ISSN 2355-0082

CNN-Transformer Fusion for Indonesian Traditional Cake Recognition: An EfficientNet-ViT Approach with Grad-CAM Explainability

Tasya Yustira Agustian[1], **Aswan Supriyadi Sunge**[2]
[1]Department of Informatics Engineering, Nusa Putra University, Sukabumi, Indonesia
[2]Department of Computer Science, Pelita Bangsa University, Bekasi, Indonesia
[1]tasya.yustira@nusaputra.ac.id, [2]aswan.sunge@pelitabangsa.ac.id

Accepted 03 June 2026
Approved 29 June 2026

***Abstract*— Visual recognition of Indonesian traditional confectionery is an underexplored problem in deep learning research, partly due to high inter-class visual ambiguity and the scarcity of well-curated local food benchmarks. We address this gap by fusing EfficientNet with a Vision Transformer (ViT) encoder into a unified classification network. The rationale for this pairing is straightforward: EfficientNet's compound-scaled convolutional stack efficiently encodes low and mid-level texture cues, while the ViT's self-attention layers then relate those cues across distant image regions-a capability that convolution alone cannot replicate. Post-hoc explainability is provided through Grad-CAM, which produces class-discriminative spatial maps confirming that activations concentrate on cake surfaces rather than background. We train and evaluate on a publicly available eight-class Kaggle corpus of 1,833 images, applying a two-stage fine-tuning regimen totaling 25 epochs. The resulting system attains 94.37% accuracy, 94.57% precision, 94.37% recall, and 94.31% F1 on the reserved test split. Beyond the metrics, the Grad-CAM evidence suggests the network learns genuinely food-discriminative features, lending credibility to deployment in culinary archiving and nutrition-monitoring applications.**

***Index Terms*— deep learning; EfficientNet; food image classification; Grad-CAM; Vision Transformer**

I. INTRODUCTION

The intersection of deep learning and culinary heritage has attracted growing research attention as institutions seek scalable ways to catalogue and preserve regional food cultures. Recognition systems that can automatically identify food items from photographs can support digital menu generation, dietary tracking, and cultural documentation. However, food images are notoriously difficult to classify because the same dish may look strikingly different depending on the lighting, plating convention, and camera perspective, whereas distinct dishes can appear deceptively similar. These difficulties compound when the target domain is underrepresented in standard benchmarks, a situation faced by Indonesian traditional cakes, which receive far less systematic study than popular international food [1].

Two architectural advances make this problem more tractable. First, EfficientNet demonstrated that scaling a network along depth, width, and resolution axes simultaneously-rather than along any single axis-yields accuracy improvements without proportional parameter growth [2]. Second, the ViT established that treating an image as a sequence of linearly embedded patches and applying multi-head self-attention can model global feature correlations that CNN receptive fields miss [3]. A hybrid that routes feature maps from EfficientNet into a ViT encoder therefore inherits both advantages-local discriminative power and global contextual awareness [2], [4].

Alongside predictive performance, this work integrates Gradient-weighted Class Activation Mapping (GradCAM) to make predictions auditable. Rather than trusting accuracy figures alone, we verify that the model attends to the food object itself-not to incidental background cues-before drawing conclusions about its suitability for real-world use [5], [6].

Three gaps motivate this work specifically: (i) hybrid CNN-Transformer classifiers have rarely been benchmarked on traditional food datasets; (ii) explainability analysis via Grad-CAM in such hybrid settings is nearly absent from the food-recognition literature; and (iii) Indonesian confectionery classification remains a relatively open problem compared with globally popular food benchmarks [7]. The contributions of this paper directly target all three gaps.

II. LITERATURE REVIEW

A. Food Image Recognition with Deep Learning

Convolutional networks have served as the workhouse of food image recognition for over a decade, progressively delivering higher accuracy as architectures matured from shallow networks to VGG, ResNet, and beyond [4]. Prior work has consistently demonstrated that CNN-based models can learn discriminative representations are capturing texture gradients, contour patterns, and color distributions which are directly from food imagery without handcrafted feature engineering [8]. However, the majority of these investigations focus on internationally recognized food categories, and localized cuisines such as Indonesian traditional cakes remain underserved. Dataset scarcity continues to constrain the development of robust classifiers for regional food varieties [9].

B. Hybrid Architecture: CNN-Transformer Models

Purely convolutional classifiers excel at encoding local spatial patterns but are bounded by the receptive field of their filters. EfficientNet exemplifies this trend through compound scaling that simultaneously optimizes depth, width, and resolution, surpassing conventional CNNs in accuracy-to-parameter ratio [2]. ViT has equally gained traction by segmenting images into fixed-size patches and applying transformer-style self-attention to capture global semantic relationships across the entire image [3]. Studies integrating CNN local feature extraction with transformer global context modelling consistently report performance improvements over either paradigm individually [10]. However, systematic evaluation of such hybrid models on traditional food datasets remains sparse.

C. Explainable AI with Grad-CAM

As model complexity increases, interpretability has become an essential counterpart to raw predictive performance. Grad-CAM addresses this need by computing class-specific gradient signals flowing into the final convolutional feature maps and aggregating them into a spatial activation map that highlights discriminative image regions [11]. This visualization allows researchers to verify that a model focuses on semantically relevant areas rather than spurious background patterns, supporting more rigorous validation of classification behavior [5]. Applications of Grad-CAM within hybrid CNN-Transformer pipelines targeting food image recognition, however, remain limited in the existing literature.

III. METHOD

A. System overview

The study implements a Hybrid EfficientNet-ViT model for Indonesian cake image classification. EfficientNet extracts local visual representations while ViT captures long-range inter-feature dependencies, together enabling robust recognition across visually diverse food categories. Fig.1 schematizes the end-to-end pipeline proceeds from raw image input through preprocessing and augmentation, EfficientNet-based feature extraction, ViT-based global relational modelling, softmax-based classification, and Grad-CAM visualization.

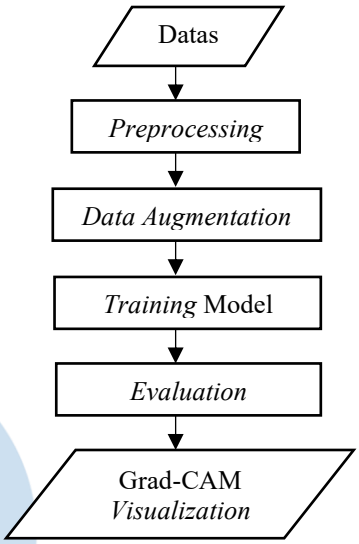

Fig. 1. Research design pipeline of the proposed model.

B. Dataset

The dataset was sourced from the Kaggle platform (https://www.kaggle.com/datasets/ilhamfp31/kue-indonesia/data) and serves as the experimental corpus. It consists of approximately 1,846 JPEG images (~303 MB) organized into eight traditional Indonesian cake categories. Each class exhibits natural variation in appearance, color, and lighting, providing conditions suitable for training deep learning classifiers. Table I details the image distribution per category.

TABLE I. CORPUS STATISTICS

Category	Number of Images
Dadar Gulung	±200
Kastengel	±200
Klepon	±200
Lapis	±200
Lumpur	±200
Putri Salju	±200
Risoles	±200
Serabi	±200

Source: Kaggle – Kue Indonesia (2024)

Fig. 2 Representative images from each cake category in the dataset.

C. Data Preprocessing and Augmentation

Every image is resized to 224×224 pixels before entering the network, prior to training to satisfy EfficientNet's input requirements, followed by pixel intensity normalization to promote stable gradient propagation [2], [4]. The spatial mapping from original image I(x,y) to normalized output I'(x',y') is expressed as shown in Eq. (1).

$$I'(x', y') = I(T(x, y)) \quad (1)$$

To expand training diversity and reduce overfitting risk, online augmentation was applied throughout training. The selected transformations synthesize plausible visual variants while preserving each class's defining characteristics, an approach with well-documented benefits for generalization on constrained datasets [12]. Table II summarizes the augmentation configuration.

TABLE II. DATA AUGMENTATION TECHNIQUES

No	Technique	Description
1	Rotation	Angular rotation by a random degree
2	Horizontal Flip	Mirror reflection along the horizontal axis
3	Zoom	Randomized scale enlargement or reduction
4	Width Shift	Lateral translation along the image width
5	Height Shift	Vertical translation along the image height

D. Model Architecture

The proposed architecture unifies EfficientNet with ViT into a single classification backbone. EfficientNet fulfills the role of a convolutional feature extractor that encodes local visual patterns, while ViT subsequently models global inter-token relationships through self-attention [2], [3]. A fully connected softmax classifier projects the resulting embedding to per-class probabilities, with Grad-CAM applied post-hoc to produce interpretable spatial activation maps [11]. Fig. 3 illustrates the complete architecture.

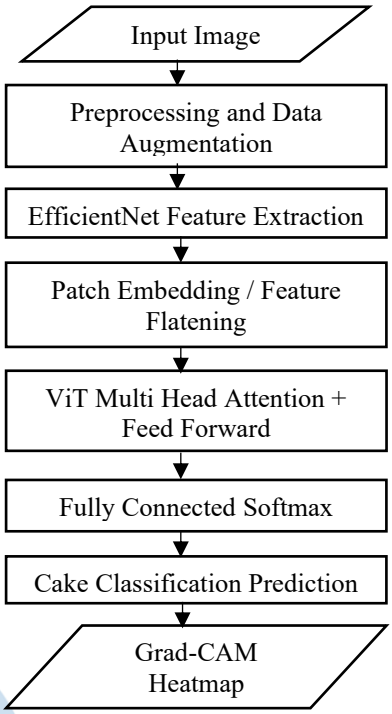

Fig. 3 Architecture of the proposed Hybrid EfficientNet-ViT model

EfficientNet-B0 is deployed as the initial visual encoder. Its compound scaling strategy jointly optimizes network depth, width, and resolution, achieving strong feature capacity at reduced computational cost compared with equivalently accurate conventional CNNs [2]. Hierarchical convolutional processing encodes key visual attributes from each cake image, including surface texture, geometric contour, and dominant color. The convolution operation generating feature map F at spatial position (i, j) is given by Eq. (2).

$$F(i,j) = \sum_m \sum_n I\,(i + m, j + n).K(m,n) \quad (2)$$

The output feature vector from EfficientNet-B0 is then fed into the ViT encoder. In this implementation, the ViT component follows the base configuration (ViT-Base) with the following specifications, patch size of 16×16, embedding dimension of 768, 12 transformer layers, and 12 attention heads. The feature vector is projected into the ViT embedding space before being processed by the transformer encoder. The ViT encoder transforms each feature token into Query (Q), Key (K), and Value (V) projections and computes inter-token affinities through scaled dot-product attention [3] as shown in Eq. (3).

$$Attention\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Finally, a fully connected layer maps the encoder output to a class score vector. Softmax normalization

converts these scores into a valid probability distribution [13] as shown in Eq. (4).

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \quad (4)$$

E. Training Protocol

Training proceeded in two sequential phases to maximize the utility of ImageNet pre-training. Phase 1 froze the EfficientNet-B0 backbone and optimized only the classification head for 15 epochs at a learning rate of 0.001, enabling stable initial feature adaptation. Phase 2 then unlocked the final 30 backbone layers and continued training for up to 30 epochs at a reduced learning rate of 0.00001, preserving pre-trained low-level representations while adapting higher-level features to the target domain [2]. This staged approach has demonstrated effectiveness for constrained datasets by exploiting ImageNet priors prior to domain-specific fine-tuning [14]. Table III lists all training hyperparameters.

The Adam optimizer was chosen for its per-parameter adaptive learning rate adjustment, which facilitates convergence in deep network optimization [15]. Categorical cross-entropy was used as the loss function, as defined in Eq. (5) [16].

$$L = -\sum_{i=1}^{n} y_i \log \hat{y}_i \quad (5)$$

TABLE III. TRAINING CONFIGURATION

Parameter	Value
Architecture	Hybrid EfficientNet-ViT
Input Image Size	224×224 pixels
Phase 1 - Epochs	15
Phase 2 - Epochs	30
Batch Size	32
Phase 1 - Learning Rate	0.001
Phase 2 - Learning Rate	0.00001
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Output Activation	Softmax
Dataset Split	80% / 10% / 10%

F. Evaluation Protocol

Four complementary metrics were used to evaluate classification performance across all eight categories: accuracy in Eq. (6), precision in Eq. (7), recall Eq. (8), and F1-score in Eq. (9) [17], [18].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1 = 2\,X\,\frac{Precision\ x\ Recall}{Precision + Recall} \quad (9)$$

G. Grad-CAM Explainability

To provide visual interpretability, Grad-CAM was applied to generate class-discriminative activation maps that localize image regions with the greatest influence on each prediction [11], [19]. Importance weights for each feature map channel are derived via global average pooling of class-specific gradient signals as shown in Eq. (10) and Eq. (11).

$$\propto_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \quad (10)$$

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \propto_k^c A^k\right) \quad (11)$$

These weights are combined with their corresponding feature activations and passed through ReLU to generate the final heatmap.

IV. RESULT AND DISCUSSION

A. Dataset Partitioning and Training Results

The full dataset was split into training (80%), validation (10%), and test (10%) subsets to ensure evaluation on previously unseen data. Table IV details the resulting partition sizes.

TABLE IV. DATA PARTITIONING

Subset	Proportion	Images
Training	80%	1,514
Validation	10%	159
Test	10%	160

Training terminated at epoch 25 when early stopping was triggered, with the peak validation accuracy of 0.9434 recorded at epoch 9 of the first phase. Fig. 4 plots the accuracy and loss trajectories across both training phases.

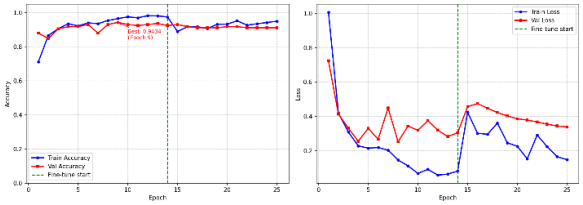

Fig. 4 Accuracy and loss progression across training epochs

B. *Preprocessing and Augmentation Results*

All images were successfully resized to 224×224 pixels and normalized before training. The applied augmentation operations included rotation, horizontal flipping, zoom, and directional shift. These operations produced diverse training variants without distorting the defining visual identity of each cake class. Fig. 5 shows representative augmented samples.

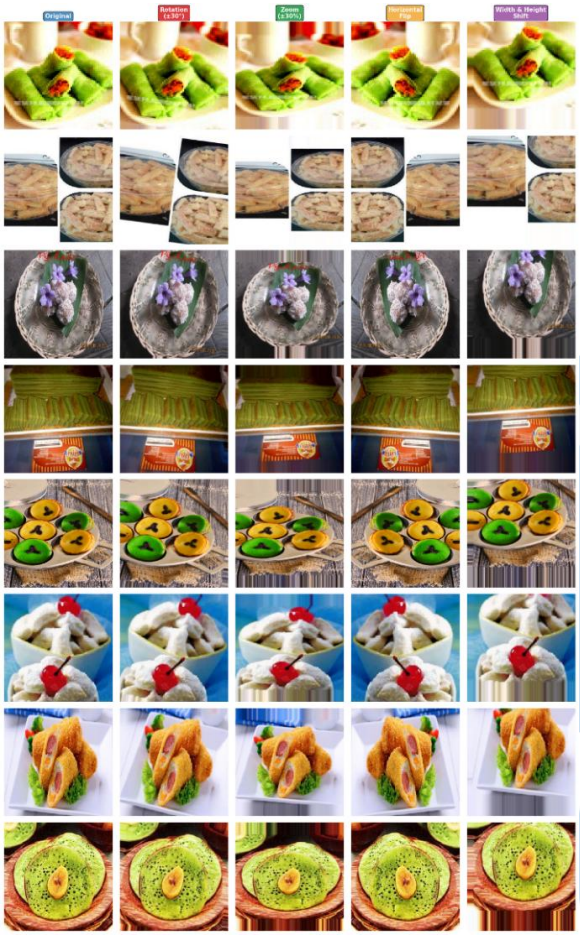

Fig. 5 Examples of augmented training images.

C. *Overall Evaluation Results*

Evaluation on the 160-image test set yielded an overall accuracy of 94.37%, corresponding to 151 correctly classified images. Table V summarizes all aggregate performance metrics.

TABLE V. AGGREGATE TEST METRICS

Metric	Value
Accuracy	94.37%
Precision	94.57%
Recall	94.37%
F1-score	94.31%

D. *Class-Level Breakdown*

Per-class results are reported in Table VI. Kastengel, klepon, and lapis achieved perfect recall (1.0000), indicating complete retrieval of all test samples in those classes. Putri salju and risoles attained perfect precision (1.0000), signifying an absence of false positives. Dadar gulung recorded the lowest recall at 0.8000, likely owing to visual similarity with adjacent categories, including overlapping surface texture and color profile.

TABLE VI. PER-CLASS RESULTS

Class	Prec.	Recall	F1	Supp.
Dadar Gulung	0.9412	0.8000	0.8649	20
Kastengel	0.9565	1.0000	0.9778	22
Klepon	0.9524	1.0000	0.9756	20
Lapis	0.9091	1.0000	0.9524	20
Lumpur	0.9474	0.9000	0.9231	20
Putri Salju	1.0000	0.9444	0.9714	18
Risoles	1.0000	0.9500	0.9744	20
Serabi	0.8636	0.9500	0.9048	20
Wtd. Avg	0.9457	0.9437	0.9431	160

The evaluation results demonstrate that the Hybrid EfficientNet-ViT model effectively classifies Indonesian traditional cake images. Per-class performance is illustrated in Fig. 6, which presents the precision, recall, and F1-score for each class.

As shown in the figure, the Kastengel, Klepon, and Lapis classes achieved perfect recall scores of 1.0000, indicating that all images in these categories were correctly classified. The Putri Salju and Risoles classes also exhibited excellent performance with perfect precision scores of 1.0000. In contrast, the Dadar Gulung class recorded the lowest recall at 0.8000, suggesting some misclassifications with other categories.

The overall prediction distribution is visualized in the confusion matrix presented in Fig. 7. Most classification errors occurred in the Dadar Gulung class, with one image each misclassified as Klepon, Lapis, Lumpur, and Serabi. Additionally, two images from the Lumpur class were misclassified as Kastengel and Lapis. Overall, the dominant diagonal values in the confusion matrix indicate strong classification performance across most traditional cake categories.

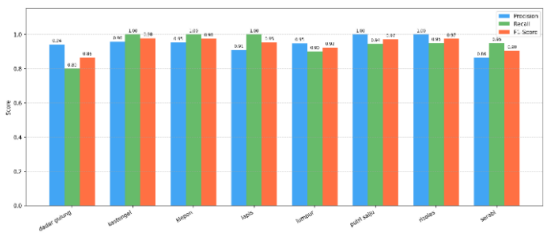

Fig. 6. Per-class precision, recall, and F1-score

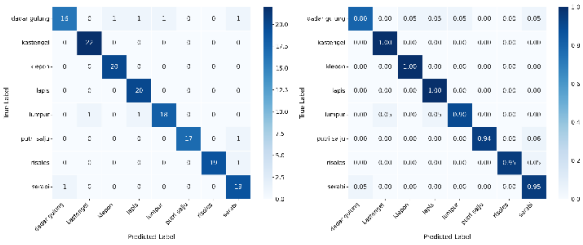

Fig. 7. Confusion matrix of test set predictions

E. Grad-CAM Analysis

Grad-CAM heatmaps were generated for all eight categories to validate the spatial reasoning of the classifier. Fig. 8 presents three-column panels for each class: (i) the original image, (ii) the standalone activation map, and (iii) the overlay. Across all categories, the heatmaps consistently localize attention to the cake object itself rather than surrounding background regions. This spatial alignment between model attention and object location confirms that the network has acquired class-discriminative feature representations and that its predictions are visually grounded.

F. Discussion

The 94.37% test accuracy achieved by the proposed hybrid model is competitive with recent works on fine-grained food classification using single-paradigm architectures on datasets of similar scale. The model also demonstrates the ability to focus on semantically relevant regions, as evidenced by the Grad-CAM visualizations.

Dadar gulung's lower recall (0.80) indicates a need for targeted improvement in future work. The confusion pattern, which is distributed across four different classes, suggests that this category does not exhibit strong visual overlap with any single class but rather lacks a well-defined cluster in the feature space. Future improvements could involve class-specific data augmentation or rebalancing strategies to enhance discrimination for this class.

The Grad-CAM analysis serves as an important validation tool. Correctly classified images generally produce compact activation maps centered on the food object, whereas misclassified samples tend to exhibit more diffuse or background-focused activations. This behavior provides useful insight into model uncertainty and supports the practical value of Grad-CAM beyond its role as an explainability tool.

V. LIMITATIONS

One limitation of this study is the absence of direct experimental comparisons with EfficientNet-only and ViT-only baselines, making it difficult to isolate the contribution of each architectural component. Although additional baseline comparisons are provided in Appendix A, future work will focus on more comprehensive ablation studies under a unified experimental setup. Another limitation is the use of a single train-validation-test split for evaluation. While commonly practiced, this approach may not fully demonstrate the model's robustness. Future research will therefore incorporate repeated experiments or k-fold cross-validation to better assess the generalizability and stability of the proposed approach.

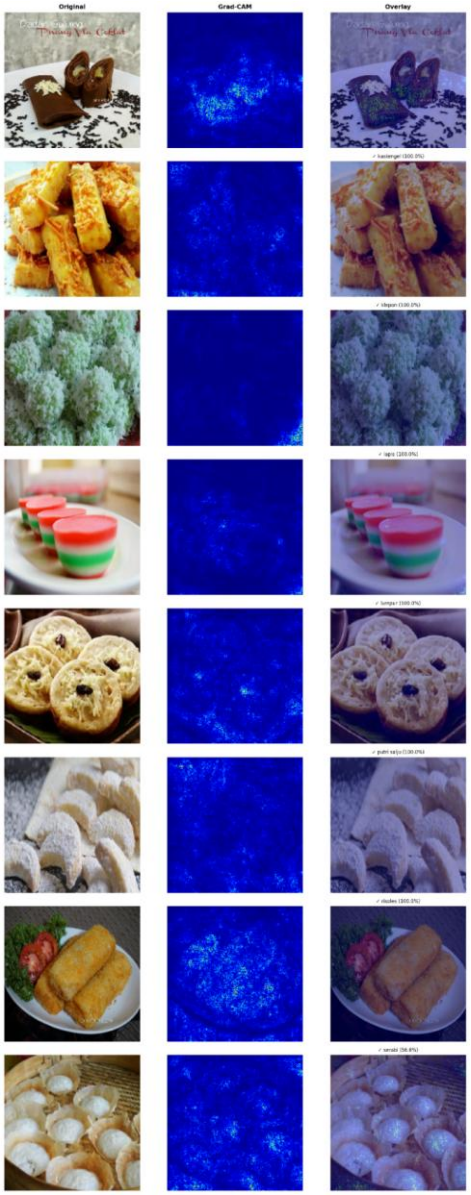

Fig. 8 Grad-CAM activation maps for all eight cake categories.

VI. CONCLUSION

We presented a recognition system for Indonesian traditional cake photography that fuses EfficientNet's convolutional feature extraction with a ViT encoder's global self-attention, trained under a warm-start two-stage schedule and audited through Grad-CAM. Across an eight-class, 1833 image benchmark, the system achieves 94.37% accuracy, 94.57% precision, 94.37% recall, and 94.31% of F1. Grad-CAM heatmaps confirm that high activations cluster on cake

surfaces rather than background content, supporting the trustworthiness of the predictions for downstream applications.

Several directions are open for future investigations. Expanding coverage-both in number of classes and images per class-would test the scalability of the fusion design. Multi-scale attention, guided by the confusion analysis reported here, could specifically improve discrimination among visually ambiguous categories. Finally, knowledge-distillation into a compact mobile-friendly model would enable on-device inference for culinary documentation applications.

APPENDIX A

Additional Baseline Comparison

Additional baseline experiments were conducted to provide a clearer understanding of the proposed hybrid model's performance relative to several well-known architectures. All models were evaluated using the same dataset split and training protocol. The results are presented below as supplementary material.

TABLE A1. PERFORMANCE COMPARISON WITH BASELINE MODELS

Model	Accuracy	Precision	Recall	F1-Score
MobileNet V2	95.62	95.73	95.62	95.59
EfficientNet_B0	98.12	98.18	98.12	98.12
ResNet50	96.88	96.99	96.88	96.89
ViT_B16	85.62	86.28	85.62	85.65
Hybrid EfficientNet-ViT	95.00	95.23	95.00	95.01

ACKNOWLEDGEMENT

The authors thank Universitas Nusa Putra and Universitas Pelita Bangsa for institutional support. Appreciation is also extended to the Kaggle community for providing the open-access Kue Indonesia dataset that underpinned this study.

REFERENCES

[1] D. Koteshwar Rao, K. Sahiti, J. Vamshi, K.Uday Kiran, and T. Srikar, "Deep Learning-Based Food Image Classification Using Enhanced CNN Architecture and Data Augmentation Techniques," *Int. J. Comput. Eng. Res. Trends*, vol. 11, no. 1s SE-Research Articles, pp. 16–22, Dec. 2024, doi: 10.22362/ijcert/2024/v11/i1/v11i1s03.

[2] M. Tan and Q. V Le, "EfficientNetV2: Smaller Models and Faster Training," *CoRR*, vol. abs/2104.00298, 2021, [Online]. Available: https://arxiv.org/abs/2104.00298

[3] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *CoRR*, vol. abs/2010.11929, 2020, [Online]. Available: https://arxiv.org/abs/2010.11929

[4] L. Bu, C. Hu, and X. Zhang, "Recognition of food images based on transfer learning and ensemble learning," *PLoS One*, vol. 19, no. 1, p. e0296789, Jan. 2024, doi: 10.1371/journal.pone.0296789.

[5] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks," *CoRR*, vol. abs/1710.11063, 2017, [Online]. Available: http://arxiv.org/abs/1710.11063

[6] G. Deng, D. Wu, and W. Chen, "Attention Guided Food Recognition via Multi-Stage Local Feature Fusion," 2024. doi: 10.32604/cmc.2024.052174.

[7] D. Liu, E. Zuo, D. Wang, L. He, L. Dong, and X. Lu, "Deep Learning in Food Image Recognition: A Comprehensive Review," 2025. doi: 10.3390/app15147626.

[8] Z. Zhu and Y. Dai, "A New CNN-Based Single-Ingredient Classification Model and Its Application in Food Image Segmentation," 2023. doi: 10.3390/jimaging9100205.

[9] B. N. Jagadesh *et al.*, "Enhancing food recognition accuracy using hybrid transformer models and image preprocessing techniques," *Sci. Rep.*, vol. 15, no. 1, p. 5591, 2025, doi: 10.1038/s41598-025-90244-4.

[10] J. Yoo, T. Kim, S. Lee, S. H. Kim, H. Lee, and T. H. Kim, "Enriched CNN-Transformer Feature Aggregation Networks for Super-Resolution," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 4956–4965.

[11] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization," *CoRR*, vol. abs/1610.02391, 2016, [Online]. Available: http://arxiv.org/abs/1610.02391

[12] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019, doi: 10.1186/s40537-019-0197-0.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[14] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *CoRR*, vol. abs/1411.1792, 2014, [Online]. Available: http://arxiv.org/abs/1411.1792

[15] D. S. Brar, B. Singh, and V. Nanda, "Application of deep learning and explainable artificial intelligence (XAI) for detecting red chilli powder adulteration," *J. Food Compos. Anal.*, vol. 146, p. 107947, 2025, doi: https://doi.org/10.1016/j.jfca.2025.107947.

[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *CoRR*, vol. abs/1611.03530, 2016, [Online]. Available: http://arxiv.org/abs/1611.03530

[17] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: https://doi.org/10.1016/j.ipm.2009.03.002.

[18] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *CoRR*, vol. abs/2010.16061, 2020, [Online]. Available: https://arxiv.org/abs/2010.16061

[19] S. Srinivas and F. Fleuret, "Full-Jacobian Representation of Neural Networks," *CoRR*, vol. abs/1905.00780, 2019, [Online]. Available: http://arxiv.org/abs/1905.00780