OPEN ACCESS

International Journal of Advances in Artificial Intelligence and Machine Learning Vol 2 No 2 (2025) https://doi.org/10.58723/ijaaiml.v2i2.405

Check for updates

Research Article

Comparative Study of CNN and Vision Transformers on Indonesian Tradisional Cakes Classification

Received: April 15, 2025	Revised: July 10, 2025	Accepted: July 11, 2025	Publish: July 13, 2025

Dedi Trisnawarman*, Adolf Asih Supriyanto, Viny Christanti Mawardi, Ugochi A Okengwu

Abstract:

Background of study: Food image classification is a challenging task in computer vision, particularly when dealing with traditional food items that exhibit subtle visual variations. While Convolutional Neural Networks (CNNs) have long been the standard for image recognition, their limitation in capturing long-range dependencies has led to the emergence of Vision Transformers (ViTs). In this context, the classification of Indonesian traditional cakes offers a culturally rich yet complex problem for automated image recognition systems.

Aims and scope of paper: This study aims to conduct a comparative analysis between EfficientNet-B0 (CNN-based) and ViT-B/16 (Transformer-based) architectures in classifying eight categories of Indonesian traditional cakes. The research evaluates not only classification accuracy but also the strengths and limitations of each model in handling fine-grained visual distinctions.

Methods: Both models were fine-tuned using the "Kue Indonesia" dataset from Kaggle. The methodology includes image preprocessing, model training with consistent parameters, and evaluation using accuracy, precision, recall, and F1-score. A confusion matrix was also used to visualize misclassifications and analyze per-class performance.

Result: ViT-B/16 achieved slightly higher accuracy (96.25%) compared to EfficientNet-B0 (95.62%). ViT performed better in classes with subtle variations, such as *kue lapis* and *kue dadar gulung*, while EfficientNet-B0 showed superior efficiency and high accuracy on visually distinct cakes.

Conclusion: Both CNN and ViT models demonstrate strong performance in traditional food classification. ViT is more robust in fine-grained visual analysis, whereas EfficientNet-B0 is preferable for resource-constrained environments. This study highlights the role of AI in supporting digital preservation of culinary heritage.

Keywords: EfficientNet, Fine-Grained Image Recognition, **I**ndonesian Kue Tradisional, Traditional Food Classification, Vision Transformers (ViT).

1. INTRODUCTION

Computer vision has advanced significantly in recent years, primarily due to innovations in deep learning techniques. Convolutional Neural Networks (CNNs) remain a cornerstone for image classification tasks, thanks to their hierarchical feature extraction capabilities (Mienye et al., 2025). Architectures like EfficientNet have achieved an efficient balance between accuracy and computational efficiency, leading to their widespread adoption in real-world applications such as medical imaging, object detection, and food recognition (Mingxing Tan, 2019). Specifically, EfficientNet models, including EfficientNet-B0, have been shown to be highly effective in various food recognition tasks due to their scalable architecture, allowing for robust feature learning even with limited data (Taufiqurrahman et al., 2024) (Alruwaili & Mohamed, 2025).

Nevertheless, CNNs are inherently limited by their local receptive fields, which restrict their ability to model long-range dependencies in visual data (Chen et al., 2024). To overcome this, Vision Transformers (ViTs) were introduced as a novel architecture that leverages self-attention mechanisms to model global relationships between image patches (Dosovitskiy et al., 2021). Since the introduction of the original ViT by (Dosovitskiy et al., 2021), several enhanced variants such as Swin Transformer (Z. Liu et al., 2021) and Data-efficient Image Transformers (DeiT) (Touvron et al., 2021) have been proposed. These models have demonstrated competitive or superior performance across various image recognition tasks, including those involving small datasets or fine-grained distinctions (Sikdar et al., 2025).

While these architectures have proven successful in general-purpose computer vision tasks, their application in culturally specific and fine-grained domains remains underexplored (Bhatt et al., 2021). Fine-grained classification in such domains is particularly

challenging due to the subtle visual distinctions between classes, often requiring models to capture intricate details that differentiate visually similar items (Wang & Wang, 2019). One vibrant yet challenging domain is the classification of Indonesian traditional cakes. These cakes are visually diverse and culturally significant, with differences in shape, texture, color, and presentation that are often subtle (Alba-Martínez et al., 2022) (Karlita et al., 2022). This study specifically addresses these challenges by utilizing a unique dataset of Indonesian traditional cakes, allowing for an in-depth analysis of model performance on visually similar yet distinct culinary items. The ability to automatically classify them would significantly benefit various applications, including culinary tourism, by enabling personalized recommendations and virtual tours (Suanpang & Pothipassa, 2024). Furthermore, it supports digital cultural preservation efforts by creating searchable and categorized archives of traditional recipes, and enhances menu digitization and augmented reality experiences for diners. Accurate classification also aids in nutritional tracking systems by precisely identifying ingredients and portion sizes, contributing to public health initiatives (Hjalager, 2022).

However, developing models for this task is complicated by several challenges, including class imbalance, low-resolution images, and variations in lighting and background (Sampath et al., 2021). These issues make it difficult for models to generalize effectively across classes.

Previous works in food image classification have primarily focused on global datasets such as Food-101 (D. Liu et al., 2025) or Recipe1M (Boyd et al., 2024), where CNNs have achieved robust performance. While valuable, these datasets often lack the cultural specificity and fine-grained visual distinctions present in regional cuisines (Zhang et al., 2023). More recently, Transformer-based models have also begun to show promise in food-related tasks (Nfor et al., 2025). For instance, (Isinkaye et al., 2024) demonstrated that SEViT effectively classified fine-grained visual patterns in plant disease images, a task analogous in complexity to food recognition. Similarly, (Banerjee et al., 2024) proposed a Nutritional Content Detection Using Vision Transformers-An Intelligent approach, highlighting the model's superior performance in identifying visually similar food items in varied environments. Our current study differentiates itself by focusing specifically on the "Kue Indonesia" dataset, which presents unique complexities related to subtle variations in shape, color, and texture that are not as pronounced in broader food datasets like Food-101 or Recipe1M. Furthermore, our comparative approach between CNNs and ViTs in this underrepresented cultural domain provides novel insights into their respective strengths and limitations.

This study compares EfficientNet-B0, a CNN-based model, and ViT-B/16, a baseline Vision Transformer. Both models were pretrained on the ImageNet dataset and fine-tuned on a curated dataset of Indonesian traditional cakes. We evaluate their accuracy, precision, recall, and F1-score performance, particularly their ability to distinguish between visually similar cake types. These metrics were chosen to provide a comprehensive assessment of the models' performance, offering insights into not only overall correctness (accuracy) but also their ability to correctly identify positive instances (precision and recall) and a balanced measure between them (F1-score), which is crucial for imbalanced datasets often encountered in fine-grained classification. By benchmarking these architectures in a culturally specific and visually complex setting, our study aims to (1) assess the strengths and limitations of CNNs versus ViTs in fine-grained classification tasks and (2) contribute to the growing body of research applying deep learning to underrepresented cultural domains in computer vision.

2. MATERIAL AND METHOD

This section outlines the experimental pipeline for performing a comparative analysis between Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to classify traditional Indonesian cakes. The methodology is organized into several key stages: dataset collection, pre-processing, model selection, training procedures, and evaluation metrics.

1. Data Collection

The "Kue Indonesia" dataset, created by Ilham (2020) and hosted on Kaggle, is a curated collection of images featuring a variety of traditional Indonesian cakes. This dataset, accessible at https://www.kaggle.com/datasets/ilhamfp31/kue-indonesia, is explicitly designed for image classification tasks. The dataset presents a unique challenge due to the visual similarities between different types of cakes and the diversity in color, texture, and presentation style. With a total size of approximately 303 MB and around 1,846 JPEG images, it provides a reasonably sized benchmark for developing and testing deep-learning models in food recognition.

Dataset Statistics and Class Distribution

To provide a more comprehensive understanding, here are some basic statistics about the "Kue Indonesia" dataset: The "Kue Indonesia" dataset contains 8 distinct classes of traditional Indonesian cakes. While the total number of images is approximately 1,846, the distribution across these classes is generally balanced, aiming to provide sufficient examples for each category.

Specifically, the dataset includes images for the following cake types:

1. Kue Klepon
2. Kue Lumpur
3. Kue Kastengel
4. Kue Putri Salju
5. Kue Serabi
6. Kue Dadar Gulung

7. Kue Lapis
8. Kue Risoles

Each of these classes typically has a comparable number of images, contributing to a relatively balanced class distribution. This balance is beneficial for training deep learning models, as it helps prevent bias towards dominant classes and allows models to learn features from all cake types effectively. While the exact count per class can vary slightly depending on the specific split (training/validation/testing), researchers have generally aimed for an even representation.

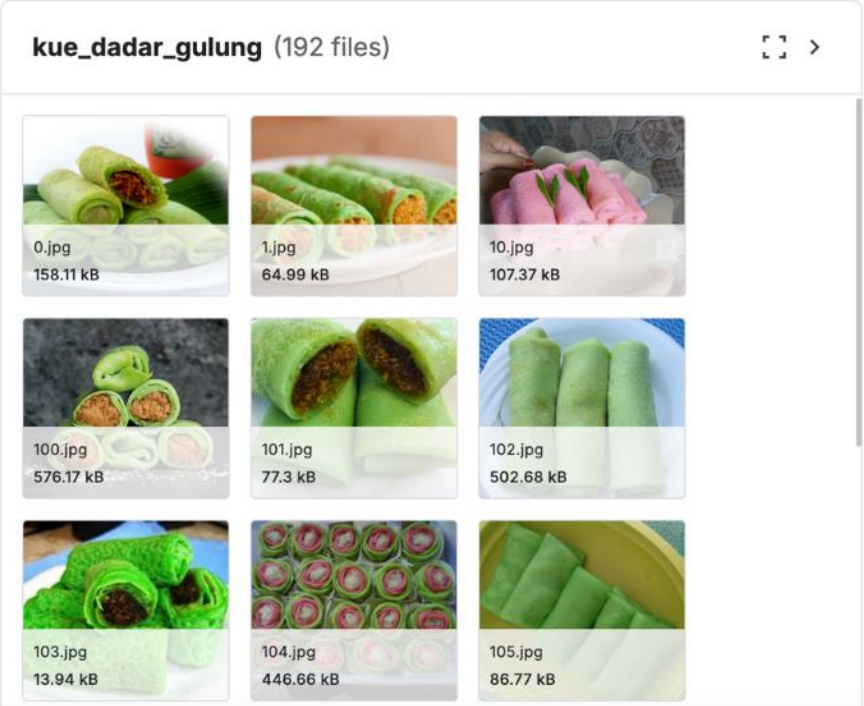

Figure 1. One of the example class, kue_dadar_gulung

In practical applications, several research studies have used the "Kue Indonesia" dataset to explore deep learning models such as ResNet50V2, EfficientNet, and even more recent architectures like Vision Transformers (ViT). (Sari & Chandra, 2025) These studies have shown promising results, with some models achieving high accuracy in identifying the various types of cakes. This demonstrates the dataset's potential as a benchmark for food recognition models, particularly within the context of Southeast Asian cuisine, which remains underrepresented in global computer vision datasets.

Overall, the "Kue Indonesia" dataset is a valuable resource for researchers and practitioners in machine learning and computer vision. It supports the development of image classification models and contributes to the digital preservation of Indonesia's rich culinary heritage. As such, it opens avenues for further innovation in culturally aware AI systems, food recommendation engines, and mobile dietary assistance and tourism applications.

2. Data Preprocessing

Several data preprocessing steps were applied before training to ensure consistency and enhance the learning capability of the models. These steps aimed to standardize the input format, normalize feature values, and improve generalization through data augmentation. All images in the dataset were first resized to a fixed resolution of 224×224 pixels, the standard input size for both EfficientNet-B0 and ViT-B/16 models. This resizing ensures compatibility with the pretrained model architectures and reduces computational costs during training.

3. Model Selection

This study investigates the effectiveness of two distinct deep learning architectures for image classification tasks: a Convolutional Neural Network (CNN) model and a Vision Transformer (ViT) model. The goal is to compare their performance in classifying images of traditional Indonesian Kue.

a. EfficientNet-B0 (CNN-based)

EfficientNet-B0 was selected as the CNN baseline due to its balance between accuracy and computational efficiency. EfficientNet introduces a compound scaling method that uniformly scales depth, width, and resolution using a predefined set of coefficients. The B0 variant represents the smallest and most lightweight model in the EfficientNet family, yet it performs competitively on various image classification benchmarks. This study used the pretrained EfficientNet-B0 weights trained on ImageNet and fine-tuned the model on the Kue Indonesia dataset. Figure 2 shows the Python code for model creation of the EfficientNet-B0.

```
import torch.nn as nn
import timm  # For pretrained models

model_cnn = timm.create_model('efficientnet_b0', pretrained=True)
model_cnn.classifier = nn.Linear(model_cnn.classifier.in_features, len(class_names))
```

Figure *2*. The EfficientNet-B0 model Python code model creation.

b. The EfficientNet-B0 model Python code model creation.

As a transformer-based model, ViT-B/16 treats image classification as a sequence modeling problem by dividing an image into non-overlapping patches and processing them with self-attention mechanisms. ViT-B/16 splits each input image into 16×16 patches and applies a transformer encoder to model the relationships between patches. This approach has shown promising results in various computer vision tasks, particularly large-scale datasets. For this study, the ViT-B/16 model was initialized with pretrained weights from ImageNet-21k and fine-tuned on the Kue dataset. Figure 3 shows the Python code for model creation of the ViT-B/16.

```
model_vit = timm.create_model('vit_base_patch16_224', pretrained=True)
model_vit.head = nn.Linear(model_vit.head.in_features, len(class_names))
```

Figure *3*. The ViT-B/16 model Python code model creation.

Both models represent two fundamentally different paradigms in image recognition: spatial inductive biases in CNNs versus long-range global attention in Transformers. By evaluating and comparing these models on the same dataset, we aim to highlight their strengths, limitations, and suitability for fine-grained food classification in Indonesian cuisine.

4. Training and Evaluation

To assess the performance of the selected models, EfficientNet-B0 and ViT-B/16 were fine-tuned on the Kue Indonesia dataset using a consistent training and evaluation pipeline. The training process was conducted using PyTorch, leveraging GPU acceleration when available to expedite the computation.

5. Training Configuration

Both models were trained using the cross-entropy loss function, which is appropriate for multi-class classification problems. The Adam optimizer was selected for its ability to adapt learning rates during training, with an initial learning rate set to 1e-4. Each model was trained for 30 epochs, with a batch size of 32, and early stopping was considered based on validation loss to prevent overfitting.

During training, the models were evaluated on the validation set after each epoch to monitor loss and accuracy progression. The best-performing model based on validation accuracy was saved for final evaluation on the test set.

6. Evaluation Metrics

The models were evaluated using accuracy as the primary metric, supported by precision, recall, and F1-score to provide a more detailed performance analysis, especially for imbalanced classes. A confusion matrix was also used to visualize misclassifications and understand patterns between similar kue types. These combined metrics offered a comprehensive view of each model's effectiveness in classifying traditional Indonesian kue.

7. Visualization and Reporting

The validation and test results included confusion matrices plotted with color gradients to illustrate prediction strengths and weaknesses. A detailed classification report with precision, recall, and F1-score per class was also generated and analyzed. These evaluations helped me understand overall model accuracy and how each model handled challenging or ambiguous examples of traditional cakes.

3. RESULT AND DISCUSSION

3.1 Results

This section presents the classification performance of the two selected models—EfficientNet-B0 and Vision Transformer (ViT-B/16)—on the Indonesian Traditional Cakes dataset. The models were evaluated based on accuracy, precision, recall, and F1-score across eight cake classes.

To evaluate and compare the performance of the two models, EfficientNet-B0 and ViT-B/16, several standard classification metrics were used: accuracy, precision, recall, and F1-score. These metrics comprehensively assess each model's ability to identify various classes of traditional Indonesian cakes correctly. The comparison results are shown in Table 1.

Table 1. The comparison results

Approaches	Accuracy (%)	Class	Precision	Recall	F1-score
EfficientNet-B0	95.62	kue_dadar_gulung	0.94	0.85	0.89
		kue_kastengel	1.00	0.95	0.97
		kue_klepon	0.95	1.00	0.98
		kue_lapis	0.86	0.95	0.90
		kue_lumpur	0.95	0.95	0.95
		kue_putri_salju	1.00	1.00	1.00
		kue_risoles	1.00	1.00	1.00
		kue_serabi	0.95	0.95	0.95
ViT-B/16	96.25	kue_dadar_gulung	0.90	0.95	0.93
		kue_kastengel	0.95	1.00	0.98
		kue_klepon	1.00	0.95	0.97
		kue_lapis	0.90	0.95	0.93
		kue_lumpur	1.00	0.90	0.95
		kue_putri_salju	1.00	0.95	0.97
		kue_risoles	0.95	1.00	0.98
		kue_serabi	1.00	1.00	1.00

Table 1 shows the classification performance of EfficientNet-B0 and ViT-B/16 across eight traditional cake classes. The classification of traditional Indonesian cakes presents a unique challenge due to the subtle differences in appearance, texture, and color among the various types. This study evaluated two state-of-the-art deep learning models: EfficientNet-B0, a convolutional neural network (CNN), and ViT-B/16, a Vision Transformer model. Both models were pre-trained on ImageNet and fine-tuned using a curated dataset of Indonesian cakes. The primary objective was to assess their ability to accurately classify eight categories of traditional cake through a comparative analysis of accuracy, precision, recall, and F1-score.

ViT-B/16 slightly outperformed EfficientNet-B0, achieving a classification accuracy of 96.25%, compared to 95.62% for EfficientNet-B0. While the margin of difference is modest (0.63%), it reflects the advantage of Vision Transformers in modeling global image dependencies and context, particularly in tasks involving fine-grained image classification. The performance of each model was further evaluated on a per-class basis to identify specific strengths and limitations.

When analyzing class-wise F1-scores, ViT-B/16 showed stronger performance in several classes such as kue dadar gulung (0.93 vs. 0.89), kue kastengel (0.98 vs. 0.97), kue lapis (0.93 vs. 0.90), and kue serabi (1.00 vs. 0.95). These results suggest that the Vision Transformer can capture these cakes' nuanced patterns and textures, especially those with layered or complex structures. Its self-attention mechanism enables it to learn from the image holistically, which is particularly beneficial for visually similar or subtly varied classes.

In contrast, EfficientNet-B0 exhibited better or equal performance in several categories. It achieved a perfect F1-score (1.00) in recognizing kue putri salju and kue risoles and performed equally well classifying kue lumpur (0.95). This indicates that CNNs maintain strong discriminative power for classes with well-defined and easily distinguishable features, especially when the visual patterns are localized and not overly complex.

Despite their differences, both models performed exceptionally well, with F1 scores above 0.89 for all classes. EfficientNet-B0 stands out for its efficiency and fast inference times, making it a viable choice for deployment on resource-constrained devices. Meanwhile, ViT-B/16 offers slightly better overall accuracy and robustness, making it suitable for applications with higher classification precision.

While ViT-B/16 edges out EfficientNet-B0 in terms of overall accuracy and class-wise performance for specific categories, both models demonstrate excellent capability in classifying traditional Indonesian cakes. These findings highlight the potential of applying advanced deep learning techniques to preserve and promote local culinary heritage through digital means. Future work may consider model ensembling or including more diverse training data to enhance classification performance.

Comparison Evaluation Metric

The following graph compares the performance of the two models used. Each model is evaluated using accuracy, precision, recall, and F1-score metrics, as shown in Figure 4.

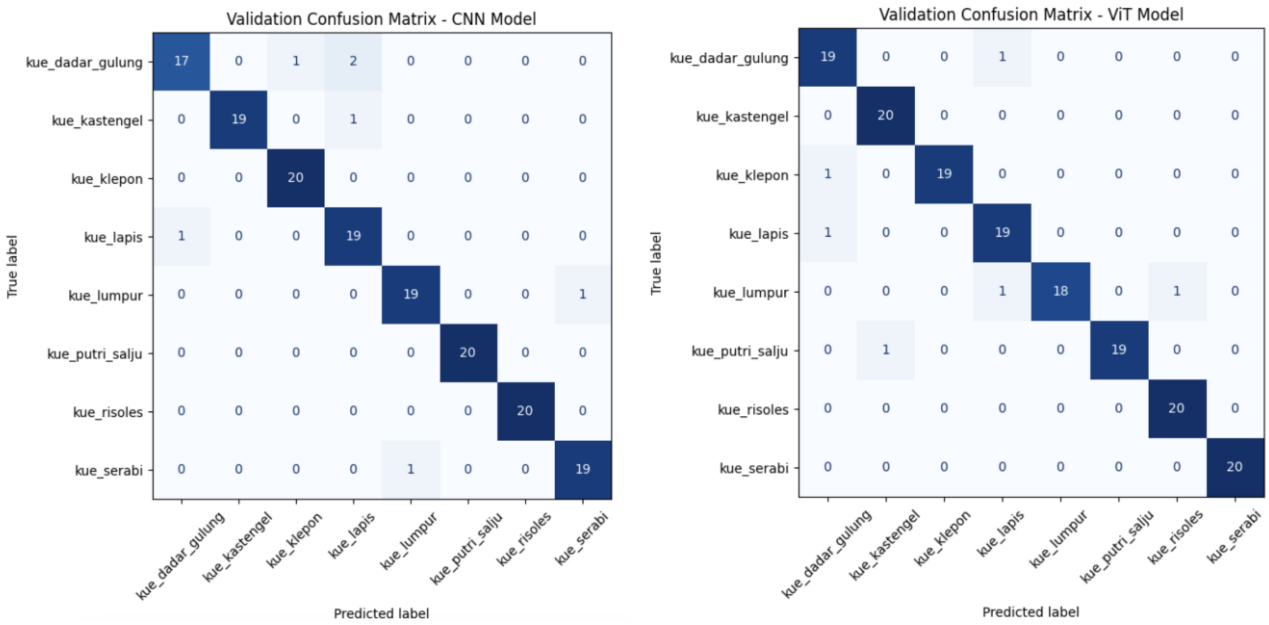

Figure 4. The confusion matrices for each different model.

These metrics were calculated per class to provide detailed insight into each model's strengths and weaknesses across different cake categories. Additionally, macro and weighted averages were used to summarize performance across all classes in a balanced and class-size-aware manner.

These metrics ensure a fair and interpretable comparison between CNN-based (EfficientNet-B0) and Transformer-based (ViT-B/16) architectures, highlighting their effectiveness in image-based food classification tasks.

3.2 Discussion

This comparative study's results illustrate that Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) are well-suited for classifying traditional Indonesian cakes, achieving high overall performance. However, ViT-B/16 slightly outperformed EfficientNet-B0 in terms of both accuracy and consistency across multiple classes. While the margin of improvement in overall accuracy was modest (96.25% vs. 95.62%), the class-wise precision, recall, and F1 scores suggest that the Vision Transformer exhibits greater robustness in capturing subtle intra-class variations and generalizing across complex visual patterns.

This performance advantage can largely be attributed to the inherent self-attention mechanism in ViTs, which enables the model to consider relationships between all parts of an image simultaneously. Unlike CNNs, which primarily focus on local receptive fields and hierarchical feature extraction, ViTs model global context more effectively, allowing them to distinguish subtle differences in color gradients, textures, and shapes. This property is beneficial in datasets like this, where many traditional cakes share similar appearances due to overlapping ingredients, presentation styles, or lighting conditions in the images.

For example, ViT-B/16 achieved higher F1 scores for kue dadar gulung and kue lapis, two classes known to be visually similar due to their cylindrical and layered green appearances, respectively. This suggests that the global feature representation of ViT helps disambiguate between classes where local textures alone may be insufficient. In contrast, EfficientNet-B0's slightly lower recall in some classes indicates that it may occasionally misclassify these cakes due to over-reliance on local or lower-level patterns.

Interestingly, both models performed exceptionally well on categories with distinct visual identities, such as kue risoles and kue putri salju, each achieving perfect or near-perfect scores. These cakes typically feature unique external traits, such as crumb coatings or powdered sugar finishes, which are relatively easy for both architectures to recognize. This aligns with findings in other food classification studies, where high accuracy is typically achieved on items with consistent and distinguishable shapes and textures.

Despite ViT-B/16's superior performance, EfficientNet-B0 remains a strong contender, particularly when considering computational efficiency. EfficientNet is known for its compound scaling method, which optimizes model depth, width, and resolution to balance accuracy and resource usage better. This makes EfficientNet-B0 highly suitable for real-world deployments on edge devices or mobile applications where computational power is limited.

From a practical standpoint, the results underscore that Vision Transformers represent a promising direction for food classification, especially in domains requiring fine-grained visual discrimination. However, the trade-off

regarding training complexity, model size, and inference time must also be considered. A lightweight and fast model like EfficientNet-B0 might still be preferable for real-time applications or large-scale deployment scenarios, even if it offers slightly lower accuracy.

This comparative analysis validates the growing potential of Vision Transformers in traditional food classification while also reinforcing the utility of optimized CNNs like EfficientNet-B0. As datasets become more diverse and complex, especially incultural or culinary domains, leveraging models that capture global contextual cues will be increasingly important. Future research could explore hybrid models, data augmentation techniques tailored to fine-grained food categories, or model ensembling to combine the strengths of both architectures for even more accurate and efficient classification systems.

This section contains the conclusions of the research that has been conducted. Conclusions contain answers to the research questions, and state your conclusions clearly and concisely.

3.2.1 Implications

This comparative study illustrates that Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) are well-suited for classifying traditional Indonesian cakes, achieving high overall performance. ViT-B/16 slightly outperformed EfficientNet-B0 in terms of both accuracy and consistency across multiple classes. The margin of improvement in overall accuracy was modest (96.25% vs. 95.62%). This performance advantage can largely be attributed to the inherent *self-attention* mechanism in ViTs, which enables the model to consider relationships between all parts of an image simultaneously and model global context more effectively, allowing them to distinguish subtle differences in color gradients, textures, and shapes. This property is beneficial in datasets like this, where many traditional cakes share similar appearances due to overlapping ingredients, presentation styles, or lighting conditions in the images.

For example, ViT-B/16 achieved higher F1 scores for *kue dadar gulung* and *kue lapis*, two classes known to be visually similar due to their cylindrical and layered green appearances, respectively. This suggests that the global feature representation of ViT helps disambiguate between classes where local textures alone may be insufficient. In contrast, EfficientNet-B0's slightly lower recall in some classes indicates that it may occasionally misclassify these cakes due to over-reliance on local or lower-level patterns. Both models performed exceptionally well on categories with distinct visual identities, such as *kue risoles* and *kue putri salju*, each achieving perfect or near-perfect scores. These cakes typically feature unique external traits, such as crumb coatings or powdered sugar finishes, which are relatively easy for both architectures to recognize. This aligns with findings in other food classification studies, where high accuracy is typically achieved on items with consistent and distinguishable shapes and textures.

3.2.2 Research contribution

This research provides valuable contributions in several aspects:

1. Comprehensive Comparison: The study presents a comprehensive comparative analysis of EfficientNet-B0 (a CNN-based model) and ViT-B/16 (a Transformer-based model) for the classification of Indonesian traditional cakes, an area that remains underexplored in computer vision.
2. Application in Culturally Specific Domain: The research demonstrates the potential of applying advanced deep learning techniques to preserve and promote local culinary heritage through digital means. It also addresses the underrepresentation of Southeast Asian cuisine in global computer vision datasets.
3. Insights into Model Capabilities: The study highlights the strengths and limitations of both CNNs and ViTs in fine-grained classification tasks, with ViTs showing an advantage in capturing long-range dependencies and global visual features, which is crucial for discriminating between visually similar items.

3.2.3 Limitations

Despite providing significant insights, the study has several limitations:

1. Dataset Size and Diversity: Although the "Kue Indonesia" dataset is of reasonable size , future research could explore larger and more diverse datasets to further improve classification performance and model generalization.
2. Focus on Two Architectures: The study is limited to comparing EfficientNet-B0 and ViT-B/16. Other architectures or newer variants of CNNs and Transformers might exhibit different performances.
3. Computational Considerations: While EfficientNet-B0 stands out for its efficiency and fast inference times , the trade-off regarding training complexity, model size, and inference time for ViT-B/16 must also be considered for practical deployment.

3.2.4 Suggestions

Based on the results and limitations of this study, several suggestions for future research include:

1. Model Ensembling and Hybrid Architectures: Future work could explore model ensembling or include more diverse training data to enhance classification performance. Exploring hybrid models that leverage the complementary strengths of both CNNs and transformers could also be a promising direction for even more accurate and efficient classification systems.
2. Domain-Specific Data Augmentation Strategies: Developing data augmentation techniques tailored to fine-grained food categories could help improve model performance, especially in handling diverse and complex datasets.
3. Deployment on Resource-Constrained Devices: For real-time applications or large-scale deployment

scenarios, optimizing models to run efficiently on edge devices or mobile applications with limited computational power is an important area for further research.
4. Exploration of Broader Datasets: Future research could utilize broader and more diverse datasets to improve the generalization and robustness of traditional food classification models.

4. CONCLUSION

This study presented a comprehensive comparative analysis of two prominent deep learning architectures EfficientNet-B0 and Vision Transformer (ViT-B/16) for classifying traditional Indonesian cakes. Using a diverse and culturally rich image dataset, we demonstrated that both models achieved high classification performance, with ViT-B/16 slightly outperforming EfficientNet-B0 in terms of overall accuracy and consistency across visually similar classes.

The findings indicate that ViT-B/16's ability to model long-range dependencies and capture global visual features gives it a distinct advantage in handling fine-grained classification tasks where local texture variations may not be sufficient for accurate discrimination. Nonetheless, EfficientNet-B0 remains a competitive and computationally efficient baseline, particularly suitable for deployment in environments with limited resources.

In practical terms, this research underscores the effectiveness of Vision Transformers for traditional food classification and highlights the ongoing relevance of CNNs in real-world applications. Future work could explore more extensive and diverse datasets, domain-specific augmentation strategies, or hybrid architectures that leverage the complementary strengths of CNNs and transformers. Ultimately, integrating advanced AI models into culinary and cultural heritage preservation holds significant promise, and this study offers valuable insights for researchers and developers working in that direction.

5. ACKNOWLEDGEMENT

This research was supported by the authors' respective institutions. The authors also acknowledge the creators of the "Kue Indonesia" dataset for providing a valuable resource for this study.

6. AUTHOR CONTRIBUTION STATEMENT

DT, AAS, VCM, and UAO contributed to the conceptualization and methodology of the study. DT was responsible for the formal analysis, investigation, and writing of the original draft. All authors contributed to the review and editing of the manuscript.

AUTHOR INFORMATION

Corresponding Authors

Dedi Trisnawarman, Universitas Tarumanagara, Jakarta, Indonesia
https://orcid.org/0000-0002-9994-249X
Email: dedit@fti.untar.ac.id

Authors

Adolf Asih Supriyanton, Politeknik Enjinering Indorama, Purwakarta, Indonesia
https://orcid.org/0009-0009-7945-8754
Email: adolf@pei.ac.id

Viny Christanti Mawardi, Universitas Tarumanagara, Jakarta, Indonesia
https://orcid.org/0000-0001-6260-406X
Email: viny@untar.ac.id

Ugochi A Okengwu, Department of Computer Science, University of Port Harcourt, Nigeria
https://orcid.org/0000-0003-1695-0660
Email: ugochi.okengwu@uniport.edu.ng

REFFERENCE

Alba-Martínez, J., Bononad-Olmo, A., Igual, M., Cunha, L. M., Martínez-Monzó, J., & García-Segovia, P. (2022). Role of Visual Assessment of High-Quality Cakes in Emotional Response of Consumers. *Foods*, *11*(10), 1–15. https://doi.org/10.3390/foods11101412

Alruwaili, M., & Mohamed, M. (2025). An Integrated Deep Learning Model with EfficientNet and ResNet for Accurate Multi-Class Skin Disease Classification. *Diagnostics*, *15*(5), 1–18. https://doi.org/10.3390/diagnostics15050551

Banerjee, S., Palsani, D., & Mondal, A. C. (2024). Nutritional Content Detection Using Vision Transformers- An Intelligent Approach. *International Journal of Innovative Research in Engineering and Management*, *11*(6), 21–27. https://doi.org/10.55524/ijirem.2024.11.6.3

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Modi, K., & Ghayvat, H. (2021). Cnn variants for computer vision: History, architecture, application, challenges and future scope. *Electronics (Switzerland)*, *10*(20), 1–28. https://doi.org/10.3390/electronics10202470

Boyd, L., Nnamoko, N., & Lopes, R. (2024). Fine-Grained Food Image Recognition: A Study on Optimising Convolutional Neural Networks for Improved Performance. *Journal of Imaging*, *10*(6), 1–25. https://doi.org/10.3390/jimaging10060126

Chen, J., Ma, X., Li, S., Ma, S., Zhang, Z., & Ma, X. (2024). A Hybrid Parallel Computing

Architecture Based on CNN and Transformer for Music Genre Classification. *Electronics (Switzerland)*, *13*(16), 1–13. https://doi.org/10.3390/electronics13163313

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale. *ICLR 2021 - 9th International Conference on Learning Representations*, *2*(1), 1–22. https://doi.org/10.48550/arXiv.2010.11929

Hjalager, A.-M. (2022). Digital Food and the Innovation of Gastronomic Tourism. *Journal of Gastronomy and Tourism*, *7*(1), 35–49. https://doi.org/10.3727/216929722x16354101932186

Isinkaye, F. O., Olusanya, M. O., & Singh, P. K. (2024). Deep learning and content-based filtering techniques for improving plant disease identification and treatment recommendations: A comprehensive review. *Heliyon*, *10*(9), e29583. https://doi.org/10.1016/j.heliyon.2024.e29583

Karlita, T., Afif, B. P., & Prasetyaningrum, I. (2022). Indonesian Traditional Cake Classification Using Convolutional Neural Networks. *Proceedings of the International Conference on Applied Science and Technology on Social Science 2021 (ICAST-SS 2021)*, *647*, 924–929. https://doi.org/10.2991/assehr.k.220301.153

Liu, D., Zuo, E., Wang, D., He, L., Dong, L., & Lu, X. (2025). Deep Learning in Food Image Recognition : A Comprehensive Review. *Applied Sciences*, *15*(14), 1–18. https://doi.org/10.3390/app15147626

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., & Ilono, P. (2025). Deep Convolutional Neural Networks in Medical Image Analysis: A Review. *Information (Switzerland)*, *16*(3), 1–28. https://doi.org/10.3390/info16030195

Mingxing Tan, Q. V. Le. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks Mingxing. *International Conference on Machine Learning*, *1*(5). https://doi.org/10.48550/arXiv.1905.11946

Nfor, K. A., Theodore Armand, T. P., Ismaylovna, K. P., Joo, M. Il, & Kim, H. C. (2025). An Explainable CNN and Vision Transformer-Based Approach for Real-Time Food Recognition. *Nutrients* , *17*(2), 1–24. https://doi.org/10.3390/nu17020362

Sampath, V., Maurtua, I., Aguilar Martín, J. J., & Gutierrez, A. (2021). A survey on generative adversarial networks for imbalance problems in computer vision tasks. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00414-0

Sari, R. P., & Chandra, A. Y. (2025). Analisis Perbandingan Akurasi Model EfficientNetB0 dan Vision Transformer Dalam Klasifikasi Citra Motif Batik Giriloyo. *Building of Informatics, Technology and Science*, *7*(1), 252–263. https://doi.org/10.47065/bits.v7i1.7343

Sikdar, A., Liu, Y., Kedarisetty, S., Zhao, Y., Ahmed, A., & Behera, A. (2025). Interweaving Insights: High-Order Feature Interaction for Fine-Grained Visual Recognition. *International Journal of Computer Vision*, *133*(4), 1755–1779. https://doi.org/10.1007/s11263-024-02260-y

Suanpang, P., & Pothipassa, P. (2024). Integrating Generative AI and IoT for Sustainable Smart Tourism Destinations. *Sustainability (Switzerland)*, *16*(17), 1–34. https://doi.org/10.3390/su16177435

Taufiqurrahman, Sari, I. C., & Manurung, M. K. (2024). Integrasi Model Deep Learning Efficientnet-B0 Untuk Deteksi Penyakit Daun Tomat Pada Aplikasi Seluler Berbasis Flutter. *Djtechno: Jurnal Teknologi Informasi*, *5*(2), 332–346. https://doi.org/10.46576/djtechno.v5i2.4651

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *Proceedings of Machine Learning Research*, 1–22. https://doi.org/10.48550/arXiv.2012.12877

Wang, Y., & Wang, Z. (2019). A survey of recent work on fine-grained image classification techniques. *Journal of Visual Communication and Image Representation*, *59*, 210–214. https://doi.org/10.1016/j.jvcir.2018.12.049

Zhang, Y., Deng, L., Zhu, H., Wang, W., Ren, Z., Zhou, Q., Lu, S., Sun, S., Zhu, Z., Gorriz, J. M., & Wang, S. (2023). Deep learning in food category recognition. *Information Fusion*, *98*(March), 101859. https://doi.org/10.1016/j.inffus.2023.101859