JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 11, No. 3, SEPTEMBER, 2026, Pp. 396 – 407
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Improving Multi-Class Classification Performance with Multi-Scale Feature Extraction CNNs on a Limited Indonesian Spices Dataset

Ahmad Zein Al Wafi [(1)], Alena Ghinna Kirana [(2)*]
Independent Researcher, Indonesia
e-mail : {ahmadzeinalwafi,alenaghinn}@outlook.com.
* Corresponding author.
This article was submitted on 29 March 2026, revised on 23 August 2026, accepted on 24 August 2026, and published on 25 September 2026.

Abstract

This study evaluates whether multi-scale feature extraction provides an advantage over single-scale CNN feature extraction under limited-data conditions. Using the Indonesian Spices Dataset as a controlled benchmarking testbed, which contains 31 distinct categories with only 210 images per class, we propose a multi-scale Convolutional Neural Network (CNN) architecture that captures features across varying spatial scales through three parallel branches with 3x3, 5x5, and 7x7 kernel sizes. The proposed model was evaluated against single-scale baselines using training and validation accuracy and loss, macro-averaged precision, recall, F1-score, and a confusion matrix. Experimental results show that the multi-scale approach provides the strongest learning capacity among the evaluated architectures, achieving the highest training accuracy (91.44%) and the lowest training loss (0.2713), while reaching a peak validation accuracy of 64.98% (62.90% final-epoch accuracy, 64.22% macro precision, and 62.59% macro F1-score) with a competitive training time of 3839.7s. The validation-accuracy advantage over the strongest single-scale baseline is modest; the clearer advantage is observed in training capacity and the broader validation metric profile. The late-epoch divergence between training and validation loss indicates that the additional representational capacity remains constrained by the limited number of training samples, highlighting the need for stronger regularization to improve generalization in scarce multi-class data.

Keywords: Multi-Scale CNN, Image Classification, Fine-Grained Recognition, Limited Data, Feature Extraction

Abstrak

Penelitian ini mengevaluasi apakah ekstraksi fitur multiskala memberikan keunggulan dibandingkan dengan ekstraksi fitur CNN skala tunggal pada kondisi data terbatas, dengan menggunakan Indonesian Spices Dataset sebagai testbed benchmarking terkontrol. Dataset ini terdiri dari 31 kategori dengan 210 citra per kelas. Arsitektur CNN multiskala memproses citra melalui tiga cabang paralel dengan ukuran kernel 3×3, 5×5, dan 7×7. Kinerja model dibandingkan dengan baseline skala tunggal menggunakan akurasi dan loss pelatihan serta validasi, macro-precision, macro-recall, macro-F1 score, dan confusion matrix. Hasil eksperimen menunjukkan bahwa pendekatan multiskala memberikan kapasitas pembelajaran terkuat di antara arsitektur yang dievaluasi, dengan akurasi pelatihan tertinggi (91.44%) dan loss pelatihan terendah (0.2713), serta akurasi validasi puncak 64.98% (62.90% akurasi akhir, 64.22% macro-precision, dan 62.59% macro-F1 score) dengan waktu pelatihan yang kompetitif (3839.7 detik). Keunggulan akurasi validasi terhadap baseline skala tunggal terbaik relatif kecil; keunggulan yang lebih jelas terlihat pada kapasitas pembelajaran dan profil metrik validasi secara keseluruhan. Divergensi loss pelatihan dan validasi pada epoch akhir menunjukkan bahwa kapasitas representasi tambahan belum sepenuhnya menghasilkan generalisasi yang stabil pada data terbatas, sehingga diperlukan regularisasi yang lebih kuat dan evaluasi pada test set terpisah pada penelitian selanjutnya.

Kata Kunci: CNN Multiskala, Klasifikasi Citra, Pengenalan Detail Tinggi, Data Terbatas, Ekstraksi Fitur

1. INTRODUCTION

Image recognition is a foundational task in computer technologies that enables machines to identify and interpret objects, patterns, and scenes within visual data (Zhang, 2022). Image recognition plays a crucial role in numerous real-world applications, as its ability to interpret and analyze images has vast implications across industries such as healthcare (Li et al., 2023), autonomous driving (Jun et al., 2023), robotics (Xiao et al., 2023), and security (Raharja et al., 2021). In medical imaging, for example, image classification helps detect diseases such as COVID-19 from radiographs and CT scans, supporting doctors in formulating clinical judgments as soon as possible (Ullah et al., 2023). Similarly, in autonomous vehicles, recognizing pedestrian behavior, random objects, and road types and their settings ensures safer navigation (Parekh et al., 2022). Beyond these domains, image recognition underpins applications in e-commerce, where product recommendations are assisted with image feature extraction (Indira et al., 2022), and in security, where it supports facial recognition for identity verification (Rukhiran et al., 2023).

Historically, image recognition has been a challenging task due to the high complexity of visual data, which is often noisy, unstructured, and subject to variation in illumination, scale, and orientation. Early attempts to solve image recognition problems relied on handcrafted features and traditional machine learning models (Ma et al., 2021). However, the emergence of deep learning, particularly through Convolutional Neural Networks (CNNs), has revolutionized this field, achieving state-of-the-art performance on various benchmark datasets such as ImageNet (Zoph et al., 2018).

The introduction of CNNs in the late 20th century marked a transformative leap, enabling automatic feature extraction and hierarchical learning from raw pixel data (Taye, 2023). This architecture has demonstrated remarkable success across numerous datasets and application domains (Tuggener et al., 2022). One of the major breakthroughs in CNNs came with the success of AlexNet, which significantly outperformed traditional machine learning models in the 2012 ImageNet competition (Krizhevsky et al., 2017). This success was further amplified by the development of deeper and more complex networks, such as GoogLeNet (Subba & Sunaniya, 2025) and ResNet (Wu et al., 2025), which introduced innovations like residual connections and more efficient architectures. CNNs have been instrumental in solving many visual recognition tasks, achieving remarkable levels of accuracy in detecting objects across large and varied datasets.

Despite their success, CNNs face significant challenges when applied to datasets with limited samples (D'souza et al., 2020). One major issue is the model's reliance on large amounts of labeled data for generalization (Younesi et al., 2024). In practical scenarios, acquiring enough labeled data can be difficult. This problem is especially acute in domains like medical imaging, where the availability of annotated data may be scarce due to the need for expert knowledge (Salehi et al., 2023). These challenges underscore the need for novel methods to improve the efficiency and robustness of CNNs, especially in situations where data is limited and the number of classes is very large.

Overcoming the limitations of CNNs in scenarios with limited data and a large number of classes is a persistent challenge in machine learning. Various strategies have been developed to address these issues, including data augmentation (Maharana et al., 2022), transfer learning (Maray et al., 2023), and semi-supervised learning (Fini et al., 2023). Data augmentation artificially increases the diversity of the training dataset through transformations like rotation, resizing, and flipping to improve performance (Anaya-Isaza & Mera-Jimenez, 2022). Transfer learning, which leverages pre-trained models trained on large-scale datasets, has been a popular method for improving performance on limited data (Xu et al., 2023). Similarly, semi-supervised learning combines a small amount of labeled data with a larger pool of unlabeled data, allowing the model to utilize unlabeled samples to support the learning process (Duarte & Berton, 2023).

Despite the progress made, these approaches have limitations in handling datasets with numerous classes, where the ability to discriminate between highly similar classes becomes critical. Prior studies indicate that transfer learning from ImageNet may provide limited improvements on certain classification tasks, suggesting that pretrained representations do not always transfer effectively to datasets requiring highly detailed category distinctions (Kornblith et al., 2019). Although data augmentation increases training diversity, excessive or insufficient augmentation may introduce noise or fail to expose models to sufficient variability, which can limit generalization (Kumar et al., 2024). Existing architectures also tend to focus on extracting features at a single scale, which may overlook essential details at varying levels of granularity. While multi-branch networks such as Inception architectures exist, they predominantly utilize asymmetric 1x1 bottleneck compressions and complex factorizations specifically tailored for large-scale benchmarks like ImageNet (Subba & Sunaniya, 2025). However, the present study is not intended to reproduce a full Inception-style architecture or establish state-of-the-art performance on the Indonesian Spices Dataset; instead, it isolates whether combining different receptive-field sizes provides an advantage over single-scale feature extraction under the same limited-data training conditions. To address this question, we propose a streamlined multi-scale CNN architecture that preserves uncompressed, direct parallel convolutional streams across distinct receptive fields (3x3, 5x5, and 7x7). This design enables direct coarse-to-fine spatial feature capture while keeping the comparison structurally controlled.

The primary objective of this paper is to determine whether multi-scale feature extraction provides an advantage over single-scale CNN feature extraction under a limited-data setting, using the Indonesian Spices Dataset as a controlled benchmarking testbed. The contributions of this work are: (1) proposing an uncompressed parallel multi-scale CNN tailored for low-sample fine-grained classification, (2) providing a transparent empirical analysis comparing single-scale and multi-scale feature dynamics, and (3) offering a comprehensive error analysis across 31 spice classes using macro-averaged evaluation metrics and confusion matrices.

2. METHODS

2.1 Dataset

The Indonesian Spices Dataset, sourced from Kaggle, is a balanced dataset comprising 31 classes, each containing 210 images. These classes represent a variety of Indonesian spices, many of which exhibit subtle visual differences, making them an ideal example of fine-grained classification. This fine-grained nature poses a significant challenge for traditional classification models, as distinguishing between visually similar classes often requires capturing intricate and nuanced features. In the context of deep learning, this dataset is considered small compared to standard computer vision benchmarks; while large-scale datasets such as ImageNet provide upwards of 1,000 images per class, the Indonesian Spices Dataset offers only 210 images per class (yielding approximately 168 training images per category under an 80/20 split). This low-sample regime presents a significant generalization challenge, making standard deep architectures particularly vulnerable to overfitting.

For image width, the mean is 616.69 pixels, with a median and mode of 500 pixels. The standard deviation of 523.48 and variance of 274,029.00 indicate considerable variation in image sizes. The high positive skewness (4.54) and kurtosis (25.46) indicate a strongly right-skewed distribution with heavy tails, suggesting the presence of several images with substantially larger widths than the majority of the dataset. Similarly, for image height, the mean is 543.72 pixels, with the median and mode also at 500 pixels. The standard deviation of 431.82 and variance of 186,464.78 indicate slightly lower variability compared to the width, while the skewness (4.91) and kurtosis (30.24) reveal a similarly right-skewed distribution with extreme values. A visual representation of the data distribution, shown in Figure 1, further supports this observation, highlighting the concentration of images around the mode and the presence of a long tail corresponding to larger image dimensions.

This dataset's balanced nature, coupled with its limited amount, challenging fine-grained classes, and variation in image sizes, makes it highly suitable for evaluating the effectiveness of multi-scale CNNs. The ability of multi-scale architectures to extract features at varying levels of granularity is particularly relevant for addressing the nuances of fine-grained classification, where small, subtle features may be crucial for accurate differentiation.

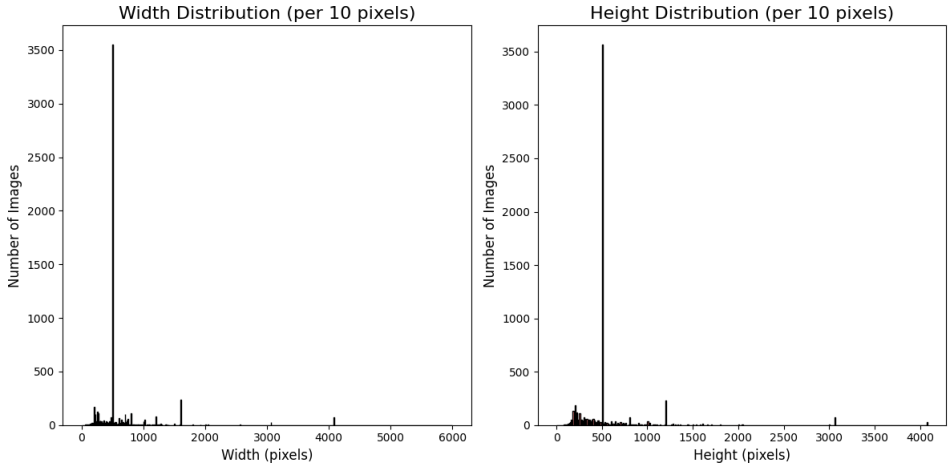

Figure 1 Width & Height Distribution

2.2 Data Preprocessing

The dataset was preprocessed to standardize input images and enhance the diversity of training data through augmentation techniques. All images were rescaled by dividing pixel values by 255, normalizing the intensity to a range of [0, 1]. This step ensures consistent input scaling, which is critical for optimizing the learning process.

To improve generalization and mitigate overfitting, several augmentation techniques were applied to the training data. These included random shear transformations (up to 20%) and small random rotations (up to 10 degrees) to account for variations in orientation. Brightness levels were adjusted within a range of 80% to 120% of the original intensity, simulating changes in lighting conditions. Random zooming (up to 20%) was also implemented to introduce variability in object size, and horizontal flipping was applied to capture mirrored features. These augmentations created a more diverse training set while preserving the original class labels.

To guarantee scientific validity and eliminate the risk of data leakage, data partitioning was strictly executed before any augmentation procedures. The raw dataset was split into an 80% training set (5,208 images) and a 20% validation set (1,302 images) using a fixed random seed (42). Data augmentation was applied strictly and dynamically only to the training generator pipeline during training. Crucially, the validation dataset was processed through a separate, deterministic generator without any geometric or photometric augmentations, and without shuffling (shuffle=False), ensuring unbiased and consistent metric evaluation. No separate isolated test set was used in this study; therefore, the reported classification metrics are validation-set metrics, and held-out test evaluation is identified as a limitation for future work.

2.3 Model Architecture

The design of the convolutional neural network (CNN) architectures in this study aims to evaluate and compare the performance of single-scale and multi-scale approaches for fine-grained image classification in datasets with limited samples and many classes. These challenges necessitate extracting detailed and varied features to distinguish between visually similar categories effectively. To this end, two architectural paradigms are explored: single-scale CNNs, which utilize

a uniform kernel size throughout the network, and multi-scale CNNs, which integrate multiple kernel sizes to capture features at different levels of granularity.

2.3.1 Single-Scale CNN

The single-scale convolutional neural network (CNN) architectures are designed to evaluate the performance of individual kernel sizes 3x3, 5x5, and 7x7 on the task of fine-grained classification with limited data and numerous classes. Each single-scale CNN employs a single convolutional kernel size throughout its convolutional layers, followed by a fully connected layer for classification. The network begins with an input layer accommodating images resized to 224×224×3, ensuring compatibility with the dataset's preprocessing pipeline. The convolutional layers are configured with progressively increasing numbers of filters (16, 32, and 64), with ReLU activation and "same" padding. Each convolutional layer is followed by max-pooling to reduce spatial dimensions and extract salient features. For instance, in the 7x7 kernel model, three convolutional layers of 7x7 kernels are used sequentially, while analogous configurations are adopted for the 5x5 and 3x3 models.

The flattened output of the final max-pooling layer serves as the feature representation of the input, which is passed to a dense layer with 31 output neurons (equal to the number of classes) and a softmax activation function for multi-class classification. This architecture allows for the direct comparison of different kernel sizes in capturing discriminative features. The single-scale models serve as baselines for understanding the efficacy of kernel sizes when applied independently. Their performance provides insights into how feature granularity impacts the classification of visually similar (fine-grained) objects under the constraints of limited data and numerous categories.

2.3.2 Multi-Scale CNN

The multi-scale CNN architecture integrates feature extraction from multiple kernel sizes—3x3, 5x5, and 7x7—within a unified framework. The motivation is to test whether complementary receptive fields provide an advantage over a single kernel size under identical training conditions. Direct 5x5 and 7x7 convolutions are retained rather than factorized into stacked 3x3 operations or replaced with dilated convolutions so that the experiment isolates the effect of explicitly combining distinct receptive-field sizes without simultaneously changing network depth or introducing additional architectural mechanisms.

The network begins with an input layer identical to the single-scale models. Three separate convolutional branches process the input data in parallel, each utilizing a distinct kernel size. Each branch follows a similar design, with ReLU activation, "same" padding, and max-pooling layers to downsample the feature maps. The number of filters in all branches increases progressively (16, 32, 64) to enhance the capacity of feature extraction. The outputs of the three branches are concatenated along the channel dimension, resulting in a multi-scale feature representation that captures information at varying levels of granularity. The concatenated features are flattened and passed through a dense layer with 31 output neurons and a softmax activation function for classification. The concatenated feature maps are flattened before the final classification layer, producing a 150,528-dimensional feature representation prior to the dense layer.

The three branches are intended to capture complementary fine, intermediate, and coarse spatial information. Because the study is a controlled architectural comparison, this multi-scale configuration is treated as a hypothesis to be tested rather than as an assumption of superior performance. Its effectiveness is therefore assessed directly against the corresponding single-scale 3x3, 5x5, and 7x7 baselines.

2.4 Training Configuration

The training process utilized a configuration optimized for efficient learning and robust evaluation. The categorical cross-entropy loss function was chosen for its effectiveness in multi-class

classification tasks, as it quantifies the divergence between the predicted class probabilities and the true labels across the dataset's 31 categories. The Adam optimization algorithm was employed with its default parameter values, including a learning rate of 0.001, β1=0.9, and β2=0.999. These defaults are well-established for providing a stable and adaptive optimization process, particularly beneficial for datasets with fine-grained distinctions and limited samples. Training was conducted for 30 epochs, a duration selected to balance learning progress and mitigate overfitting risks, especially critical for datasets with limited data.

2.5 Training Environment

The experiments were conducted using Google Colab, which provides access to a free GPU, specifically the Tesla T4, equipped with 15GB of VRAM and 12.7GB of system RAM. This setup offered an efficient environment for training deep learning models, particularly for computationally intensive tasks such as multi-scale CNNs. The Tesla T4 was chosen for its capability to accelerate model training, leveraging its hardware support for tensor operations, which is crucial for deep neural network computations.TensorFlow was utilized as the deep learning framework, providing a robust and flexible platform for constructing, training, and evaluating the CNN models.

3. RESULTS AND DISCUSSION

To determine whether multi-scale feature extraction provides an advantage over single-scale architectures, the models were assessed across training and validation accuracy and loss, macro-averaged precision, recall, F1-score, and computational efficiency. The consolidated results are presented in Table 1, while Figures 2–7 provide visual evidence for the training dynamics, validation behavior, runtime, and class-level errors.

Table 1 Comprehensive Performance Comparison of CNN Models

Model	Best Train Acc	Best Train Loss	Final Val Acc	Peak Val Acc	Macro Prec	Macro Recall	Macro F1	Total Time (s)
CNN3x3 Baseline	88.61%	0.3740	60.29%	60.29%	0.6012	0.6029	0.5988	3957.7s
CNN5x5 Baseline	90.44%	0.3020	62.29%	62.29%	0.6210	0.6229	0.6195	3792.5s
CNN7x7 Baseline	86.39%	0.4280	60.14%	60.14%	0.5990	0.6014	0.5972	3888.6s
Multi-Scale CNN (Proposed)	91.44%	0.2713	62.90%	64.98%	0.6422	0.6290	0.6259	3839.7s

3.1 Training Dynamics and Feature Learning Capacity

The training phase demonstrates the strongest learning capacity for the Multi-Scale CNN. As illustrated by the Comparison of Training Accuracy in Figure 2 and Comparison of Training Loss in Figure 3, the multi-scale architecture consistently outperformed all single-scale variants throughout the 30 epochs. It achieved the highest overall training accuracy of 91.44% and the lowest training loss of 0.2713. Among the single-scale baselines, the CNN 5x5 showed the strongest learning capability, while the CNN 7x7 recorded the lowest training accuracy (86.39%) and highest training loss (0.4280). These results indicate that combining receptive fields improves the model's capacity to fit discriminative training representations; however, this training advantage must be interpreted together with the validation behavior shown in Figures 4 and 5.

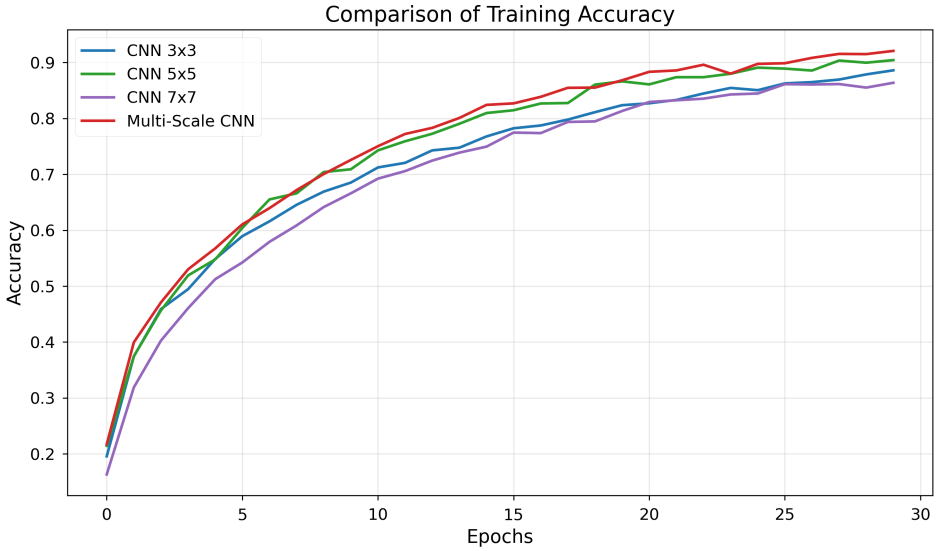

Figure 2 Comparison of Training Accuracy

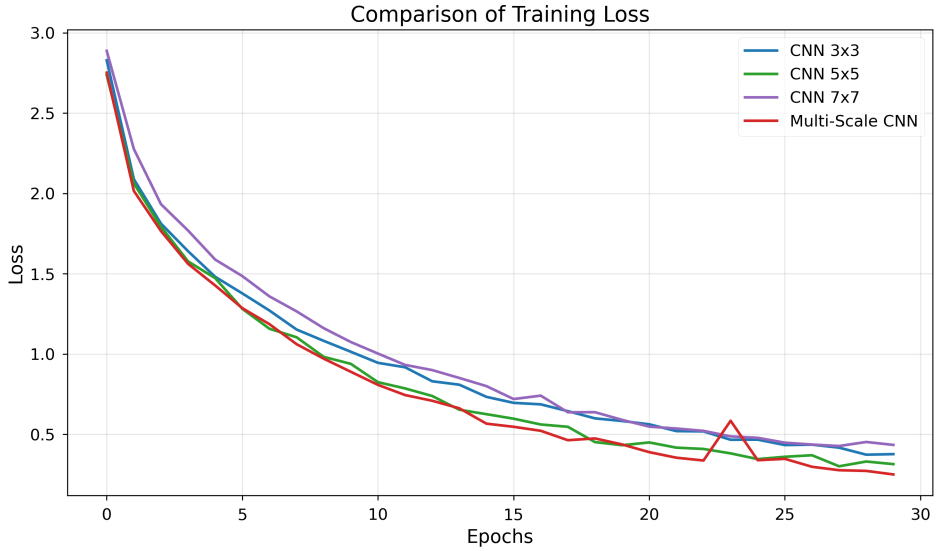

Figure 3 Comparison of Training Loss

3.2 Validation Performance and Overfitting Trends

While the Multi-Scale CNN dominated the training metrics, the validation metrics reveal the inherent generalization challenges of the limited dataset. As shown in the Comparison of Validation Accuracy in Figure 4 and Comparison of Validation Loss in Figure 5, validation accuracies for all models plateau relatively early. The Multi-Scale CNN attained a final validation accuracy of 62.90% with a peak validation accuracy of 64.98% at epoch 27, compared with 62.29% for the strongest single-scale baseline, CNN 5x5. Thus, the validation-accuracy advantage is modest rather than substantial. Crucially, the validation loss exhibits a distinct U-shaped curve across the models: it decreases steadily until approximately epoch 8–10 and then rises, signaling the onset of overfitting. The divergence between improving training accuracy and increasing validation loss indicates that the additional representational capacity does not fully translate into stable generalization when only approximately 168 training images are available per class.

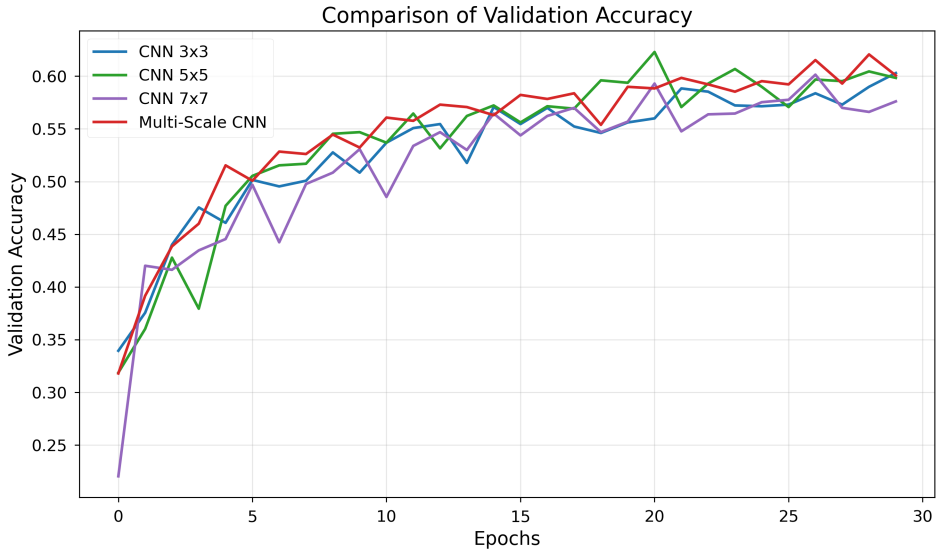

Figure 4 Comparison of Validation Accuracy

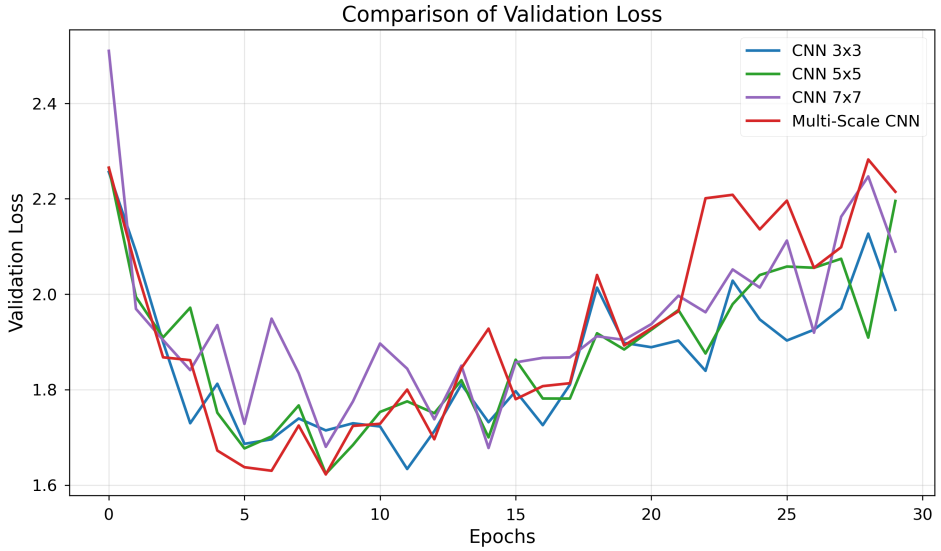

Figure 5 Comparison of Validation Loss

The training and validation curves also provide insight into the learning dynamics of the evaluated architectures. During the early and middle stages of training, the multi-scale CNN maintained a favorable learning trajectory, with training accuracy continuing to improve while the corresponding loss decreased. This behavior is consistent with the model progressively learning complementary feature representations from different receptive fields. As training progressed, the separation between training and validation behavior became more apparent, indicating the emergence of overfitting. Thus, the learning curves suggest that the advantage of the multi-scale architecture is reflected not only in its validation performance but also in its learning dynamics and representation capacity during training.

3.3 Computational Efficiency and Training Runtime

A common concern with multi-branch network architectures is the potential increase in computational overhead. However, the Comparison of Total Training Time in Figure 6 shows that the Multi-Scale CNN remains computationally competitive under the tested environment. Requiring 3839.7 seconds to complete 30 epochs, the proposed model trained faster than the single-scale 3x3 baseline (3957.7s) and 7x7 baseline (3888.6s), but slightly slower than the 5x5 baseline (3792.5s). Thus, in this experiment, the parallel multi-scale configuration did not introduce a prohibitive computational bottleneck, although it should not be interpreted as universally faster than single-scale alternatives.

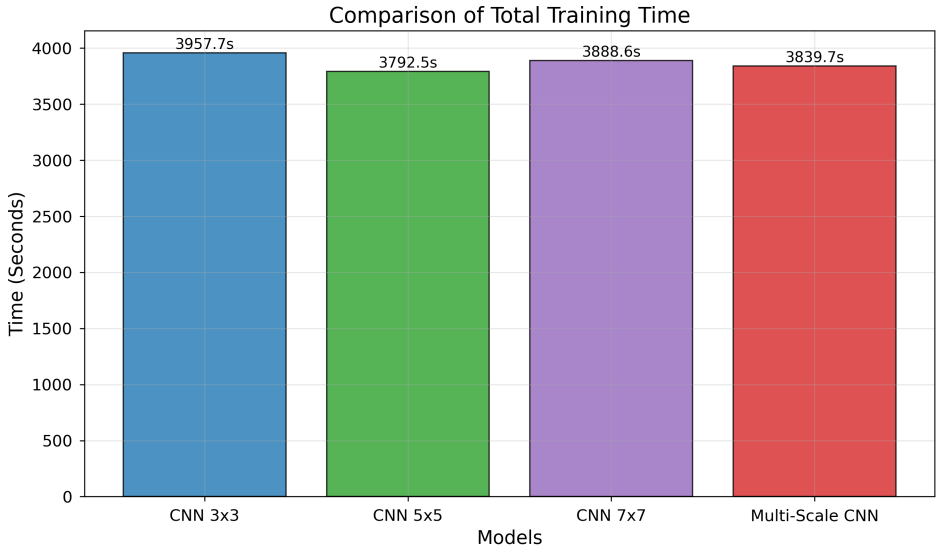

Figure 6 Comparison of Training Time

3.4 Fine-Grained Error Analysis via Confusion Matrix

To evaluate class-level discrimination among the 31 spice categories, Figure 7 presents the confusion matrix for the Multi-Scale CNN. The model achieved high classification recall on categories with distinct geometric silhouettes and fibrous textures, notably *kayu secang* (95.24%), *saffron* (90.48%), *asam jawa* (83.33%), and *serai* (80.95%). Conversely, primary misclassifications were concentrated in two distinct clusters: (1) leafy herbs (*daun kemangi* vs. *daun jeruk* and *daun ketumbar*), and (2) rhizome roots (*jahe*, *kencur*, *lengkuas*, and *kemiri*). These classes share almost identical color palettes and ambiguous surface contours, illustrating the inherent challenge of fine-grained recognition when constrained to ~168 training images per class.

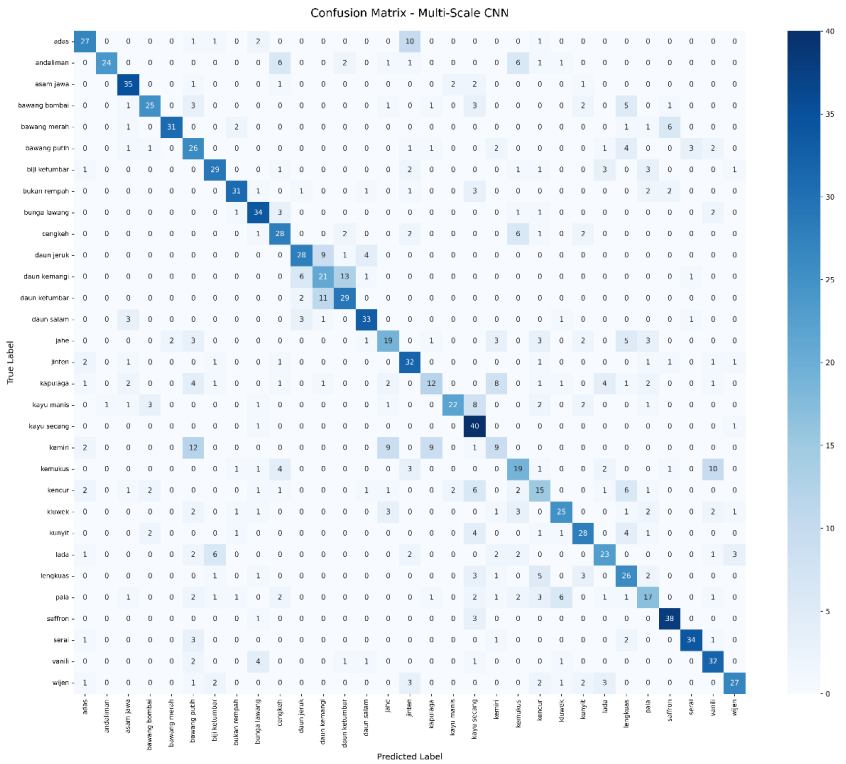

Figure 7 Confusion Matrix of the Proposed Multi-Scale CNN

The confusion patterns observed in Figure 7 provide an important bridge between class-level errors and the aggregate validation behavior. Although combining different receptive-field sizes increases the diversity of extracted features, the remaining errors are concentrated in visually similar classes, particularly the leafy-herb and rhizome groups. This suggests that increased feature-extraction capacity alone may not be sufficient to establish reliable class boundaries when the available training samples are limited. The result motivates examining whether the model's high-dimensional classification representation contributes to the observed generalization gap.

3.5 Linking Class-Level Errors to Architectural Capacity

The architectural behavior is consistent with this interpretation. Direct concatenation of the multi-scale feature maps yields a 150,528-dimensional vector prior to classification, producing approximately 4.66 million dense weights. This large classification bottleneck provides substantial representational capacity during training, consistent with the superior training loss of 0.2713, but it also increases the risk of over-parameterization under the limited-data setting. The structural choice therefore helps explain why stronger training performance does not automatically resolve the fine-grained confusions observed in Figure 7. Rather than treating additional capacity as sufficient, the results indicate a need to control the classification bottleneck and regularize the learned representation.

Taken together, these findings show that the proposed multi-scale structure can enrich spatial representation from scratch, but architectural scaling alone cannot fully offset severe data scarcity. In particular, the study's controlled comparison isolates the benefit of combining receptive fields, while the absence of a separate held-out test set limits claims about out-of-sample generalization beyond the validation split. Future iterations can investigate replacing direct flattening with Global Average Pooling (GAP), applying stronger regularization such as Dropout (0.4–0.5), and evaluating the resulting models on an isolated test set to determine whether these changes improve generalization without obscuring the multi-scale effect.

4. CONCLUSIONS

This study evaluated whether multi-scale feature extraction provides an advantage over single-scale CNNs under limited-data conditions using the Indonesian Spices Dataset as a controlled benchmarking testbed. The proposed multi-scale CNN demonstrated stronger learning capacity than the single-scale architectures, reflected in its higher training performance and its ability to learn complementary visual representations through parallel 3×3, 5×5, and 7×7 convolutional branches. The training and loss curves further indicate favorable learning dynamics during the early and middle stages of training, supporting the stronger learning capacity of the multi-scale architecture. These findings suggest that the primary benefit of multi-scale feature extraction in this setting lies in enriching feature learning and representation capacity, while maintaining competitive validation performance.

Overall, the results support multi-scale feature extraction as a promising approach for improving multi-class fine-grained classification under limited-data conditions. However, the findings are based on a single dataset and experimental setting, with evaluation performed using a validation set rather than an independent held-out test set. Further studies across diverse datasets and experimental settings, including independent test evaluation, can provide additional evidence regarding the reliability and generalizability of these findings.

REFERENCES

Anaya-Isaza, A., & Mera-Jimenez, L. (2022). Data Augmentation and Transfer Learning for Brain Tumor Detection in Magnetic Resonance Imaging. *IEEE Access*, *10*, 23217–23233. https://doi.org/10.1109/ACCESS.2022.3154061

D'souza, R. N., Huang, P.-Y., & Yeh, F.-C. (2020). Structural Analysis and Optimization of Convolutional Neural Networks with a Small Sample Size. *Scientific Reports*, *10*(1), Article ID: 834. https://doi.org/10.1038/s41598-020-57866-2

Duarte, J. M., & Berton, L. (2023). A Review of Semi-Supervised Learning for Text Classification. *Artificial Intelligence Review*, *56*(9), 9401–9469. https://doi.org/10.1007/s10462-023-10393-8

Fini, E., Astolfi, P., Alahari, K., Alameda-Pineda, X., Mairal, J., Nabi, M., & Ricci, E. (2023). *Semi-Supervised Learning Made Simple with Self-Supervised Clustering*. http://arxiv.org/abs/2306.07483

Indira, D. N. V. S. L. S., Markapudi, B. R., Chaduvula, K., & Jyothi Chaduvula, R. (2022). Visual and Buying Sequence Features-Based Product Image Recommendation Using Optimization Based Deep Residual Network. *Gene Expression Patterns*, *45*, Article ID: 119261. https://doi.org/10.1016/j.gep.2022.119261

Jun, W., Son, M., Yoo, J., & Lee, S. (2023). Optimal Configuration of Multi-Task Learning for Autonomous Driving. *Sensors*, *23*(24), Article ID: 9729. https://doi.org/10.3390/s23249729

Kornblith, S., Shlens, J., & Le, Q. V. (2019). *Do Better ImageNet Models Transfer Better?* http://arxiv.org/abs/1805.08974

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kumar, T., Brennan, R., Mileo, A., & Bendechache, M. (2024). Image Data Augmentation Approaches: A Comprehensive Survey and Future Directions. *IEEE Access*, *12*, 187536–187571. https://doi.org/10.1109/ACCESS.2024.3470122

Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical Image Analysis Using Deep Learning Algorithms. *Frontiers in Public Health*, *11*, Article ID: 1273253. https://doi.org/10.3389/fpubh.2023.1273253

Ma, J., Jiang, X., Fan, A., Jiang, J., & Yan, J. (2021). Image Matching from Handcrafted to Deep Features: A Survey. *International Journal of Computer Vision*, *129*(1), 23–79. https://doi.org/10.1007/s11263-020-01359-2

Maharana, K., Mondal, S., & Nemade, B. (2022). A Review: Data Pre-Processing and Data Augmentation Techniques. *Global Transitions Proceedings*, *3*(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

Maray, N., Ngu, A. H., Ni, J., Debnath, M., & Wang, L. (2023). Transfer Learning on Small Datasets for Improved Fall Detection. *Sensors*, *23*(3), Article ID: 1105. https://doi.org/10.3390/s23031105

Parekh, D., Poddar, N., Rajpurkar, A., Chahal, M., Kumar, N., Joshi, G. P., & Cho, W. (2022). A Review on Autonomous Vehicles: Progress, Methods and Challenges. *Electronics*, *11*(14), Article ID: 2162. https://doi.org/10.3390/electronics11142162

Raharja, N. M., Fathansyah, M. A., & Chamim, A. N. N. (2021). Vehicle Parking Security System with Face Recognition Detection Based on Eigenface Algorithm. *Journal of Robotics and Control (JRC)*, *3*(1), 78–85. https://doi.org/10.18196/jrc.v3i1.12681

Rukhiran, M., Wong-In, S., & Netinant, P. (2023). IoT-Based Biometric Recognition Systems in Education for Identity Verification Services: Quality Assessment Approach. *IEEE Access*, *11*, 22767–22787. https://doi.org/10.1109/ACCESS.2023.3253024

Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Siddiqui, T., & Mellit, A. (2023). A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. *Sustainability*, *15*(7), Article ID: 5930. https://doi.org/10.3390/su15075930

Subba, A. B., & Sunaniya, A. K. (2025). Computationally Optimized Brain Tumor Classification Using Attention Based GoogLeNet-style CNN. *Expert Systems with Applications*, *260*, Article ID: 125443. https://doi.org/10.1016/j.eswa.2024.125443

Taye, M. M. (2023). Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. *Computation*, *11*(3), Article ID: 52. https://doi.org/10.3390/computation11030052

Tuggener, L., Schmidhuber, J., & Stadelmann, T. (2022). Is It Enough to Optimize CNN Architectures on ImageNet? *Frontiers in Computer Science*, *4*, Article ID: 1041703. https://doi.org/10.3389/fcomp.2022.1041703

Ullah, N., Khan, J., El-Sappagh, S., El-Rashidy, N., & Khan, M. (2023). A Holistic Approach to Identify and Classify COVID-19 from Chest Radiographs, ECG, and CT-Scan Images Using ShuffleNet Convolutional Neural Network. *Diagnostics*, *13*(1), Article ID: 162. https://doi.org/10.3390/diagnostics13010162

Wu, W., Huo, L., Yang, G., Liu, X., & Li, H. (2025). Research into the Application of ResNet in Soil: A Review. *Agriculture*, *15*(6), Article ID: 661. https://doi.org/10.3390/agriculture15060661

Xiao, F., Wang, H., Li, Y., Cao, Y., Lv, X., & Xu, G. (2023). Object Detection and Recognition Techniques Based on Digital Image Processing and Traditional Machine Learning for Fruit and Vegetable Harvesting Robots: An Overview and Review. *Agronomy*, *13*(3), Article ID: 639. https://doi.org/10.3390/agronomy13030639

Xu, X., Zhang, H., Ma, Y., Liu, K., Bao, H., & Qian, X. (2023). TranSDet: Toward Effective Transfer Learning for Small-Object Detection. *Remote Sensing*, *15*(14), Article ID: 3525. https://doi.org/10.3390/rs15143525

Younesi, A., Ansari, M., Fazli, M., Ejlali, A., Shafique, M., & Henkel, J. (2024). A Comprehensive Survey of Convolutions in Deep Learning: Applications, Challenges, and Future Trends. *IEEE Access*, *12*, 41180–41218. https://doi.org/10.1109/ACCESS.2024.3376441

Zhang, X. (2022). Application of Artificial Intelligence Recognition Technology in Digital Image Processing. *Wireless Communications and Mobile Computing*, *2022*(1), Article ID: 7442639. https://doi.org/10.1155/2022/7442639

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). *Learning Transferable Architectures for Scalable Image Recognition*. http://arxiv.org/abs/1707.07012