Research Article

Balancing Accuracy and Efficiency: A Comparative Study of Deep Learning Architectures on a Limited Thai Food Dataset

Bhurisub Dejpipatpracha[a], Korakot Matarat[b], Suwat Gluaythong[c], Narodom Kittidachanupap[d,] *

[a] *Faculty of Science and Technology, Phuket Rajabhat University, Phuket, 83000, Thailand*
[b] *Faculty of Science and Technology, Sakon Nakhon Rajabhat University, Sakon Nakhon, 47000, Thailand*
[c] *Faculty of Science and Technology, Surindra Rajabhat University, Surin, 32000, Thailand*
[d] *Faculty of Science, Technology and Agriculture, Yala Rajabhat University, Yala, 95000, Thailand*

***Corresponding authors** *: narodom.k@yru.ac.th*
https://doi.org/10.55674/ias.v15i2.266431

Received: 28 February 2026; **Revised:** 18 April 2026; **Accepted:** 20 April 2026; **Available online:** 01 May 2026

Abstract

This study presents a structured benchmarking framework for evaluating five widely adopted convolutional neural network (CNN) architectures —namely DenseNet201, MobileNetV2, ResNet50, VGG19, and NASNetMobile, all initialized with ImageNet pre-trained weights and fine-tuned using transfer learning—under limited-data conditions for Thai food image classification using a curated dataset of 3,000 images across ten food categories. Rather than focusing solely on predictive accuracy, the proposed evaluation integrates computational efficiency metrics—training time, inference latency, and model size—to provide a balanced assessment of model suitability in resource-constrained contexts. Experimental results reveal a consistent trade-off between classification performance and computational demand, highlighting the importance of aligning architectural complexity with dataset scale and operational constraints. Notably, DenseNet201 achieved the highest test accuracy (0.94), while MobileNetV2 provided the most favorable efficiency profile with competitive accuracy (0.91) and the lowest computational overhead. The findings demonstrate that lightweight architectures can achieve competitive accuracy with substantially lower computational overhead, while high-capacity networks deliver marginal performance gains at increased resource cost. By introducing a normalized multi-criteria evaluation strategy, this work advances balanced benchmarking for culturally specific image classification tasks, particularly addressing the visual similarity among Thai dishes and limited availability of high-quality public datasets, and provides practical guidance for model selection in deployment-limited environments.

Keywords: Thai food classification; Convolutional neural network; Resource-constrained learning; Model efficiency; Deployment-oriented evaluation

1. Introduction

Food image classification has become an important research domain within computer vision, driven by applications such as automated calorie estimation, personalized nutrition systems, digital marketing, and culinary heritage preservation. Recent advances in deep convolutional neural networks (CNNs) have substantially improved recognition performance across diverse visual domains, particularly on large-scale benchmark datasets such as ImageNet and Food-101 [1 − 5]. These improvements are largely attributed to deeper architectures and transfer learning strategies enabled by residual and densely connected learning mechanisms [6, 7]. However, while high accuracy has been achieved in controlled large-scale datasets, performance on culturally specific and limited datasets—such as regional cuisine recognition—remains underexplored [8 − 10]. Thai cuisine presents additional challenges due to high inter-class similarity and intra-class variability [11, 12]. Furthermore, practical implementations frequently target mobile or embedded platforms, where memory footprint, computational capacity,

and inference latency are critical constraints [13, 14]. Consequently, selecting an appropriate model requires balancing predictive accuracy with computational efficiency, forming a multi-objective trade-off problem. Many dishes exhibit high intra-class variability while simultaneously sharing strong inter-class similarity, particularly among foods with overlapping ingredients, textures, or preparation methods. Unlike general object recognition tasks, food classification is culturally specific and less transferable across regions, limiting the effectiveness of models trained on generic large-scale datasets. Consequently, achieving reliable performance under small-scale data conditions remains a technically demanding yet comparatively underexplored problem. Existing literature demonstrates that proposed training of the models using cyclical learning rates and representations significantly enhances food recognition accuracy, especially when trained on large annotated datasets [11, 16]. Some papers have reported that the features obtained from deep convolutional neural networks can significantly improve the accuracy of food recognition [15].

More broadly, research consistently indicates that complex architectures benefit from large, high-quality datasets to achieve robust generalization. Nevertheless, far fewer studies systematically examine model behavior when both dataset size and computational resources are constrained. To address this research gap, the present study conducts a structured comparative evaluation of five widely adopted CNN architectures—DenseNet201, MobileNetV2, ResNet50, VGG19, and NASNetMobile on a curated Thai food image dataset comprising ten classes and approximately 3,000 images. All models are trained using transfer learning under identical experimental settings to ensure methodological fairness. By controlling data splits, preprocessing procedures, and training configurations, the study isolates architectural effects from implementation variability. Unlike prior work that focuses predominantly on classification accuracy, this study integrates both predictive performance and computational efficiency metrics to provide a comprehensive and deployment-oriented evaluation of each architecture.

The novelty of this study lies in three main aspects. First, it provides a controlled, architecture-level benchmark comparison under small-scale data conditions specific to Thai cuisine. Second, it introduces an application-oriented evaluation perspective that jointly considers predictive reliability and computational practicality. Third, it proposes a normalized analytical framework that supports balanced model selection beyond accuracy-centric assessment. By explicitly quantifying the interaction between architectural complexity, dataset scale, and computational constraints, this work advances efficiency-aware benchmarking for culturally specific food recognition. The findings offer practical guidance for selecting deep learning architectures suited to resource-constrained settings, emphasizing that model suitability should be determined not only by predictive performance but also by operational viability in real-world scenarios.

Despite the rapid advancement of deep learning architectures, systematic benchmarking under small-scale, culturally specific datasets remains underexplored. Thai cuisine, in particular, presents unique challenges due to its high inter-class visual similarity (e.g., among curry-based dishes) and the absence of a large-scale, publicly available annotated dataset. These characteristics make Thai food classification a compelling and practically relevant testbed for investigating the behavior of CNN architectures under limited-data conditions. Accordingly, this study addresses the need for a deployment-conscious and scale-sensitive evaluation framework tailored to moderate-sized datasets.

2. Literature Review

Recent advances in food image recognition have been largely driven by the rapid development of deep convolutional neural networks (CNNs) [37]. Prior studies can be broadly categorized into four major research directions: (1) accuracy-driven architectural enhancement, (2) optimization and training strategy improvement, (3) lightweight and deployment-oriented modeling, and (4) culturally specific or small-scale dataset adaptation [10].

2.1 Accuracy-Driven Architectural Enhancement

A substantial body of research has focused primarily on improving classification accuracy through architectural modifications and attention mechanisms. For example, the integration of Convolutional Block Attention Module (CBAM) into CNN architectures such as MobileNetV2, VGG16, and ResNet50 significantly improved feature discrimination in Asian food classification tasks, achieving a Top-1 accuracy of 87.33% [6]. Similarly, comparative evaluations on benchmark datasets such as UEC Food-100 demonstrated that deeper architectures such as ResNet, DenseNet, and ResNeXt consistently outperform simpler CNN models when trained with large-scale pretraining [14, 20, 32, 33]. Other studies explored hybrid approaches combining deep feature extraction with traditional classifiers. For instance, pretrained AlexNet and VGG16 features followed by SVM classification achieved strong results on FOOD-5K and FOOD-11 datasets, though performance decreased substantially on larger and more diverse datasets such as FOOD-101 [9, 23]. These findings suggest that while high-capacity models can achieve impressive results, their performance is strongly influenced by dataset scale and diversity. Collectively, these works demonstrate that architectural complexity and attention

mechanisms can substantially improve predictive accuracy—particularly when trained on large, well-annotated datasets. However, most evaluations remain centered on accuracy metrics alone.

2.2 Optimization and Training Strategy Improvements

Beyond architectural design, several studies have investigated optimization strategies to enhance model performance. Particle Swarm Optimization (PSO) has been employed to fine-tune data augmentation parameters and hyperparameters, resulting in measurable accuracy improvements [20]. Data classifications were then classified using various machine learning algorithms. From the experiment, the Random Forest classifier achieved the highest accuracy compared to other classical machine learning methods [21]. Transfer learning remains a dominant paradigm in food recognition research. EfficientNet-B0 and other pretrained architectures have been successfully fine-tuned to classify over 100 food categories with competitive performance [19, 24]. Data augmentation strategies, feature dimensionality reduction techniques (e.g., ReliefF), and training regularization methods have further contributed to improved generalization, particularly under moderate dataset sizes [25, 26]. While these approaches improve classification accuracy, they often assume the availability of sufficient computational resources and do not systematically evaluate the associated training or inference costs.

2.3 Lightweight and Deployment-Oriented Models

With the increasing demand for mobile and embedded applications, lightweight architectures have gained significant attention. Models such as SqueezeNet and MobileNetV2 have demonstrated competitive performance with significantly fewer parameters [14, 27, 28]. Enhanced versions of MobileNet incorporating additional normalization and regularization layers have further improved performance on datasets like ETH Food-101 [26]. In robotic or real-time applications, ResNet-based architectures have achieved over 90% accuracy [29], indicating their robustness in controlled environments. However, such studies rarely quantify inference latency or storage footprint in detail. Although lightweight models are frequently promoted as suitable for real-world deployment, systematic comparisons between high-capacity and mobile-optimized architectures—particularly under identical experimental conditions—remain limited.

2.4 Cultural and Dataset-Specific Studies

Several studies have focused on culturally specific cuisines, including Korean [30], Indonesian Bangladeshi and Vietnamese food datasets. These works highlight the unique challenges of food classification within specific culinary contexts, such as visual similarity between dishes and dataset imbalance. Notably, some studies report exceptionally high accuracy (e.g., above 98%) [31], Some research is expected to contribute to the field of computer vision-related algorithms that are used to solve the problem in image classification, with the state of optimal hyperparameter and validation accuracy level above 90% [33, 34]. But these results often rely on relatively small or well-controlled datasets. In contrast, performance significantly decreases on more complex or diverse datasets such as ETHZ FOOD-101 or UECFOOD256 [32]. This discrepancy suggests that dataset characteristics strongly influence reported performance outcomes. Moreover, classical machine learning approaches using handcrafted features (e.g., Haralick features, Hu Moments, histograms) combined with traditional classifiers have been explored [21], but deep learning models generally outperform such methods when sufficient training data are available. There is a notable lack of detailed research specifically focused on Thai food in the field of food image recognition. Thai cuisine poses unique challenges due to high intra-class variability, complex ingredient compositions, and visually similar dishes. Most existing studies rely on generic or international datasets, which often fail to capture these distinctive characteristics. Therefore, more domain-specific research using well-curated Thai food datasets and context-aware approaches is needed to improve classification performance and ensure cultural relevance.

2.5 Identified Research Gap

Despite extensive research on improving food classification accuracy, several methodological limitations remain evident. First, predictive accuracy is frequently treated as the primary evaluation criterion, with limited integration of computational cost analysis. Second, cross-architecture comparisons are often conducted under inconsistent experimental conditions, reducing benchmarking reliability. Third, systematic examination of performance–efficiency trade-offs in small-scale and culturally specific datasets remains scarce. These limitations motivate the need for a controlled, multi-dimensional evaluation framework capable of jointly assessing predictive reliability and computational feasibility.

In particular, the balance between predictive performance and operational feasibility remains underexplored. High-capacity models may achieve superior accuracy but incur substantial computational overhead, while lightweight models offer efficiency advantages at the potential expense of classification performance.

2.6 Positioning of the Present Study

To address these limitations, the present study conducts a structured and controlled comparative evaluation of five widely adopted CNN architectures—DenseNet201, MobileNetV2, ResNet50, VGG19, and NASNetMobile—on a curated Thai food dataset comprising approximately 3,000 images across ten classes. All models are trained using transfer learning under identical experimental settings to ensure methodological consistency. Unlike prior work that emphasizes accuracy as the primary evaluation criterion, this study introduces a multi-dimensional assessment framework integrating predictive metrics (accuracy, precision, recall, and F1-score) with deployment-oriented indicators (training time, inference latency, and model size). By normalizing heterogeneous metrics into a unified comparison space, the study provides a systematic analysis of the performance–efficiency trade-off.

Previous research studies have explored Vision Transformers (ViT) and hybrid CNN–transformer models for food image recognition, demonstrating improved performance through the use of self-attention mechanisms that capture global contextual information. However, these architectures typically require large datasets and substantial computational resources, resulting in higher training and inference costs. Consequently, their applicability in resource-constrained settings is limited, thereby motivating the focus on CNN-based architectures in this study due to their more efficient balance between performance and computational cost [38]. This critical synthesis emphasizes the need to move beyond accuracy-centric evaluation toward balanced, resource-conscious benchmarking in food image recognition research, particularly for small-scale and culturally specific applications.

3. Materials and Methods

This section describes the datasets, model architectures, training procedures, and evaluation metrics used in this study to investigate the effectiveness and limitations of convolutional neural networks (CNNs) for classifying food images. The proposed evaluation framework extends beyond conventional single-metric comparison by integrating normalized predictive and computational indicators into a unified analytical perspective. This design enables cross-architecture comparison on heterogeneous metrics without bias toward accuracy dominance, thereby supporting structured and transparent decision-making in applied deep learning scenarios.

3.1 Dataset

To comprehensively evaluate the performance of convolutional neural networks (CNNs) for Thai food image classification under limited-data conditions, a curated multi-class dataset was constructed based on previous image processing studies [4]. The dataset comprises approximately 3,000 high-resolution images organized into ten distinct categories: Green_Curry, Khao_phat, Khao_soi, Massaman_Curry, Pad_Thai, Phanaeng_Curry, Phat_kaphrao, Roti_canai, Tom_kha_gai, and Tom_yum. To ensure clarity regarding class distribution and to support the balanced dataset design, a summary of the number of images per class has been provided. The dataset (or a representative subset) has been made publicly available to facilitate validation and reproducibility, and the access URL has been included in the manuscript, referencing the THFOOD-50 dataset hosted on GitHub [39]. A manual curation process was subsequently applied to ensure data quality, including the removal of duplicate images, exclusion of blurred or low-resolution samples, correction of mislabeled instances, and elimination of images with significant occlusion or intrusive watermarks. These procedures ensure that the dataset is representative of real-world conditions and suitable for robust model evaluation. These classes were selected to represent a diverse range of Thai culinary characteristics, including variations in color composition, texture patterns, ingredient structures, and presentation styles (e.g., rice-based dishes, noodle dishes, soups, curries, and desserts/snacks). Notably, Roti Canai—locally known as “roti” in Thailand—reflects culinary adaptation influenced by historical interactions with the Malay Peninsula, particularly in southern regions and urban street food contexts. This diversity introduces inter-class similarity in certain categories (e.g., curry-based dishes) while preserving meaningful visual distinctions, thereby creating a realistic and challenging multi-class classification task.

Although the dataset includes ten food categories, the overall sample size remains relatively small compared to large-scale benchmarks commonly used in food recognition research. This design intentionally reflects resource-constrained conditions, enabling a systematic investigation of how different CNN architectures balance predictive performance, generalization capability, and computational efficiency when trained on limited data.

Fig 1. Image of the difference between the foods used in training.

3.2 Model Architectures

In this study, pre-trained convolutional neural networks were employed, with lower layers responsible for general feature extraction frozen, while higher-level layers were selectively unfrozen and fine-tuned to adapt to the target dataset. A gradual unfreezing strategy was adopted to improve training stability and mitigate overfitting, and a lower learning rate was applied to the pre-trained layers to prevent excessive weight updates. This approach facilitates effective knowledge transfer while preserving adaptability to domain-specific characteristics. Furthermore, to systematically examine the trade-off between predictive performance and computational efficiency under limited-data conditions, five widely used CNN architectures—DenseNet201, MobileNetV2, ResNet50, VGG19, and NASNetMobile—were selected. These models reflect diverse architectural paradigms, ranging from deep residual networks to lightweight, mobile-oriented designs, thereby enabling a comprehensive comparative analysis.

All models were initialized with weights pre-trained on the ImageNet dataset to leverage transfer learning. The original classification heads were removed and replaced with a task-specific fully connected layer corresponding to the ten Thai food categories in this study. A Global Average Pooling (GAP) layer was applied prior to the final dense layer to reduce the number of trainable parameters and mitigate overfitting. The final output layer employed a Softmax activation function to produce normalized class probabilities for multi-class classification.

DenseNet201 adopts a densely connected convolutional architecture in which each layer receives feature maps from all preceding layers within the same dense block. This connectivity pattern promotes feature reuse, strengthens gradient flow, and reduces the vanishing gradient problem. By concatenating feature maps instead of summing them, DenseNet improves parameter efficiency while maintaining strong representational capacity. However, the dense connectivity also increases memory usage during training. The DenseNet architecture was originally introduced by Huang et al. [7].

ResNet50 is a deep residual network composed of 50 layers that incorporates identity shortcut connections. These residual connections enable the network to learn residual mappings rather than direct transformations, thereby facilitating the training of deeper architectures. Residual learning alleviates degradation problems commonly observed in deep networks and has demonstrated strong generalization performance in large-scale image classification tasks. In this study, ResNet50 serves as a high-capacity baseline for performance comparison. The residual learning framework was first proposed by He et al. [6].

VGG19 is a deep convolutional architecture characterized by a simple and uniform design consisting of sequential 3 × 3 convolutional layers followed by max-pooling operations. Although VGG19 contains a large number of parameters due to its fully connected layers, its architectural simplicity makes it a valuable benchmark for evaluating classical deep CNN designs. In resource-constrained settings, however, its large parameter count may increase the risk of overfitting and computational overhead. The VGG architecture was introduced by Simonyan and Zisserman [36].

MobileNetV2 is a lightweight architecture specifically designed for mobile and embedded applications. It utilizes depthwise separable convolutions to significantly reduce computational cost compared to standard convolutions. Additionally, it introduces inverted residual blocks with linear bottlenecks, which preserve representational power while maintaining parameter efficiency. Owing to its compact design, MobileNetV2 is particularly suitable for real-time inference and deployment in resource-constrained environments. MobileNetV2 was proposed by Sandler et al. to enable efficient mobile vision applications [14].

NASNetMobile is derived from Neural architecture search (NAS) methods automatically discover optimized network topologies through reinforcement learning-based strategies [35]. The mobile variant is tailored for efficiency, combining normal and reduction cells to balance accuracy and computational cost. Compared to manually designed architectures, NASNetMobile aims to achieve competitive performance with fewer parameters, making it an appropriate candidate for resource-constrained operational environments. NASNet was introduced by Zoph et al. using neural architecture search [35].

To ensure fair comparison, all architectures were trained under identical experimental conditions, including input resolution (224 × 224), data augmentation strategy, optimizer configuration, and evaluation protocol. Only the final classification layers were fine-tuned during initial training phases, followed by selective unfreezing of deeper layers for further optimization. This standardized setup ensures that performance differences primarily reflect architectural characteristics rather than training discrepancies. Collectively, the selected architectures provide a balanced spectrum of model complexity, depth, parameter size, and computational requirements. This design enables a rigorous assessment of how architectural choices influence classification accuracy, generalization capability, inference speed, and model compactness in small-scale Thai food image classification tasks.

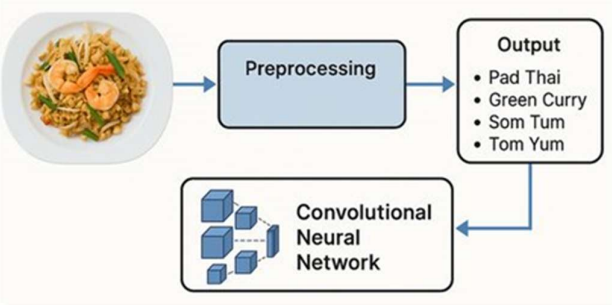

Fig 2. Food image flow diagram.

3.3 Evaluation Framework

This study employed a comprehensive quantitative analysis framework to evaluate and compare the performance of the selected convolutional neural network (CNN) architectures under limited-data conditions. The analysis focused not only on predictive accuracy but also on computational efficiency, reflecting the study's objective of identifying models suitable for practical implementation in resource-constrained environments.

3.3.1 Performance Metrics

Model performance was evaluated on the held-out test set using standard multi-class classification metrics, including overall accuracy, precision, recall, and F1-score. Overall accuracy was defined as the proportion of correctly classified samples among all test instances. Given the multi-class nature of the problem (ten Thai food categories), macro-averaged precision, recall, and F1-score were computed to ensure equal weighting across classes, thereby mitigating bias toward any dominant category.

3.3.2 Computational Efficiency Assessment

In addition to predictive performance, this study systematically evaluates practical implementation factors to assess the operational feasibility of each architecture. Key indicators of computational efficiency were considered, including training time, inference time, model size, and the hardware utilized.

The experimental framework was implemented in Python using the TensorFlow deep learning framework (version 2.x) with the Keras high-level API (tf.keras). Widely adopted libraries, including NumPy, Pandas, Matplotlib, and Seaborn, were utilized for data processing and visualization. Pre-trained convolutional neural network architectures were employed

via the tensorflow.keras.applications module. Image preprocessing and augmentation were performed using the ImageDataGenerator class, and model customization was achieved through the integration of GlobalAveragePooling2D and fully connected (Dense) layers. All experiments were conducted on a workstation equipped with dual Intel® Xeon® CPU E5-2630 v3 processors (2.40 GHz), 64 GB of RAM (2133 MHz), and an NVIDIA GeForce GTX 1060 GPU with 6 GB of dedicated memory, operating on a 64-bit architecture. Key indicators of computational efficiency were systematically evaluated, including total training time, average inference time per image (computed over the test dataset), model size (in megabytes), and the underlying hardware configuration. All models were assessed under identical experimental conditions, enabling a fair and controlled comparison of performance and efficiency across different architectures.

3.3.3 Comparative Trade-off Analysis

To systematically examine the trade-off between predictive performance and computational cost, a multi-criteria comparative analysis was conducted. Performance indicators (accuracy and F1-score) were jointly interpreted with efficiency metrics (inference time and model size). Predictions were executed using a fixed batch size of 32. For visualization and comparative purposes, metrics were normalized to a common scale prior to graphical representation (e.g., radar chart analysis), enabling balanced comparison across heterogeneous measurement units. This trade-off analysis was central to the study, as high predictive accuracy alone may not justify practical deployment if computational demands are excessive. Conversely, lightweight models with slightly lower accuracy may provide superior suitability for real-time or edge-device applications.

3.3.4 Statistical Considerations and Reproducibility

To ensure consistency with the stratified sampling strategy and balanced class representation, the dataset was constructed such that each class contained an equal number of samples (approximately *N* per class), proportionally divided into 70% for training, 20% for validation, and 10% for testing. This design mitigates class imbalance and supports fair model evaluation across all categories. To maintain methodological rigor and prevent data leakage, the subsets were strictly mutually exclusive, with the test set fully held out during model development to provide an unbiased estimate of generalization performance. All images were resized to 224 × 224 pixels and normalized to the range [0, 1] to ensure numerical stability. Data augmentation—including random horizontal flipping, small-angle rotations, zoom scaling, and minor brightness adjustments—was applied exclusively to the training subset to reduce overfitting, while validation and test sets remained unchanged.

For reproducibility and fair cross-architecture comparison, identical data splits, preprocessing procedures, and training configurations were applied across all models. Training was conducted using the Adam optimizer with a learning rate of 0.001, a batch size of 32, and 50 epochs, with early stopping (patience = 10) and dropout (rate = 0.5) to enhance generalization. The categorical cross-entropy loss function was employed due to its suitability for multi-class classification with mutually exclusive labels, ensuring effective optimization and compatibility with the softmax output layer. Hyperparameters were standardized to eliminate confounding factors, and model performance was evaluated through predictive accuracy and computational efficiency. Given the comparative and applied nature of the study, results were interpreted based on practical significance and consistent performance trends rather than formal hypothesis testing, thereby supporting transparent and reliable benchmarking in line with best practices in empirical deep learning research.

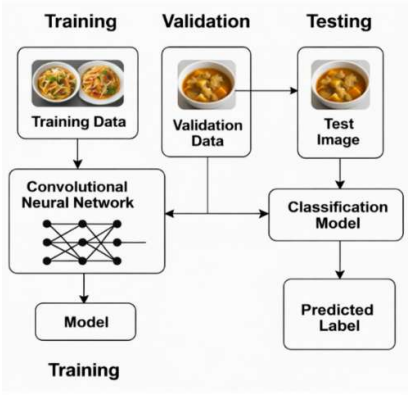

Fig 3. Overview diagram showing the classification model training process using CNN, the testing and validation phase.

4. Results and Discussions

4.1 Predictive Performance Analysis

In this section, aggregate classification metrics—including accuracy, precision, recall, and F1-score—are calculated to provide a balanced view of each model's strengths and weaknesses. These results are presented using tables and graphs to enable a straightforward comparison among DenseNet201, MobileNetV2, ResNet50, VGG19, and NASNetMobile.

By integrating predictive performance metrics with practical deployment considerations—namely training time, inference speed, and model size—this evaluation framework provides a comprehensive assessment of each model's real-world applicability for food image classification. Such a multi-dimensional analysis is essential to ensure that model selection is guided not only by predictive accuracy but also by operational feasibility in resource-constrained environments.

In the equations below, Tp (true positives) and Tn (true negatives) denote the number of correctly classified positive and negative instances, respectively, while Fp (false positives) and Fn (false negatives) denote the number of incorrectly classified instances. These quantities are computed per class under the macro-averaging scheme described above.

$$\text{Accuracy} = \frac{(\text{Tp}+\text{Tn})}{(Tp+Tn+Fp+Fn)} \tag{1}$$

$$\text{Precision} = \frac{\text{Tp}}{(\text{Tp}+\text{Fp})} \tag{2}$$

$$\text{Recall} = \frac{\text{Tp}}{(\text{Tp}+\text{Fn})} \tag{3}$$

$$F1\ Score = \frac{2*(\text{Precision}*\text{Recall})}{(Precision+Recall)} \tag{4}$$

Model performance was evaluated using a multi-dimensional framework to ensure a comprehensive assessment of both predictive capability and computational feasibility. Classification effectiveness was measured using accuracy, precision, recall, and F1-score, while computational efficiency was assessed through training time, inference latency, and model size. Inference time was measured under a standardized CPU environment to approximate realistic deployment conditions, particularly for mobile or resource-constrained systems. Collectively, these metrics establish a structured basis for analyzing the trade-off between predictive accuracy and resource consumption.

4.1.1 Training Time

Training time represents the total wall-clock duration required to complete the full learning process. It can be approximated as:

$$T_{train} = E * N_{batch} * t_{batc} \tag{5}$$

Where E denotes the number of training epochs, N_{batc} is the number of batches per epoch, and, t_{batch} represents the average processing time per batch. This formulation provides a practical estimation of computational cost during model optimization.

4.1.2 Inference Time (per sample)

Inference time for a single sample corresponds to the forward propagation through the trained network and can be expressed as:

$$T_{inference} = t_{forward} \tag{6}$$

Where $t_{forward}$ denotes the time required to complete one forward pass for a single input sample. For N input samples, the total inference time is given by:

$$T_{total} = \text{N} * t_{forward} \tag{7}$$

This formulation assumes a consistent forward-pass time per sample and provides a scalable estimate of deployment latency.

4.1.3 Model Size

Model size reflects storage requirements and is determined by the total number of trainable parameters. It can be estimated as:

$$S_{model} = N_{param} * s_{param} \tag{8}$$

Where N_{param} represents the total number of parameters in the model, and s_{param} denotes the storage size per parameter. In standard 32-bit floating-point representation, $s_{param} = 4$ bytes.

These formulations provide a transparent and reproducible framework for quantifying computational demands, enabling fair comparison of model efficiency alongside predictive performance.

4.2 Computational Efficiency Analysis

The comparative results demonstrate clear trade-offs among training time, inference efficiency, and model size across the evaluated architectures. Lightweight models, particularly MobileNetV2, consistently achieve superior computational efficiency with minimal resource consumption, whereas deeper architectures such as VGG19 and ResNet50 incur higher training or memory costs. These findings highlight the importance of selecting model architectures based on application-specific constraints, especially in scenarios requiring a balance between computational resources and deployment efficiency.

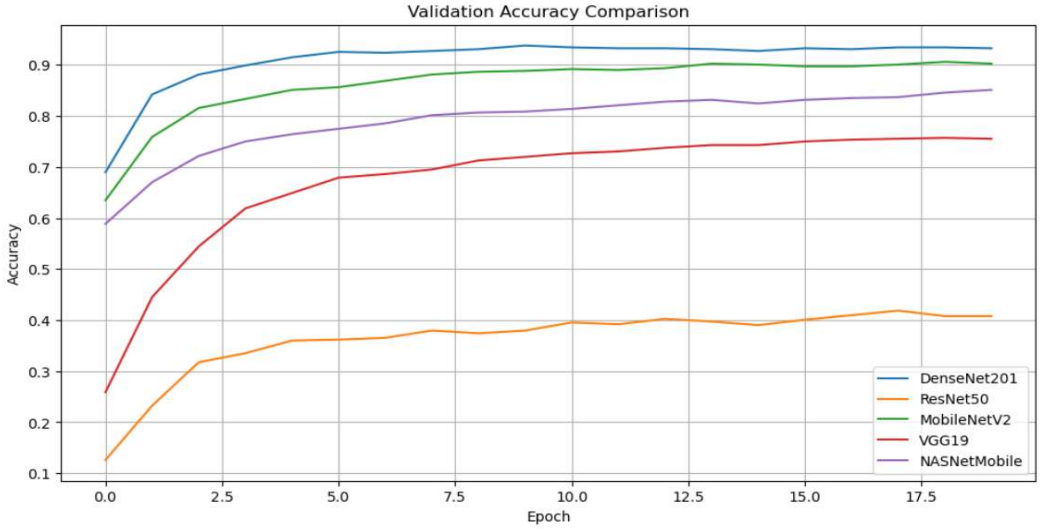

Fig 4. Normalized computational resource usage of deep learning models.

Fig. 4 presents a comparative analysis of computational resource utilization across five deep learning architectures—DenseNet201, MobileNetV2, ResNet50, VGG19, and NASNetMobile—using normalized metrics for training time, inference time, and model size. Normalization enables fair comparison by reducing scale disparities among heterogeneous resource indicators.

The training and validation accuracy curves illustrate model convergence under limited-data conditions. All models exhibit rapid improvement in the initial epochs followed by stabilization, indicating effective convergence. DenseNet201 achieves the highest and most stable validation accuracy, reflecting strong generalization capability, while MobileNetV2 demonstrates consistent and smooth convergence. In contrast, VGG19 and NASNetMobile show slower convergence, and ResNet50 exhibits comparatively lower performance. The absence of significant fluctuations in the validation curves suggests that overfitting is effectively controlled.

From a computational perspective, MobileNetV2 consistently shows the lowest normalized values across all metrics, confirming its suitability for resource-constrained environments. ResNet50 exhibits the largest normalized model size with moderate training and inference times, indicating higher memory demands without proportional efficiency gains. DenseNet201 represents a balanced trade-off, achieving strong predictive performance with moderate computational cost. Overall, these results highlight the trade-offs among training complexity, inference efficiency, and model compactness, providing practical guidance for model selection under varying resource constraints.

4.3 Performance–Efficiency Trade-off Analysis

Before presenting the quantitative results, it is essential to evaluate both predictive performance and computational efficiency in a systematic and balanced manner. In the context of limited data availability and constrained computational resources, assessing model performance solely based on classification accuracy may lead to incomplete conclusions. High-accuracy models may incur substantial training time, inference latency, and memory requirements, which can limit their suitability for real-time or resource-constrained deployment environments. Accordingly, this study adopts a multi-dimensional evaluation framework that incorporates training time, inference time per image, model size, test loss, and test accuracy. These metrics collectively provide a comprehensive perspective on both generalization capability and operational feasibility. All models were trained and evaluated under identical experimental conditions to ensure a fair and unbiased comparison.

The following table summarizes the quantitative comparison of the five evaluated CNN architectures, highlighting the trade-offs between predictive performance and computational efficiency.

Table 1 Comparative performance and computational efficiency of CNN architectures for Thai food image classification.

	Model	Training Time (s)	Inference Time / Image (ms)	Model Size (MB)	Test Loss	Test Accuracy
4	DenseNet201	2324.26	64.55	72.99	0.18	0.94
0	MobileNetV2	444.55	17.11	10.21	0.29	0.91
1	NASNetMobile	478.31	31.43	18.65	0.49	0.83
2	VGG19	2192.90	49.03	76.97	0.69	0.78
3	ResNet50	1521.48	39.42	92.36	1.58	0.44

Table 1 provides a comprehensive comparison of the five evaluated convolutional neural network (CNN) architectures—DenseNet201, MobileNetV2, NASNetMobile, VGG19, and ResNet50—across both predictive performance and computational efficiency metrics. The reported indicators include total training time (seconds), average inference time per image (milliseconds), model size (megabytes), test loss, and test accuracy. All measurements were obtained under identical experimental conditions to ensure a fair and controlled comparison.

Among the evaluated models, DenseNet201 achieved the highest test accuracy (0.94) and the lowest test loss (0.18), indicating strong generalization capability on the held-out dataset. However, this superior predictive performance is accompanied by relatively high computational cost, reflected in extended training time (2324.26 s) and moderate inference latency (64.55 ms per image), consistent with its deep and densely connected architecture.

In contrast, MobileNetV2 demonstrated the most favorable efficiency profile. It achieved competitive test accuracy (0.91) while requiring substantially shorter training time (444.55 s), the fastest inference speed (17.11 ms per image), and the smallest model size (10.21 MB). These characteristics highlight its suitability for real-time applications and deployment in resource-constrained environments, where latency and memory footprint are critical considerations.

NASNetMobile exhibited moderate predictive performance (test accuracy = 0.83) with balanced computational requirements. Although more efficient than heavier architectures such as VGG19 and ResNet50, it did not surpass MobileNetV2 in either accuracy or efficiency.

VGG19 showed lower classification performance (test accuracy = 0.78) combined with a relatively large model size (76.97 MB) and prolonged training time (2192.90 s), reflecting its high parameter count and traditional fully connected design. To contextualize the performance–efficiency trade-off in Thai food classification, it is important to consider the high intra-class variability and inter-class similarity among dishes, particularly in visually similar categories such as curry-based foods. The superior performance of DenseNet201 can be attributed to its ability to capture fine-grained visual features, enabling better discrimination between similar classes, although at a high computational cost. In contrast, MobileNetV2 achieves a strong balance between accuracy and efficiency, indicating its suitability for real-world applications, especially in resource-constrained environments such as mobile or edge devices. The comparatively lower performance of NASNetMobile, VGG19, and ResNet50 suggests that increased architectural complexity does not necessarily lead to improved results under limited-data conditions. Overall, these findings emphasize the importance of selecting models that balance predictive performance with computational efficiency in practical Thai food classification scenarios. Finally, ResNet50 yielded the lowest test accuracy (0.44) and highest test loss (1.58), indicating limited generalization under the present configuration. It also exhibited the largest model size (92.36 MB) and substantial training time (1521.48 s), resulting in the least favorable performance-to-efficiency ratio in this resource-constrained setting. This unexpectedly low performance warrants further investigation. Possible explanations include suboptimal convergence due to inappropriate learning rate scheduling, insufficient fine-tuning of deeper layers in the residual blocks, or incompatibility between the residual architecture's identity shortcut connections and the limited training signal available from the small dataset.

Overall, the results reveal a clear trade-off between predictive accuracy and computational efficiency. While DenseNet201 achieved the highest classification performance, MobileNetV2 provided the most balanced combination of accuracy, inference speed, training time, and compactness. These findings reinforce the importance of jointly considering predictive reliability and resource constraints when selecting models for real-world implementation.

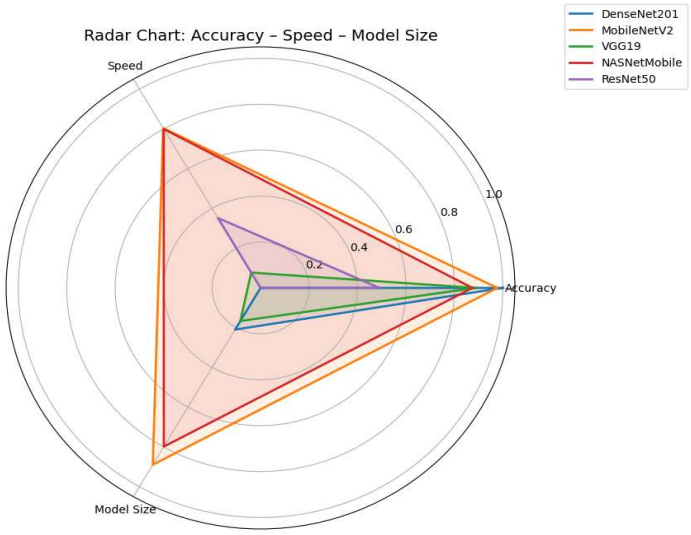

Fig 5. Normalized multi-criteria comparison of CNN architectures for Thai food image classification.

The radar chart illustrates the comparative performance of DenseNet201, MobileNetV2, NASNetMobile, VGG19, and ResNet50 across three normalized evaluation criteria: classification accuracy, inference efficiency, and model compactness. All metrics were min–max normalized to the range [0,1], where higher values consistently indicate better overall performance (i.e., higher accuracy, faster inference, and smaller model size). The visualization clearly demonstrates the trade-off between predictive performance and computational efficiency. DenseNet201 achieves the highest normalized accuracy, reflecting its superior generalization capability on the test set. However, its efficiency scores are moderate due to longer training and inference times and a relatively large model size. MobileNetV2 exhibits the most balanced overall profile, combining competitive classification accuracy with the highest normalized inference efficiency and strongest model compactness. This confirms its suitability for deployment in resource-constrained environments. NASNetMobile presents intermediate performance across all criteria, offering moderate accuracy and efficiency without excelling in any single dimension. VGG19 achieves lower accuracy and moderate efficiency, while ResNet50 demonstrates the weakest overall profile, characterized by low classification accuracy and the largest model size, resulting in limited suitability under the current resource-constrained setting.

Overall, the radar visualization reinforces that no single architecture simultaneously maximizes both predictive accuracy and computational efficiency. Model selection should therefore be guided by application-specific constraints, particularly when practical feasibility is a primary concern.

5. Conclusion

This study presented a controlled benchmark evaluation of five CNN architectures—DenseNet201, MobileNetV2, ResNet50, VGG19, and NASNetMobile—under limited-data conditions for Thai food classification. Among the evaluated models, DenseNet201 achieved the highest test accuracy (0.94) with the lowest test loss (0.18), demonstrating superior generalization capability. In contrast, MobileNetV2 delivered the most favorable efficiency profile, achieving competitive accuracy (0.91) with the shortest training time (444.55 s), fastest inference speed (17.11 ms/image), and smallest model size (10.21 MB). VGG19 and NASNetMobile showed moderate performance, while ResNet50 yielded unexpectedly low accuracy (0.44), warranting further investigation into its convergence behavior under the present experimental configuration. The findings confirm a consistent trade-off between predictive performance and computational efficiency, highlighting the importance of aligning architectural complexity with dataset scale and system constraints. From a practical standpoint, the proposed normalized multi-criteria framework provides a structured methodology for balanced model selection, enabling informed decision-making beyond accuracy-centric evaluation. The central implication of this research is that robust performance in culturally specific, small-scale classification tasks requires integrative assessment of both predictive reliability and computational feasibility.

6. Future Work

Although this study demonstrates structured evaluation of CNN architectures under limited-data conditions, several research directions remain open for further investigation.

First, expanding the dataset both in scale and diversity would enhance generalization robustness and enable deeper analysis of class imbalance and intra-class variability. Future work may incorporate additional Thai food categories, cross-regional variations, and real-world deployment images captured under uncontrolled conditions. A larger dataset would also allow investigation of scaling behavior, examining how model performance evolves as dataset size increases. Second, model compression and efficiency optimization techniques should be explored to further improve deployment feasibility. Approaches such as pruning, weight quantization, knowledge distillation, and low-rank approximation could significantly reduce memory footprint and inference latency without substantial degradation in accuracy. Given the strong baseline performance of MobileNetV2, distillation from high-capacity models such as DenseNet201 into lightweight student networks represents a promising direction. Third, hybrid and attention-based architectures could be explored, alongside the implementation of K-fold cross-validation, to improve discriminative capability while preserving computational efficiency. For example, lightweight attention modules or transformer-inspired feature refinement mechanisms may improve inter-class separability among visually similar Thai dishes (e.g., curry-based categories) without dramatically increasing parameter count. Fourth, future studies should incorporate robustness and cross-domain generalization analysis. Evaluating models under domain shifts—such as varying lighting conditions, camera devices, or restaurant environments—would provide deeper insight into real-world reliability. Domain adaptation techniques or self-supervised pretraining strategies may help mitigate performance degradation in such scenarios. Fifth, multi-objective optimization frameworks could be developed to automatically balance accuracy and efficiency during model selection. Instead of evaluating architectures post hoc, future research may formulate architecture selection as a constrained optimization problem, integrating computational budgets directly into training objectives.

Finally, extending the system toward practical applications—such as real-time dietary monitoring or mobile-based food recognition platforms—would further validate the real-world applicability of the proposed evaluation framework. Field implementation studies measuring user-level latency, energy consumption, and performance under natural usage conditions would provide valuable empirical evidence beyond controlled experimental settings. Collectively, these directions aim to advance application-oriented food image recognition by narrowing the gap between high-performance deep learning models and resource-constrained operational environments.

References

[1] Y. He, S. Yin, Food Images Classification based on Improved Convolutional Neural Network, International Conference on Electronic Communication and Artificial Intelligence (ICECAI), IEEE, 12 – 14 May 2023, 290 – 293.

[2] A. N. M. Zulfikri, F. Y. A. Rahman, S. Shabuddin, R. Mohamad, Food Recognition based on Deep Learning Algorithms, 2022 IEEE Symposium on Industrial Electronics & Applications (ISIEA), Langkawi Island, Malaysia, IEEE,16 – 17 July 2022, 1 – 4.

[3] L. Luo, Research on Food Image Recognition of Deep Learning Algorithms, 2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE), Ottawa, ON, Canada, IEEE, 26 – 28 August 2023, 733 – 737.

[4] O. Russakovsky et al., ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput. Vis. 115 (2015) 211 – 252.

[5] L. Bossard et al., Food-101 – Mining Discriminative Components with Random Forests, ECCV2014, Lecture Notes in Computer Science, Computer Vision (ECCV), Zurich, Switzerland, 6 – 12 September 2014, 446 – 461.

[6] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27 June 2016, 770 – 778.

[7] G. Huang, Z. Liu, L. V. D. Maaten, Densely Connected Convolutional Networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21 – 26 July 2017, 4700 – 470.

[8] T.-H. Do, D.-D.-A. Nguyen, H.-Q. Dang, H.-N. Nguyen, P.-P. Pham and D.-T. Nguyen, 30VNFoods: A Dataset for Vietnamese Foods Recognition, 2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), Purwokerto, Indonesia, 17 – 18 July 2021, 311 – 315.

[9] T. Trong Nguyen, T. Q. Nguyen, D. Vo, V. Nguyen, N. Ho, N. D. Vo, K. V. Nguyen, K. Nguyen, VinaFood21: A Novel Dataset for Evaluating Vietnamese Food Recognition, 2021 RIVF International Conference on Computing and Communication Technologies (RIVF), Hanoi, Vietnam, 19 – 21 Aug. 2021, 1 – 6.

[10] G. Ciocca, P. Napoletano, R. Schettini, Food Recognition: A New Dataset, Experiments, and Results, IEEE J. Biomed. Health Inform. 21(3) (2017), 588 – 598.

[11] N. Theera-Ampornpunt, P. Treepong, Thai Food Recognition Using Deep Learning with Cyclical Learning Rates, IEEE Access, 12 (2024) 174204 – 174221.

[12] T. Chakkrit, K. Surachet, NU-InNet: Thai food image recognition using convolutional neural networks on smartphone, Journal of Telecommunication, JTEC. 9(2-6) (2017) 63 – 67.

[13] O. M. Salim, R. M. Zeebaree, A. M. Sadeeq, A. H. Radie, Study for Food Recognition System Using Deep Learning, J. Phys.: Conf. Ser. 1963(1) 2021, 012014.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18 – 23 June 2018, 4510 – 4520.

[15] Y. Kawano, K. Yanai, Food image recognition with deep convolutional features, Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct), New York, NY, USA, 13 September 2014, 589 – 593.

[16] Z. Shen, A. Shehzad, S. Chen, H. Sun, J. Liu, Machine Learning Based Approach on Food Recognition and Nutrition Estimation, Procedia Comput. Sci. 174 (2020), 448 – 453.

[17] B. Xu, X. He, Z. Qu, Asian Food Image Classification Based on Deep Learning, JCC. 9(03) (2021), 10 – 28.

[18] S. Memiş, B. Arslan, O. Z. Batur, E. B. Sönmez, A Comparative Study of Deep Learning Methods on Food Classification Problem, 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 15 – 17 October 2020, 1 – 4.

[19] P. K. Singh, S. Susan, Transfer Learning using Very Deep Pre-Trained Models for Food Image Classification, 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 06 – 08 July 2023, 1 – 6.

[20] S. Sreetha, G. Naveen Sundar, D. Narmadha, Enhancing Food Image Classification with Particle Swarm Optimization on NutriFoodNet and Data Augmentation Parameters, IJCESEN. 10(4) (2024).

[21] P. K. Fahira, Z. P. Rahmadhani, P. Mursanto, A. Wibisono, H. A. Wisesa, Classical Machine Learning Classification for Javanese Traditional Food Image, 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 10 – 11 November 2020, 1 – 5.

[22] A. Şengür, Y. Akbulut, Ü. Budak, Food Image Classification with Deep Features, 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 21 – 22 September 2019, 1 – 6.

[23] G. Ciocca, G. Micali, P. Napoletano, State Recognition of Food Images Using Deep Features, IEEE Access. 8 (2020), 32003 – 32017.

[24] G. VijayaKumari, P. Vutkur, P. Vishwanath, Food classification using transfer learning technique, Glob. Transit. Proc. 3(1) 2022, 225 – 229.

[25] K. Sanjeev, M. Meleet, Food Recognition Using Extreme Learning Machines, IJRASET. 10(XI) (2022), 2321 – 9653.

[26] S. Phiphiphatphaisit, O. Surinta, Food Image Classification with Improved MobileNet Architecture and Data Augmentation, ICISS '20: Proceedings of the 3rd International Conference on Information Science and Systems, 20 April 2020, 51 – 56.

[27] S. Yadav, Alpana, S. Chand, Automated Food image Classification using Deep Learning approach, 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 19 – 20 March 2021, 542 – 545.

[28] S. Yadav, S. Chand, Food image recognition based on MobileNetV2 using support vector machine, AIJR Proceedings, Proceedings of International Conference on Women Researchers in Electronics and Computing (WREC 2021), 22 – 24 April 2021, 92 – 200.

[29] Y. Xiong, Food image recognition based on ResNet, Appl. Comput. Eng. 8 (2023), 605 – 611.

[30] M. Chun, H. Jeong, H. Lee, T. Yoo, H. Jung, Development of Korean Food Image Classification Model Using Public Food Image Dataset and Deep Learning Methods, IEEE Access, 08 December 2022, 128732 – 128741.

[31] A. A. Jeny, M. S. Junayed, I. Ahmed, M. T. Habib, M. R. Rahman, FoNet - Local Food Recognition Using Deep Residual Neural Networks, 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, 9 – 21 December 2019, 184 – 189.

[32] Z. Zahisham, C. P. Lee, K. M. Lim, Food Recognition with ResNet-50, 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, 26 – 27 September 2020, 1 – 5.