hira Lal Gope, and Hidekazu Fukai, "Normal and Peaberry Coffee Beans Classification from Green Coffee Bean Images Using Convolutional Neural Networks and Support Vector Machine," *International Journal of Computer and Information Engineering*, vol. 14, no. 6, pp. 189-196, 2020. [Google Scholar] [Publisher Link]