ICIC Express Letters
Volume **20**, Number **4**, April **2026**

ICIC International ©2026 ISSN 1881-803X
pp. **465–474**

CLASSIFICATION OF GREEN ARABICA COFFEE BEANS USING CNN-TRANSFORMER HYBRID MODEL

Kelvin Andreas, Clarice Arlin Wijaya, Dave Christian Thio and Rhio Sutoyo

Computer Science Department
School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ kelvin.andreas; clarice.wijaya; dave.thio }@binus.ac.id; rsutoyo@binus.edu

Received May 2025; accepted August 2025

Abstract. *Indonesia, as one of the largest coffee producers, faces challenges in assessing the quality of green arabica coffee beans before roasting. Conventional classification methods rely on manual inspection by experts, which is subjective, time-consuming, and inconsistent. Previous research predominantly used single-model architectures, either CNN-based for local feature extraction or Transformer-based for capturing global representations. However, single model approaches often struggle to effectively combine both feature types, limiting their ability to achieve optimal classification performance. To address this limitation, this research proposes a hybrid model combining CNN and Transformers to enhance classification accuracy by leveraging both local and global feature representations. Hyper-parameter tuning was also performed to optimize model performance. Using the USK-Coffee dataset, experimental results show that the hybrid model outperforms single-model approaches, achieving the highest accuracy of 91.88%. These findings highlight the effectiveness of the hybrid model in capturing local and global features, leading to improved classification performance and a more robust quality assessment system. By integrating local and global features, this approach provides a more accurate and reliable method for classifying green arabica coffee beans, offering significant potential for automation in the coffee industry and reducing dependence on manual inspection.*

Keywords: Deep learning, Computer vision, Coffee bean image classification, CNN, Transformer

1. **Introduction.** In global trade, coffee is among the most economically valuable agricultural products. As one of the world's largest coffee producers, Indonesia exported coffee worth 1.49 billion dollars from January to September 2024. Among the many varieties produced, arabica coffee stands out as one of Indonesia's signature varieties and is well-known worldwide [1].

However, the Indonesian coffee industry still faces challenges in sorting and classifying coffee beans before the roasting process, which significantly impacts quality and market value [2]. At the moment, sorting is done largely manually, relying on human vision. This manual process is time-consuming, labor intensive, and prone to errors, as the quality of the product is spoilt due to overworked and stressed staff [2]. Furthermore, currently available coffee bean sorting machines are mainly color sorters and density sorters. Color sorters utilize Charge-Coupled Device (CCD) sensor technology and high-resolution cameras to detect defects based on color and size [3]. While density sorters separate beans based on their specific gravity using airflow or vibrations to distinguish high-quality beans from defective ones [4]. These machines are typically designed for binary classification (i.e., normal or defective), making them insufficient for multi-class classification, such as

DOI: 10.24507/icicel.20.04.465

differentiating peaberry, longberry, premium, and defect beans [2, 5]. Their effectiveness heavily depends on predefined parameters, making it challenging to adapt to changing quality standards or new coffee bean varieties [6]. This limitation affects production efficiency and quality, particularly in meeting the growing export demand.

To address this challenge, several studies have explored the use of the USK-Coffee dataset, which consists of 8,000 images of green arabica coffee beans categorized into four classes: peaberry, longberry, premium, and defect. Research by Febriana et al. using ResNet-18 and MobileNet-V2 achieved an accuracy of 81.31%, slightly better than ResNet-18, which achieved 81.13% for multi-class classification [2]. However, for binary classification, in a study conducted by Islamy et al., ResNet-18 excelled with an accuracy of 91.50% on a balanced dataset [7]. In addition, another study by Nasution et al. using MobileNet-V2 with hyperparameter adjustments resulted in an accuracy of 88.19% [8].

A different approach was taken by Neto et al., who evaluated several CNN architectures such as AlexNet, ResNet-50, MobileNet-V3, and EfficientNet-B4. From the results of the research, EfficientNet-B4 recorded the highest accuracy of 88.44% after data augmentation was applied [9]. On the other hand, Izza and Kusuma explored the potential of Transformer-based models such as Vision Transformer (ViT), Data-efficient Transformer (DeiT), and Swin Transformer. From the research results, Swin Transformer recorded the highest accuracy of 84.75% on test data [10].

Based on our findings, previous research is still limited to the use of a single model to classify these green coffee beans, without exploring a combination of architectures that can capture local and global features more optimally [2, 7, 8, 9, 10]. This paper aims to address this gap by proposing a hybrid model approach to enhance multi-class classification performance for green arabica coffee beans. This approach combines CNN and Transformer architectures, where CNNs are used to extract local features, while Transformers capture global relationships between features within the images [11]. Integrating both architectures allows the model to not only improve accuracy, but also make it more robust in distinguishing visually similar classes and makes the approach relevant for other cases where local detail and global structure both matter.

The paper is structured as follows. The background and objectives are presented in the Introduction section. The Literature Review section reviews existing theories relevant to image classification and prior research on green arabica coffee beans classification. The Methodology section explains the methods and experiment steps in developing and evaluating the models. Then, the Results and Discussion section presents the experimental results and our findings, analyzing the effectiveness of each hybrid model. Finally, the Conclusions section provides a summary and ideas for future work.

2. Literature Review.

2.1. Hybrid deep learning model.

The hybrid deep learning model is an approach that combines two different models with the aim of complementing the advantages and disadvantages of each model so that it can get more optimal results. This research uses one model to extract local features and another model to extract global features. Local features refer to detailed and textured information derived from small areas in an image, such as edges, patterns, and contours, which are generally extracted using CNN architecture-based models. However, global features reflect the overall context and relationships between parts in an image, which are usually obtained through models with Transformer architecture [12].

2.2. Dataset comparison and reasons for choosing USK-Coffee dataset for classifying green arabica coffee beans.

There are two publicly available datasets for the classification of green arabica coffee beans, namely the USK-Coffee dataset and the Coffee Cobra dataset. The USK-Coffee dataset is one of the most complex and challenging

datasets for the classification of green arabica coffee beans because it has more varied and specific classes [2]. This dataset was developed by Syiah Kuala University, Banda Aceh, Indonesia, and consists of 8000 images divided into 4 different classes, namely longberry, peaberry, premium and defect. The longberry class represents coffee beans that have a long and large shape and are characterized by the presence of dicotyledon seeds [2]. Meanwhile, the peaberry class describes coffee beans with a smaller size and round shape and has a higher density compared to coffee beans in other classes [2, 10]. The premium class is identified from bluish green coffee beans, with a large size and a rounder shape [2, 10]. The class of defects is those of coffee beans that are defective, caused by errors in the processing process or pest attacks, resulting in beans that are not intact or are often broken [2, 8, 10].

The Cobra dataset is a coffee bean image data set used in the Open Source Coffee Sorter project, which aims to build a small-scale, low-cost coffee bean sorting machine using the transfer learning method [13]. This project focuses on optical sorting of green coffee beans using a simple computer such as the Raspberry Pi 3. This dataset consists of 4626 images divided into 2 classes, namely good and bad. The good class is high-quality coffee beans that have met the selection standards. However, the bad class is for coffee beans that have defects or do not meet quality standards.

This research uses the USK-Coffee dataset because it offers advantages over the Cobra dataset. The image quality in the USK-Coffee dataset is much better, with higher resolution and clearer details. Additionally, USK-Coffee offers a greater number of images, which enhances the model's training and results in more accurate performance. Another advantage of USK-Coffee is the diversity of classes it has. This dataset includes four different classes, namely peaberry, longberry, premium, and defect. This makes it more varied and specific than other datasets that only have two classes, good and bad.

2.3. **Recent studies on classifying green arabica coffee.** Several previous studies have explored the classification of Indonesian green arabica coffee beans using the USK-Coffee dataset, employing binary and multiclass classification approaches. In binary classification studies, coffee beans are categorized into normal and defect classes. One study [14] applied feature extraction techniques using the Gray-Level Co-occurrence Matrix (GLCM) and a Content-Based Image Retrieval (CBIR) system to measuring image similarity with reference samples. Several prior works have evaluated different deep learning architectures for classifying green Arabica coffee beans, reporting varied performance across experimental settings. Islamy et al. [7] demonstrated strong results in binary classification using CNN-based models, with ResNet-18 achieving 91.50% accuracy on a balanced dataset. In multiclass scenarios, Febriana et al. [2] compared ResNet-18 and MobileNetV2, where MobileNetV2 showed slightly better performance. Nasution et al. [8] further improved MobileNetV2 through hyperparameter tuning, indicating that optimization strategies can significantly influence model outcomes.

Other studies have broadened the comparison by evaluating multiple CNN architectures. Neto et al. [9] assessed models such as AlexNet, ResNet-50, MobileNetV3, and EfficientNetB4, highlighting the impact of architectural scaling and data augmentation on performance. In parallel, Transformer-based approaches have also been investigated. Izza and Kusuma [10] examined Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), and Swin Transformer, showing that hierarchical attention mechanisms can provide competitive results in image classification tasks.

Previous studies cited above have primarily focused on improving classification performance through dataset expansion, preprocessing techniques, and model parameter optimization. However, these efforts still rely on a single architecture, which limits the model's ability to generalize when faced with subtle visual variations among green Arabica coffee beans. Variations in lighting, surface defects, and slight color inconsistencies can introduce

ambiguity that a single-model approach may struggle to resolve consistently. Therefore, exploring alternative strategies that improve robustness and generalization remains necessary to achieve more reliable multi-class classification results [12].

Hybrid models have been widely used in related applications, such as plant disease detection. Lin et al. [15] proposed the LGNet model, which combines CNN and Transformer architectures to improve classification accuracy, outperforming existing state-of-the-art methods. Similarly, another study [16] introduced the AELGNet model, combining LocalNet and GlobalNet to achieve more detailed feature recognition and higher accuracy than previous studies.

Based on these findings, hybrid approaches present a promising alternative for improving accuracy in green arabica coffee bean classification. Moreover, the implementation of data augmentation and hyperparameter tuning has proven crucial in optimizing model performance. Therefore, this research aims to propose a new method that combines the strengths of CNN-based and Transformer-based architectures to achieve more accurate and efficient classification results.

3. **Methodology.**

This research presents an approach to enhance the accuracy of classifying green arabica coffee using hybrid deep learning models. The proposed methodology integrates feature extraction using both Convolutional Neural Networks (CNNs) and Transformer-based architectures to enhance the model's ability to capture both local and global features from images. The overview of the process is shown in Figure 1.

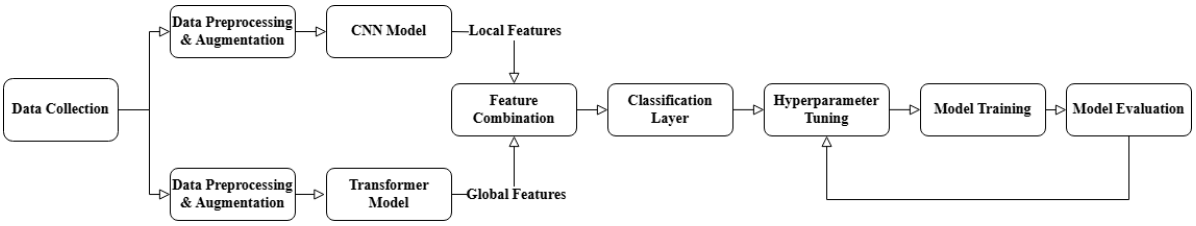


FIGURE 1. Research methodology

This architecture receives an input image and processes it through two parallel branches: a CNN-based encoder and a Transformer-based encoder. The CNN branch focuses on extracting local visual patterns such as texture and edges, while the Transformer branch captures long-range dependencies and global context. The features from both branches are then concatenated and passed to a classification head.

3.1. **Data collection.**

This research obtained data from the USK-Coffee dataset [2], which contains 8000 images of green arabica coffee. It consists of four classes: defect, longberry, peaberry, and premium (see Figure 2). The dataset is already divided into training (60%), validation (20%), and test (20%) sets, with each subset containing a balanced number of images from all classes.

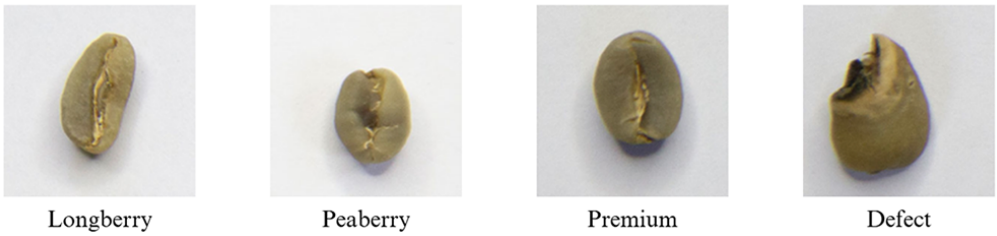


FIGURE 2. Example of USK-Coffee dataset

3.2. **Data preprocessing & augmentation.** The dataset will be processed through a preprocessing stage that is applied differently depending on the pretrained model used, where these models will follow the default preprocessing that has been trained previously with the ImageNet dataset. The training dataset was also augmented with several techniques, including random horizontal flipping and random rotation ($\pm 15°$), to improve model generalization and reduce overfitting. While the validation and testing datasets were only preprocessed without augmentation to maintain data consistency.

3.3. **Feature extraction & combination.** The hybrid model used in this research uses two pre-trained models as feature extractors (see Figure 3). The first pre-trained model based on CNN is used to extract local features from images, which includes ResNet-18, ResNet-50, MobileNet-V3, and EfficientNet-B4. Meanwhile, the second pretrained model based on the Transformer captures global context dependencies, which include ViT and Swin Transformer (see Table 1 for the list of hybrid model combinations used in this study). The selection of these models is based on previous research [2, 7, 8, 9, 10], in which these architectures have been successfully applied to the classification of green arabica coffee beans and have shown strong performance. The final classification layer in this pretrained model will be removed because it will only be used as a feature extractor. Furthermore, we use the default architecture and settings from the official torchvision models, as they are validated on ImageNet and have demonstrated strong generalization performance across various computer vision tasks. The features from those two models are combined into one concatenation of vectors representing local and global features before finally being forwarded into the classification layer. The hybrid model feature extraction and classification steps can be represented by Equation (1).

$$y = \text{Softmax}(W_2(\text{ReLU}(W_1(f_{\text{CNN}}(x) \oplus f_{\text{Transformer}}(x)) + b_1)) + b_2) \tag{1}$$

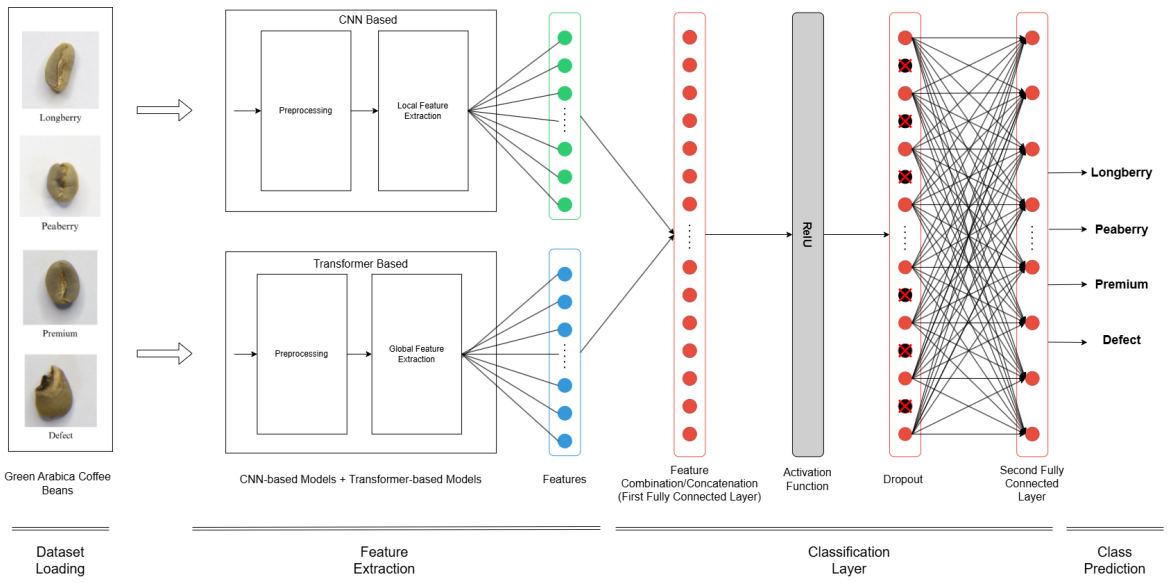


FIGURE 3. Proposed hybrid model architecture

3.4. **Classification layer.** After the local and global features are combined, this combination of features will be further processed through the fully connected layer, which acts as a classifier. The classifier consists of a linear layer with 512 hidden units, followed by a ReLU activation and a dropout of 0.5, then a final linear layer for classification. This architecture was chosen to provide a good balance between model complexity and generalization. The final output of this classification layer will predict the category of green arabica coffee beans in 4 classes, namely peaberry, longberry, premium, and defect.

Table 1. Hybrid models

Code	CNN-based model	Transformer-based model
R18-V	ResNet-18	Vision Transformer
R18-S	ResNet-18	Swin Transformer
R50-V	ResNet-50	Vision Transformer
R50-S	ResNet-50	Swin Transformer
MV3-V	MobileNetV3	Vision Transformer
MV3-S	MobileNetV3	Swin Transformer
EB4-V	EfficientNetB4	Vision Transformer
EB4-S	EfficientNetB4	Swin Transformer

3.5. **Model training & hyperparameter tuning.** The hybrid model is trained using the training dataset and validated through the validation dataset. To optimize model performance, hyperparameter tuning is performed using the grid search method, which aims to find the combination of parameters that provides the best results for the model. The parameters tested include the learning rate {1e-4, 1e-5, 1e-6} and the batch size {32, 64}. The learning rate {1e-4, 1e-5, 1e-6} is in line with common transfer-learning practice for large vision models for stable convergence speed or stability, and avoids catastrophic forgetting of pre-trained representations. The {1e-4} learning rate enables stable, yet moderately fast adaptation to the green arabica coffee dataset without destabilizing the model. The smaller values {1e-5, 1e-6} prioritize conservative fine-tuning to prevent catastrophic forgetting of pre-trained feature representations.

Meanwhile, the authors also chose batch sizes of 32 and 64. This choice follows prior best practices, batch size of 32 or 64 are a typical sweet spot in image classification as they provide enough samples per update for gradient estimation and batch normalization, yet are small enough to fit in memory and maintain good generalization (very large batches can hurt model generality). The training process is carried out for several epochs with an early stopping mechanism to avoid overfitting. With 8 hybrid models and the application of grid search on these hyperparameters, a total of 48 experiments are conducted.

3.6. **Model evaluation.** After the training process is complete, the model will be evaluated using the testing dataset. The evaluation is carried out using metrics such as accuracy, precision, recall, and F1-score, which are metrics commonly used in previous research. These metrics are chosen to allow for a direct comparison with results from previous experiments and to provide a comprehensive evaluation of model performance.

4. **Results and Discussion.** This study utilizes the Python programming language, specifically Python 3.11.11, along with the PyTorch framework version 2.5.1 for conducting experiments. The implementation was carried out on Google Colab, which provides a cloud-based environment with pre-configured dependencies, making it easier to set up and execute deep learning models. To enhance computational efficiency and reduce training time, the training process was accelerated using a GPU.

4.1. **Results for hybrid deep learning models.** Several hybrid models combining CNN and Transformer architectures were evaluated with different batch sizes and learning rates to determine the optimal configuration. The evaluation metrics included accuracy, precision, recall, and F1-score which provide detailed insights into the model's performance in classifying green arabica coffee beans. Table 2 presents the results of the experiments, where only the best hyperparameter combination is selected for comparison. The selection is based on the highest accuracy achieved among all hyperparameter configurations tested, ensuring that each model is represented by its optimal performance.

Table 2. Performance comparison of hybrid models

Model	Batch size	Learning rate	Accuracy	F1-score	Precision	Recall
R18-V	32	0.000001	90.38%	90.41%	90.75%	90.38%
R18-S	32	0.00001	87.94%	87.97%	89.25%	87.94%
R50-V	**64**	**0.00001**	**91.88%**	**91.85%**	**92.12%**	**91.88%**
R50-S	64	0.00001	89.38%	89.38%	90.13%	89.38%
MV3-V	32	0.000001	89.88%	89.84%	90.21%	89.88%
MV3-S	32	0.00001	90.25%	90.27%	90.64%	90.25%
EB4-V	32	0.000001	86.75%	86.90%	87.82%	86.75%
EB4-S	32	0.00001	90.19%	90.14%	90.67%	90.19%

In order to concentrate on high-performing configurations, only models with an accuracy greater than 85% are shown.

From the experiments, hybrid models involving EfficientNetB4 tended to have lower accuracy. Specifically, EfficientNetB4 combined with Vision Transformer produces the lowest accuracy of 86.75%. This result shows that despite EfficientNetB4 is known for its computational efficiency, this architecture might not adequately capture the essential local features required for this green coffee bean classification task. However, when EfficientNetB4 was combined with Swin Transformer, the accuracy improved significantly to 90.19%, showing that Swin Transformer might complement EfficientNetB4 feature extraction better than Vision Transformer.

The hybrid model combining MobileNetV3 and Swin Transformer model also performed well with an accuracy of 90.25%. Even though MobileNetV3 is designed as a lighter CNN architecture, this hybrid model still produced a competitive accuracy. It can be seen that lightweight CNN models can still be effective when paired with powerful Transformer.

A slightly higher accuracy of 90.38% was obtained from the hybrid combination ResNet-18 and Vision Transformer. ResNet architecture effectively captures subtle textures and patterns in coffee bean images. Hence, when combined with Vision Transformer global context awareness can enhance the classification accuracy. However, compared to deeper CNN models, ResNet-18 might lack sufficient depth and complexity to clearly differentiate between visually similar coffee bean classes. This becomes clearer when compared with ResNet50 + Vision Transformer which achieved the highest accuracy of 91.88%. This shows that deeper and more complex architecture of ResNet-50 allows richer and more discriminative local features. When paired with Vision Transformer, it maximizes both detailed local and global feature extraction capabilities to then produce better accuracy than other hybrid combinations tested in this study.

Figure 4 presents the accuracy comparison results between the best-performing hybrid deep learning model from this research and single-model approaches from previous research. Based on these comparisons, it is evident that hybrid deep learning models improve the accuracy of green arabica coffee bean classification, as they can capture both local and global features. This effectiveness is demonstrated by their superior performance compared to single-model approaches.

Figure 5 shows the accuracy and loss during the training and validation phases of the hybrid model R50-V. Based on Figure 5(a), it can be seen that both the training and validation accuracy gradually increased, reaching stability after several epochs, indicating that the model generalized well to unseen data. Furthermore, Figure 5(b) shows that both training and validation loss consistently decreased throughout training, with the validation loss stabilizing in later epochs. This shows that the model did not exhibit significant overfitting.

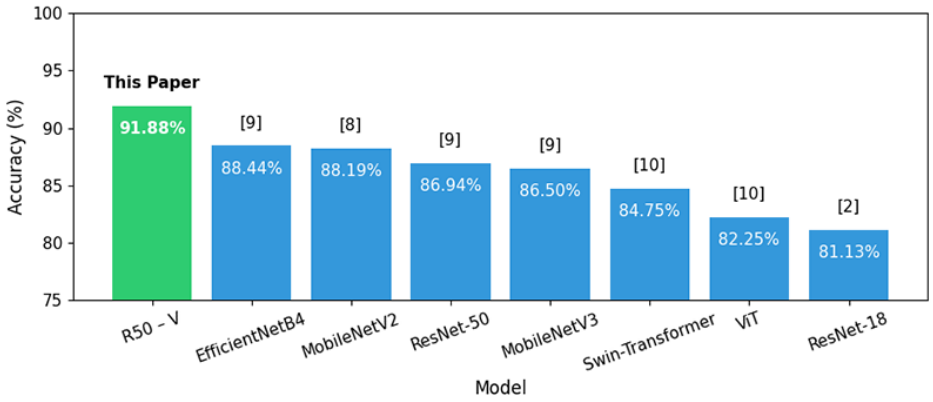
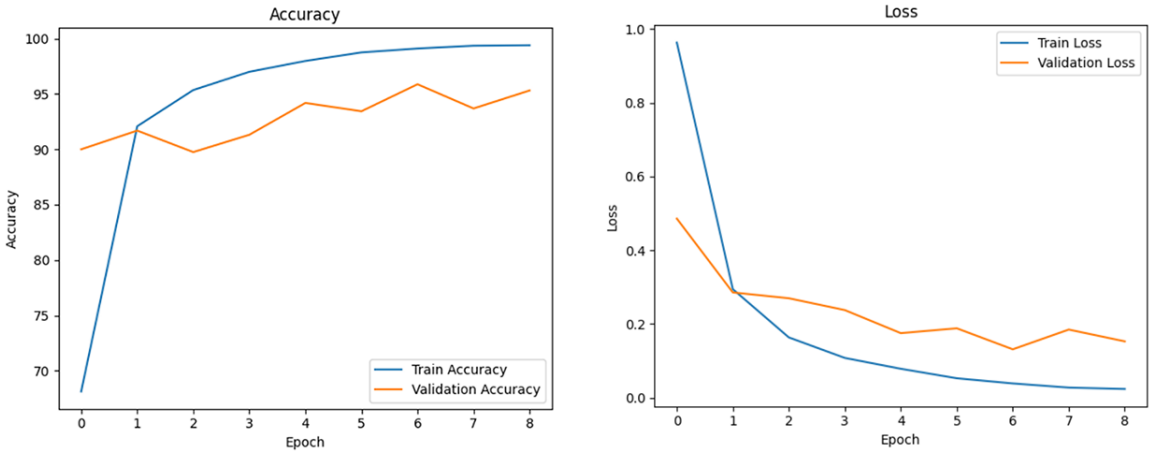


FIGURE 4. Comparison to previous research



(a) Accuracy (b) Loss

FIGURE 5. Training and validation performance of R50-V hybrid model

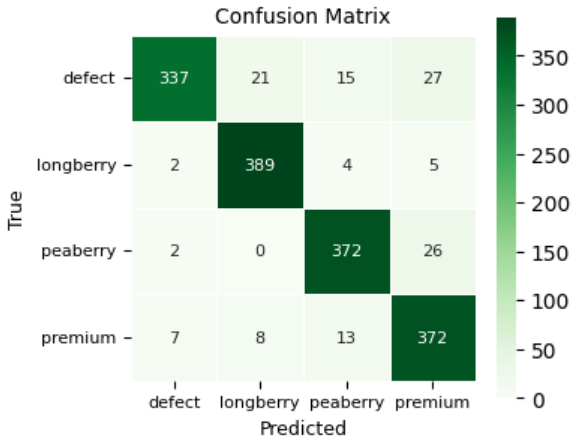


FIGURE 6. Confusion matrix of R50-V hybrid model

Based on the confusion matrix in Figure 6, the model shows strong performance in classifying peaberry, premium, and longberry coffee beans, indicated by the high numbers of correctly predicted samples. However, the defect class has a lower performance with several misclassifications into the premium and peaberry classes. These misclassifications might be caused by visual similarities between defect beans and other bean classes.

4.2. **Practical implications.** The findings from this study indicate that the hybrid modeling approach can offer meaningful improvements for the coffee industry, particularly in automating the green bean classification process. Using this method, producers can reduce the dependency on manual sorting, cut operational costs, and obtain more consistent results. These findings also offer valuable insights for developers looking to build practical and reliable classification tools which could be adapted for other image-based applications beyond coffee production.

4.3. **Implementation of R50-V hybrid model.** The best-performing hybrid model (R50-V) was deployed into a web-based application platform. As shown in Figure 7, the application first accepts an input image of a green arabica coffee bean, performs preprocessing, and then uses the R50-V hybrid model to classify the bean into one of four categories: peaberry, longberry, premium, or defect.

FIGURE 7. Implementation of R50-V hybrid model

5. **Conclusions.** The authors explored hybrid models for classifying green arabica coffee beans. Various combinations of CNN and Transformer-based models were tested to capture both local and global features from green arabica coffee bean images. The experiment shows that R50-V, which is a combination of ResNet-50 and Vision Transformer, achieved the highest accuracy of 91.88%. This hybrid model also gave better performance in precision, recall, and F1-score compared to previously studied single-model methods.

For future work, the authors recommend focusing on the implementation of the R50-V hybrid model in real-world coffee production. The hybrid model can be implemented in AI-powered sorting machines for real-time classification on production lines. This integration would automate the sorting process and replace manual labor. Additionally, it could also be connected to a central monitoring system that provides feedback to make continuous model adaptation and retraining. This ensures high accuracy despite changes in environmental conditions and bean characteristics. Moreover, future research could also explore whether deploying the hybrid model into cloud-based services would be possible.

Finally, the authors note that there are three potential ways to use the hybrid model in production environments: real-time (when information needs to be processed immediately), near real-time (when speed is important, but it is not needed immediately), or batch processing (when information processing can wait for days or longer). Each approach has distinct advantages and challenges. The authors hope that future research could address and evaluate these approaches with the R50-V hybrid model from this study.

Author Contributions. Kelvin Andreas: Conceptualization, Methodology, Software, Formal Analysis, Writing – Original Draft, Visualization. **Clarice Arlin Wijaya:** Conceptualization, Methodology, Software, Formal Analysis, Writing – Original Draft, Visualization. **Dave Christian Thio:** Conceptualization, Methodology, Software, Formal Analysis, Writing – Original Draft, Visualization. **Rhio Sutoyo:** Conceptualization, Validation, Writing – Review & Editing, Supervision, Project Administration.

REFERENCES

[1] Ministry of Agriculture of the Republic of Indonesia, *Tren 2025: Peluang Dan Daya Saing Kopi Indonesia (2025 Trends: Opportunities and Competitiveness of Indonesian Coffee)*, https://tanaman industri.bsip.pertanian.go.id/berita/tren-2025-peluang-dan-daya-saing-kopi-indonesia, 2025.

[2] A. Febriana, K. Muchtar, R. Dawood and C.-Y. Lin, USK-Coffee dataset: A multi-class green arabica coffee bean dataset for deep learning, *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pp.469-473, 2022.

[3] D. Opi, *Coffee Color Sorter: How It Works*, https://thecozycoffee.com/coffee-color-sorter/, 2023.

[4] N. Krueger, *Sorting Coffee Beans: Drying and Roasting Processes*, http://nicolekrueger.com/coffee/beans/drying-and-roasting/sorting.html, accessed in 2025.

[5] T. P. Tho, N. T. Thinh and N. H. Bich, Design and development of the vision sorting system, *Proc. of the 3rd International Conference on Green Technology and Sustainable Development (GTSD)*, pp.217-223, 2016.

[6] E. A. Garcial, *Coffee Basics: Optical Sorting in Coffee Production*, https://garciacoffee.com/coffee-basics-optical-sorting/, accessed in 2025.

[7] F. Islamy, K. Muchtar, F. Arnia, R. Dawood, A. Febriana, G. N. Elwirehardja and B. Pardamean, Performance evaluation of coffee bean binary classification through deep learning techniques, in *Innovative Technologies in Intelligent Systems and Industrial Applications, CITISIA 2022, Lecture Notes in Electrical Engineering*, S. C. Mukhopadhyay, S. M. Namal Arosha Senanayake and P. W. C. Withana (eds.), Cham, Springer, 2023.

[8] N. F. Nasution, C. Setianingsih and R. E. Saputra, Klasifikasi biji kopi arabika menggunakan convolutional neural network (Classification of arabica coffee beans using convolutional neural network), *e-Proceedings of Engineering*, vol.11, no.4, pp.3062-3069, 2024.

[9] R. P. Neto, P. Sousa, L. Moreira, P. God and J. Mari, Enhancing green coffee quality assessment through deep learning, *Anais do XVIII Workshop de Visão Computacional*, Porto Alegre, RS, Brasil, pp.84-89, 2023.

[10] M. Izza and G. Kusuma, Image classification of green arabica coffee using transformer-based architecture, *International Journal of Engineering Trends and Technology*, vol.72, pp.304-314, 2024.

[11] K. Shaheed, I. Qureshi, F. Abbas, S. Jabbar, Q. Abbas, H. Ahmad and M. Z. Sajid, EfficientRMT-Net – An efficient ResNet-50 and vision transformers approach for classifying potato plant leaf diseases, *Sensors*, vol.23, no.23, 9516, 2023.

[12] M. De Silva and D. Brown, Multispectral plant disease detection with vision transformer-convolutional neural network hybrid approaches, *Sensors*, vol.23, no.20, 2023.

[13] C. Jhamb, *Coffee Bean Quality Recognizer*, 2017, https://github.com/chirag-jhamb/coffee_bean_quality_recognizer, Accessed on March 2nd, 2025.

[14] S. Irianto, Refining content-based segmentation for prediction of coffee bean quality, *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol.14, no.2, pp.102-113, 2023.

[15] J. Lin, X. Zhang, Y. Qin, S. Yang, X. Wen, T. Cernava, Q. Migheli and X. Chen, Local and global feature-aware dual-branch networks for plant disease recognition, *Plant Phenomics*, vol.6, 0208, 2024.

[16] S. Sharma and M. Vardhan, AELGNet: Attention-based enhanced local and global features network for medicinal leaf and plant classification, *Computers in Biology and Medicine*, vol.184, 109447, 2025.