open access
Published online on: https://journal.uii.ac.id/jurnalsnati/
Check for updates
DOI: 10.20885/snati.v5.i2.48169
JURNAL SNATI Sains, Nalar, dan Aplikasi Teknologi Informasi
Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi
Vol. 5 No. 2 (2026) 108 - 117 ISSN Media Electronic: 2807-5935

Accuracy–Efficiency Trade-off Analysis of CNN Backbones for Multi-Class Waste Classification

Mochamad Rizal Fauzan[1*], Irgi Surya[2], Resa Pramudita[3]
[1,2] Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan
[3] Industrial Automation and Robotics Engineering Education Study, Universitas Pendidikan Indonesia, Bandung, Indonesia
[1]rizalfauzan2002@gmail.com, [2]irgi.surya88@gmail.com, [3]resa.pd@upi.edu

Abstract

Automated waste classification is a critical component of intelligent recycling systems, where model selection must balance predictive performance and computational efficiency. This study benchmarks three representative convolutional neural network (CNN) backbones—ResNet50, EfficientNet-B0, and MobileNetV3-Large—for eight-class waste classification under controlled augmentation and unified optimization protocols. Using a fixed 80/10/10 split (7,747 training, 969 validations, and 960 testing images), all models are evaluated across multiple random seeds to ensure statistical reliability. Performance is assessed using macro-F1, precision, recall, and GPU-based inference latency to characterize the accuracy–efficiency trade-off. EfficientNet-B0 achieves the highest macro-F1 (0.9676 ± 0.0034), while MobileNetV3-Large delivers comparable performance (0.9669 ± 0.0023) with substantially lower latency—approximately six times faster than ResNet50. Augmentation sensitivity analysis further reveals architecture-dependent robustness under occlusion-based perturbation. These results demonstrate that lightweight architectures can achieve near-optimal classification performance with significantly reduced computational cost, providing deployment-oriented guidelines for practical waste sorting systems.

Keywords: Waste classification; CNN backbones; Accuracy–efficiency trade-off; Inference latency; Model benchmarking; Deep learning

1. Introduction

The global generation of municipal solid waste has reached approximately 2.24 billion tons annually and is projected to increase to 3.88 billion tons by 2050, intensifying environmental and economic pressures on urban infrastructures [1], [2]. Efficient waste sorting plays a central role in improving recycling rates and reducing landfill dependency; however, manual segregation remains labor-intensive, inconsistent, and operationally inefficient [3], [4]. In recent years, deep learning-based computer vision systems have demonstrated promising performance in automated waste recognition tasks, leveraging convolutional neural networks (CNNs) for multi-class material classification, including plastic, paper, metal, glass, textile, and biological waste categories [5], [6], [7].

While predictive accuracy has been the dominant evaluation criterion in prior studies, real-world deployment environments such as edge-based recycling kiosks, conveyor-belt sorting systems, and embedded smart bins impose strict constraints on computational resources, inference latency, and energy consumption [8], [9]. For instance, Fauzi and Haqdu D [8] compared EfficientNetB0, MobileNetV2, and ResNet50 for a related image classification task and found that the higher-capacity ResNet50 outperformed the lighter architectures, illustrating that the accuracy–efficiency relationship can be highly task-dependent. In a deployment-focused study, Li and Grammenos [9] implemented lightweight CNN-based waste classifiers on embedded edge devices (Jetson Nano and K210), reporting strong accuracy with reduced power consumption, but without multi-seed evaluation or systematic augmentation sensitivity analysis. In such contexts, a marginal increase in accuracy may not justify a substantial increase in model complexity or latency. For instance, high-capacity architectures such as ResNet50 contain over 25 million parameters, whereas lightweight models such as MobileNetV3-Large reduce parameter counts significantly through depthwise separable convolutions and channel-wise attention mechanisms [10], [11], [12]. Despite these architectural differences, systematic investigations that jointly analyze predictive performance and computational efficiency within waste classification tasks remain limited, and existing comparisons such as [8] and [9] either target different domains or omit stability and augmentation-sensitivity analysis.

Existing literature frequently reports single-model experiments or accuracy-only comparisons without standardized training protocols, multi-seed evaluation, or controlled augmentation strategies [13], [14]. Moreover, many studies rely on single-split reporting, potentially introducing variance-induced bias that

Received: 31-03-2026 | Revised: 22-06-2026 | Accepted: 03-07-2026 | Published: 15-07-2026

obscures reliable backbone selection guidelines. The absence of multi-run stability analysis and deployment-oriented metrics such as inference time per image limits the practical interpretability of reported results. Additionally, augmentation strategies, particularly those involving spatial perturbation or occlusion simulation, may interact differently with architectural design characteristics. However, augmentation sensitivity at the backbone level in waste classification has not been systematically quantified.

Collectively, the lack of (i) unified training protocols across architectures, (ii) multi-seed stability evaluation, (iii) integration of macro-level performance metrics such as macro-F1 for balanced multi-class assessment, and (iv) computational efficiency analysis constitutes a methodological gap in the current literature. Without a structured accuracy–efficiency perspective, practitioners lack empirically grounded criteria for selecting backbone architectures suitable for deployment-constrained recycling systems.

To address these limitations, this study conducts a systematic accuracy–efficiency trade-off analysis of three representative CNN backbones ResNet50, EfficientNet-B0, and MobileNetV3-Large for eight-class waste classification. Using a standardized 80/10/10 dataset split (7,747 training, 969 validations, and 960 testing images), all models are trained under identical optimization settings and evaluated across multiple random seeds to ensure statistical stability. Performance is assessed using macro-F1 score, precision, recall, and GPU-based inference latency, enabling integrated analysis of predictive capacity and computational cost. Furthermore, controlled augmentation regimes including geometric, photometric, and occlusion-based transformations are employed to examine architecture-dependent robustness patterns.

By bridging predictive performance evaluation with deployment-oriented efficiency analysis under controlled experimental conditions, this work provides empirically grounded guidelines for backbone selection in practical multi-class waste classification systems.

2. Research Methods

Figure 1 illustrates the overall research workflow. The waste image dataset, comprising eight material categories, is initially loaded and partitioned into predefined training, validation, and testing subsets using an 80/10/10 split to ensure controlled and unbiased evaluation [15], [16]. All images undergo preprocessing, including automatic orientation correction and resizing to 224 × 224 pixels to comply with the input requirements of ImageNet-pretrained convolutional neural network (CNN) backbones. Three augmentation regimes are subsequently defined to investigate robustness and augmentation sensitivity: A0 (baseline transformation), A1 (geometric and photometric perturbations), and A2 (A1 supplemented with Random Erasing to simulate partial occlusion). For each augmentation configuration, three representative CNN backbones—ResNet50, EfficientNet-B0, and MobileNetV3-Large—are initialized with pretrained weights and adapted to the eight-class classification task. All models are trained under a unified optimization protocol using AdamW with a learning rate of 1×10^{-4}, batch size of 32, weight decay of 1×10^{-4}, and 35 epochs, while multi-seed control is enforced to mitigate stochastic variance and enhance statistical reliability. Model selection is performed based on validation macro-F1 score to ensure balanced multi-class assessment [17], [18]. The selected models are then evaluated on the held-out test set, where performance metrics—including accuracy, precision, recall, and macro-F1—are computed. In parallel, inference latency is measured in milliseconds per image under GPU execution following a warm-up phase to quantify computational efficiency. The integration of predictive performance and latency measurement enables systematic characterization of the accuracy–efficiency trade-off across backbone architectures and augmentation strategies.

2.1. Dataset Description

This study utilizes a multi-class waste image dataset comprising 9,676 images distributed across eight material categories: biological (737), cardboard (689), clothes (3,674), glass (1,080), metal (582), paper (803), plastic (663), and shoes (1,448). The dataset was compiled by the authors and made publicly available via the Roboflow Universe platform, accessible at https://universe.roboflow.com/rizalfanex/waste-classification-lm0j6-lowap.

The detailed class distribution is summarized in Table 1. As shown in Table 1, the dataset exhibits noticeable class imbalance, with the clothes category accounting for approximately 37.97% of the total samples, while categories such as metal and plastic represent less than 7% each. Rather than artificially rebalancing the dataset, the original class proportions are preserved to reflect realistic waste composition patterns encountered in practical recycling environments.

The dataset is partitioned into predefined training, validation, and testing subsets using a fixed 80/10/10 split, resulting in 7,747 training images, 969 validation images, and 960 testing images. The split is established prior to model training to ensure strict separation between development and evaluation stages, thereby preventing data leakage and preserving unbiased performance assessment.

Copyright © 2026 | Fauzan et al. | Licensee Universitas Islam Indonesia

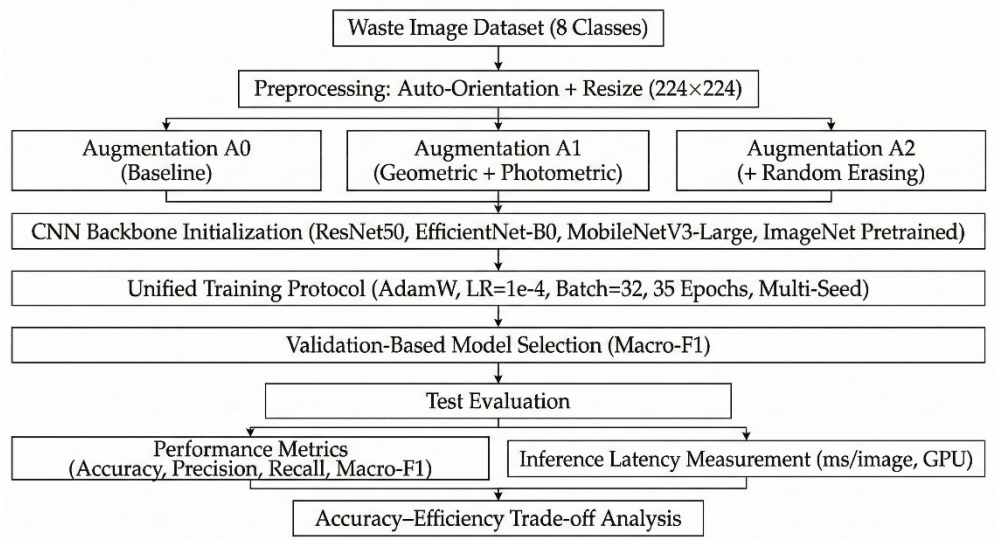

Figure 1. Research Flow

All images undergo automatic orientation correction followed by resizing to 224 × 224 pixels to comply with the input resolution requirements of ImageNet-pretrained convolutional neural network backbones. This standardization ensures consistent transfer learning initialization across architectures while maintaining computational comparability. Representative sample images from each waste category are illustrated in Figure 2, highlighting both intra-class variability and inter-class visual similarity among material-based categories such as plastic, metal, and glass. No class rebalancing, oversampling, or synthetic data generation techniques are applied at the dataset level, allowing macro-averaged metrics to meaningfully reflect classification robustness under imbalanced conditions.

Table 1. Class Distribution of the Waste Image Dataset

Class	Total Images	Percentage (%)
Biological	737	7.61%
Cardboard	689	7.12%
Clothes	3,674	37.97%
Glass	1,080	11.16%
Metal	582	6.01%
Paper	803	8.30%
Plastic	663	6.85%
Shoes	1,448	14.97%
Total	**9,676**	**100%**

Figure 1. Presents representative samples from each waste category.

2.2. Data Preprocessing and Augmentation

All input images are normalized using ImageNet mean and standard deviation statistics to ensure compatibility with pretrained backbone initialization. To examine augmentation sensitivity and robustness across architectures, three controlled augmentation regimes are evaluated: A0 (baseline), A1 (geometric–photometric perturbation), and A2 (occlusion-augmented regime). While A0 serves as a reference configuration, A1 introduces moderate spatial and illumination variability, and A2 further incorporates localized occlusion to assess architecture-dependent resilience. Validation and testing subsets undergo deterministic preprocessing without stochastic augmentation to ensure fair performance comparison.

Table 2. Augmentation Configurations

Regime	Transformations Applied
A0	RandomResizedCrop (scale 0.8–1.0)
A1	A0 + RandomHorizontalFlip (p=0.5), RandomRotation (±15°), ColorJitter (brightness/contrast/saturation ±0.2, hue ±0.05)
A2	A1 + RandomErasing (p=0.25, scale 0.02–0.15, ratio 0.3–3.3)

2.3. Model Architectures

Three representative convolutional neural network (CNN) backbones are deliberately selected to span the spectrum from high-capacity to lightweight design, each reflecting distinct design philosophies and computational profiles [19]. ResNet50 is selected as a high-capacity baseline representative of conventional residual learning; EfficientNet-B0 is selected to represent parameter-efficient architectures obtained through compound scaling; and MobileNetV3-Large is selected as a state-of-the-art lightweight architecture explicitly designed for mobile and edge deployment. These architectures differ in depth, parameterization strategy, and efficiency optimization mechanisms,

Copyright © 2026 | Fauzan et al. | Licensee Universitas Islam Indonesia

thereby enabling structured analysis of the accuracy–efficiency trade-off across a representative range of deployment scenarios relevant to practical waste-sorting systems.

ResNet50 is a deep residual network composed of 50 layers with identity skip connections that facilitate stable gradient propagation and mitigate vanishing gradient issues in deep architectures. Its residual learning framework has established strong baseline performance across diverse image recognition tasks, making it a widely adopted reference architecture [20].

EfficientNet-B0 employs compound scaling to balance network depth, width, and resolution in a parameter-efficient manner. By jointly scaling these dimensions according to predefined scaling coefficients, EfficientNet aims to achieve improved accuracy per parameter compared to conventional deep architectures.

Table 3. Backbone Architecture Characteristics

Model	Approx. Parameters	Model Size (MB)	FLOPs/MACs (G)	Design Strategy	Efficiency Focus
ResNet50	~25.6 M	~97.8	~4.1	Residual learning with skip connections	Baseline high-capacity model
EfficientNet-B0	~5.3M	~20.5	~0.39	Compound scaling (depth–width–resolution)	Parameter efficiency
MobileNetV3-Large	~5.4M	~21.1	~0.22	Depthwise separable conv + SE blocks	Lightweight deployment

MobileNetV3-Large represents a lightweight architecture optimized for computational efficiency and deployment-constrained environments. It integrates depthwise separable convolutions and squeeze-and-excitation mechanisms to reduce parameter count and floating-point operations while maintaining competitive representational capacity.

All backbones are initialized with ImageNet pretrained weights to leverage transfer learning benefits. The original classification heads are replaced with fully connected layers corresponding to the eight waste categories. Full fine-tuning is performed without freezing backbone parameters to ensure equitable optimization across architectures and avoid bias toward pretrained feature hierarchies.

2.4. Training Protocol

To ensure methodological consistency and eliminate optimization bias across backbone architectures, all experiments follow a unified and structured training pipeline, as summarized in Algorithm 1. The protocol standardizes initialization, optimization, model selection, and evaluation procedures to guarantee that performance differences arise solely from architectural characteristics rather than hyperparameter discrepancies.

Algorithm 1. Experimental Training and Evaluation Procedure

Input: Waste image dataset D, augmentation regime $A \in \{A0, A1, A2\}$, backbone architecture $M \in \{ResNet50, EfficientNet-B0, MobileNetV3-Large\}$, number of epochs $E = 35$, random seed set S

Output: Test performance metrics (Accuracy, Precision, Recall, Macro-F1) and inference latency for each backbone–augmentation configuration

Steps:

1. Load dataset D. Apply automatic orientation correction and resize all images to 224×224. Partition the dataset into training (80%), validation (10%), and testing (10%) subsets.
2. For each augmentation regime A, apply the corresponding training-time transformations (see Table 2).
   Apply deterministic preprocessing to validation and testing subsets.
3. For each backbone architecture M, initialize with ImageNet pretrained weights.
   Replace the final classification layer to match the eight output classes.
4. For each random seed $s \in S$:
   - Set deterministic seeds for Python, NumPy, and PyTorch.
   - Train model for $E = 35$ epochs using AdamW optimizer
   - (learning rate 1×10^{-4}, weight decay 1×10^{-4}, batch size 32).
   - Use CrossEntropyLoss as the objective function.
   - Adjust learning rate via ReduceLROnPlateau (factor = 0.5, patience = 3) based on validation macro-F1.
   - Save the checkpoint achieving the highest validation macro-F1.
5. Load the best-performing checkpoint.
   Evaluate the model on the held-out test set and compute Accuracy, Precision, Recall, and Macro-F1.
6. Measure inference latency (milliseconds per image) under GPU execution following a warm-up phase to remove initialization overhead.
7. Report final performance as mean ± standard deviation across all random seeds.

All experiments were conducted on an Acer Nitro laptop equipped with an NVIDIA GeForce RTX 5060 GPU and 32GB RAM, using PyTorch 2.12.1 and CUDA 13.3. All latency measurements reported in this study were obtained under this fixed hardware configuration.

2.5. Evaluation Metrics

To rigorously evaluate multi-class classification performance under imbalanced data distribution, a confusion-matrix–based metric formulation is adopted. Let $C = 8$ denote the number of classes and N the total number of test samples. For each class $c \in \{1, \ldots, C\}$, let TP_c, FP_c, FN_c, and TN_c denote the number of true positives, false positives, false negatives, and true negatives, respectively. The confusion matrix therefore provides a complete characterization of prediction outcomes across all classes.

Copyright © 2026 | Fauzan et al. | Licensee Universitas Islam Indonesia

Overall accuracy, defined as the proportion of correctly classified samples, is expressed in (1):

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} TP_c}{N} \quad (1)$$

Although accuracy provides a global correctness measure, it may be biased toward dominant categories in imbalanced datasets. Therefore, per-class precision and recall are computed as defined in (2) and (3):

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \quad (2)$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \quad (3)$$

Precision in (2) quantifies prediction reliability, whereas recall in (3) measures sensitivity to class-specific detection.

To ensure that minority classes contribute equally to performance evaluation, macro-averaging is adopted. Macro-precision and macro-recall are computed by averaging per-class metrics as defined in (4) and (5):

$$\text{Macro-Precision} = \frac{1}{C}\sum_{c=1}^{C} \text{Precision}_c \quad (4)$$

$$\text{Macro-Recall} = \frac{1}{C}\sum_{c=1}^{C} \text{Recall}_c \quad (5)$$

The macro-F1 score, defined as the harmonic mean of macro-precision and macro-recall, is expressed in (6):

$$\text{Macro-F1} = 2 \cdot \frac{\text{Macro-Precision} \cdot \text{Macro-Recall}}{\text{Macro-Precision} + \text{Macro-Recall}} \quad (6)$$

Macro-F1 in (6) is selected as the primary model selection criterion due to its robustness under uneven class distributions (Table 1). Unlike accuracy, macro-F1 assigns equal importance to each category regardless of sample frequency, thereby preventing performance inflation driven by the dominant clothes class (37.97%).

For completeness, weighted-F1 could alternatively be defined by weighting each class according to its support. However, this study prioritizes macro-F1 to emphasize balanced performance across all waste categories rather than frequency-weighted dominance.

In addition to predictive performance, computational efficiency is quantified through inference latency. Let t_i denote the forward-pass time required to process sample i. The average latency per image is defined in (7):

$$\text{Latency} = \frac{1}{N}\sum_{i=1}^{N} t_i \quad (7)$$

To ensure stable estimation, a warm-up phase is conducted prior to latency measurement to eliminate initialization overhead and GPU synchronization artifacts. The reported latency therefore reflects steady-state inference performance under practical deployment conditions.

Finally, the joint interpretation of Macro-F1 in (6) and Latency in (7) enables quantitative characterization of the accuracy–efficiency trade-off. A backbone is considered deployment-efficient if it achieves competitive macro-F1 while maintaining significantly lower inference latency relative to higher-capacity models.

3. Results and Discussions

3.1. Overall Predictive Performance

Table 4 summarizes the overall classification performance (mean ± standard deviation across three random seeds) for all backbone–augmentation configurations. Macro-F1 is adopted as the primary evaluation metric due to the imbalanced class distribution (Table 1), while overall accuracy and latency are reported for completeness.

Across all configurations, the highest Macro-F1 score is achieved by EfficientNet-B0 with A2 augmentation (0.9676 ± 0.0034). However, the performance margin relative to MobileNetV3-Large with A1 augmentation (0.9669 ± 0.0023) is minimal (Δ = 0.0007), well within the observed standard deviation range. This indicates that lightweight architectures can achieve near-equivalent predictive performance despite substantially lower parameter complexity.

ResNet50 does not demonstrate superior classification performance despite its significantly higher computational capacity. Under the best augmentation setting (A1 or A2), ResNet50 achieves Macro-F1 ≈ 0.9654, which remains slightly below EfficientNet-B0 (A2) and MobileNetV3-Large (A1). This suggests diminishing returns from increased depth and parameterization for this dataset scale.

Latency measurements reveal a pronounced computational hierarchy. MobileNetV3-Large achieves inference times around 0.50–0.53 ms per image, EfficientNet-B0 requires approximately 0.93–1.34 ms, and ResNet50 incurs over 3.1 ms per image. Notably, ResNet50 is approximately six times slower than MobileNetV3-Large while providing no statistically meaningful improvement in Macro-F1.

Augmentation effects also exhibit architecture-dependent behavior. Moderate geometric–photometric augmentation (A1) consistently improves or maintains performance across models. In contrast, the occlusion-based A2 configuration benefits EfficientNet-B0 but slightly degrades MobileNetV3-Large performance (0.9602 ± 0.0053), indicating that robustness to spatial perturbation is not solely determined by model size but also by architectural feature aggregation mechanisms.

Standard deviations across all configurations remain low (generally < 0.006 in Macro-F1), confirming strong training stability and indicating that

Copyright © 2026 | Fauzan et al. | Licensee Universitas Islam Indonesia

performance differences are systematic rather than driven by stochastic initialization.

These results demonstrate that computational efficiency mechanisms play a more decisive role than raw parameter count in balancing predictive accuracy and inference cost for multi-class waste image classification. Lightweight architectures can achieve performance parity with deeper networks while substantially reducing computational overhead.

3.2. Accuracy–Efficiency Trade-off

Table 4. Pareto-Oriented Comparison Using the Best Configuration per Backbone

Backbone	Best Aug	Macro-F1 (mean±std)	Latency (ms/img, mean±std)
EfficientNet-B0	A2	0.9676 ± 0.0034	1.341 ± 0.750
MobileNetV3-Large	A1	0.9669 ± 0.0023	0.530 ± 0.041
ResNet50	A1 (tie with A2)	0.9654 ± 0.0043	3.136 ± 0.044

To explicitly quantify the deployment-oriented trade-off between predictive performance and computational cost, Figure 3 visualizes Macro-F1 versus inference latency (ms/image) for all backbone–augmentation combinations, while Table 5 summarizes the most representative "best" configuration per backbone. Together, these results provide a compact view of which architecture lies on (or near) the Pareto-efficient frontier.

As shown in Figure 3, the three backbones form clearly separated latency regimes: MobileNetV3-Large consistently occupies the low-latency region (≈0.48–0.53 ms/image), EfficientNet-B0 lies in the mid-latency region (≈0.93–1.34 ms/image), and ResNet50 clusters in the high-latency region (≈3.12–3.14 ms/image). Importantly, this separation in computational cost is not matched by proportional gains in Macro-F1. Instead, Macro-F1 values across the best-performing settings are tightly clustered, indicating that larger capacity does not translate into commensurate predictive benefit on this dataset.

From Table 5, EfficientNet-B0 with A2 achieves the highest Macro-F1 (0.9676 ± 0.0034). However, MobileNetV3-Large with A1 attains nearly identical performance (0.9669 ± 0.0023), with an absolute difference of only 0.0007 in Macro-F1—well within the variability range—while reducing latency to 0.530 ± 0.041 ms/image. In contrast, ResNet50 does not deliver a performance advantage (Macro-F1 ≈ 0.9654) despite incurring substantially higher inference cost (≈3.136 ms/image), placing it outside the practical Pareto-efficient set for latency-sensitive deployment.

Figure 3 and Table 6 jointly demonstrate that the most deployment-efficient operating point is achieved by MobileNetV3-Large (A1), which provides near-optimal Macro-F1 at the lowest latency. EfficientNet-B0 (A2) is competitive when marginal gains in Macro-F1 are prioritized, whereas ResNet50 offers no favorable accuracy–latency trade-off under the evaluated settings.

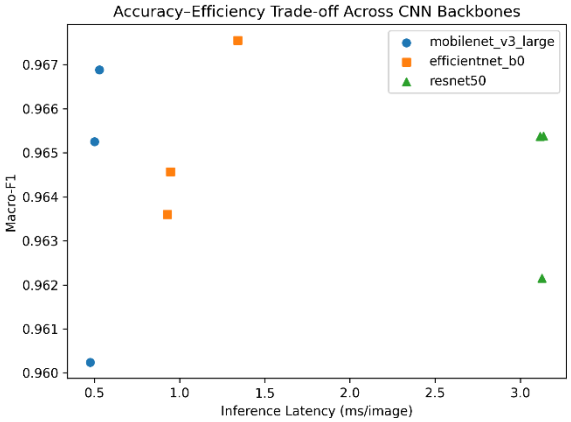

Figure 2. Accuracy–Efficiency Trade-off Across Backbone Architectures.

3.3. Trade-off Class-wise Behavior and Error Pattern

To further investigate model behavior beyond aggregate metrics, Table 6 reports the per-class precision, recall, and F1-score for the best-performing configuration (EfficientNet-B0 with A2

Table 5. Per-Class Performance for the Best Configuration (EfficientNet-B0 + A2)

Class	Precision (mean±std)	Recall (mean±std)	F1-score (mean±std)
Biological	0.9843 ± 0.0002	0.9792 ± 0.0090	0.9817 ± 0.0046
Cardboard	0.9755 ± 0.0170	0.9613 ± 0.0167	0.9684 ± 0.0169
Clothes	0.9965 ± 0.0015	0.9837 ± 0.0039	0.9901 ± 0.0015
Glass	0.9753 ± 0.0140	0.9602 ± 0.0106	0.9677 ± 0.0079
Metal	0.9603 ± 0.0192	0.9600 ± 0.0000	0.9601 ± 0.0096
Paper	0.9431 ± 0.0213	0.9574 ± 0.0067	0.9501 ± 0.0082
Plastic	0.9429 ± 0.0093	0.9649 ± 0.0176	0.9537 ± 0.0102
Shoes	0.9505 ± 0.0004	0.9877 ± 0.0085	0.9687 ± 0.0043

Copyright © 2026 | Fauzan et al. | Licensee Universitas Islam Indonesia

augmentation). Figure 4 presents the corresponding average confusion matrix across seeds.

The clothes category achieves the highest F1-score (0.9901), reflecting both high precision (0.9965) and strong recall (0.9837). This is consistent with its dominant representation in the dataset, which provides richer feature diversity during training.

The biological category also demonstrates strong discriminative performance (F1 = 0.9817), indicating that organic texture patterns are well captured by the model.

In contrast, paper and plastic exhibit relatively lower precision values (≈0.94), suggesting mild confusion with visually similar material-based classes. Notably, plastic shows higher recall (0.9649) than precision, implying that while most plastic samples are correctly detected, false positives from neighboring categories slightly degrade precision.

The metal category maintains balanced precision and recall (~0.96), though with slightly higher variability in precision (std = 0.0192), likely due to reflectance similarities with glass under certain lighting conditions.

F1-scores across all classes remain above 0.95, indicating strong balanced performance and confirming that the model does not disproportionately favor dominant categories.

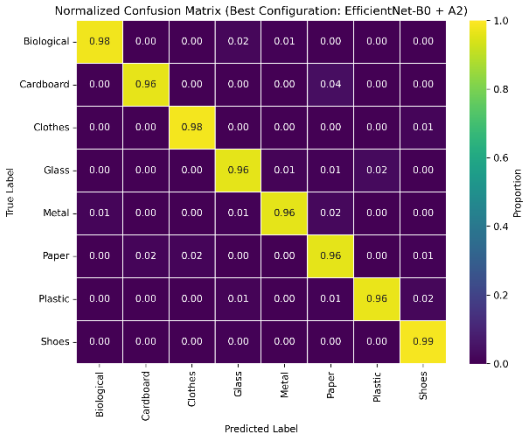

Figure 3. Confusion Matrix for the Best Configuration.

The confusion matrix in Figure 4 reveals that most misclassifications occur between material-based classes with similar surface properties, particularly plastic–glass and paper–cardboard. Importantly, there is minimal confusion between semantically distant categories (e.g., biological vs metal), indicating that classification errors are primarily driven by visual similarity rather than global feature extraction failure.

The diagonal dominance of the matrix confirms strong class separability, while off-diagonal entries remain sparse and concentrated among related materials. This pattern supports the earlier observation that model limitations are rooted in inter-material ambiguity rather than architectural capacity constraints.

3.4. Stability and Sensitivity Analysis

To further analyze experimental robustness, we examine performance variability and augmentation sensitivity across backbone architectures rather than reiterating aggregate metrics reported in Table 5.

Across all configurations, Macro-F1 standard deviation remains below 0.009, indicating strong convergence stability. MobileNetV3-Large consistently exhibits the lowest variability, particularly under A1 augmentation (±0.0023), suggesting that lightweight architectures may benefit from more stable optimization dynamics under moderate geometric–photometric augmentation.

EfficientNet-B0 demonstrates marginally higher sensitivity when Random Erasing is applied (A2), where performance variance slightly increases (±0.0034). This suggests that compound scaling architectures may respond more strongly to stochastic spatial perturbations, though without compromising mean performance.

ResNet50 shows comparatively higher variability under baseline augmentation (A0), reflecting increased optimization sensitivity due to its deeper residual structure. However, the variance remains sufficiently small to preserve ranking consistency across seeds.

Importantly, no backbone exhibits instability or collapse under any augmentation regime. The relative ordering observed in Sections 3.1 and 3.2 is preserved across seeds, confirming that the accuracy–efficiency conclusions are structurally robust rather than seed-dependent artifacts.

3.5. Discussions

The experimental findings demonstrate that architectural capacity does not scale linearly with predictive performance for multi-class waste image classification. Although EfficientNet-B0 achieves the highest Macro-F1 under the strongest augmentation setting (A2), the performance margin relative to MobileNetV3-Large remains minimal. This pattern is plausibly explained by the nature of the visual cues that distinguish waste material categories: classes such as metal, glass, and plastic are primarily separated by mid-level texture, reflectance, and color cues rather than fine-grained semantic detail. Lightweight architectures such as MobileNetV3-Large and EfficientNet-B0 are already well-suited to capturing such cues through depthwise separable convolutions and compound scaling, respectively, while ResNet50's additional depth and parameter capacity offer limited benefit because the task does not demand the highly abstract, hierarchical representations that deeper residual networks are designed to extract. In other words, increasing network depth beyond a certain threshold yields diminishing returns when the task

Copyright © 2026 | Fauzan et al. | Licensee Universitas Islam Indonesia

primarily relies on mid-level texture and material cues rather than highly abstract semantic representations.

The trade-off analysis further reinforces this observation. MobileNetV3-Large consistently operates in a substantially lower latency regime while maintaining near-parity in Macro-F1. The inference gap between MobileNetV3-Large and ResNet50 exceeds a factor of six, yet the predictive performance difference remains marginal [20]. This pattern contrasts with findings reported in other CNN comparison studies; for instance, Fauzi and Haqdu D [8] found that ResNet50 outperformed lighter architectures by a notable margin for traffic density classification, suggesting that the relative advantage of high-capacity architecture is highly task-dependent rather than universal. In the present waste classification task, depthwise separable convolutions and squeeze-and-excitation modules appear to provide sufficient representational power for structured material recognition without incurring unnecessary computational overhead, although this interpretation is specific to the evaluated dataset and would benefit from validation on additional waste classification benchmarks. From a systems perspective, this highlights the importance of evaluating model architectures within the context of deployment constraints rather than focusing exclusively on peak accuracy.

Augmentation sensitivity analysis reveals that moderate geometric–photometric transformations enhance generalization consistently across all backbones. However, the introduction of occlusion-based perturbation through Random Erasing produces architecture-dependent behavior. EfficientNet-B0 benefits from the additional regularization effect, likely due to its compound scaling design that balances resolution and channel capacity. Conversely, MobileNetV3-Large exhibits slight degradation under aggressive occlusion (Macro-F1 decreasing from 0.9669 ± 0.0023 under A1 to 0.9602 ± 0.0053 under A2, as shown in Table 4), a pattern tentatively consistent with the hypothesis that lightweight architectures possess reduced representational redundancy to compensate for erased discriminative regions. However, this interpretation remains speculative and would require targeted ablation studies, such as feature-map analysis, to confirm, since the observed degradation could equally stem from optimization sensitivity rather than an inherent architectural property. This finding nonetheless underscores that augmentation strategies should be aligned with architectural capacity rather than uniformly applied [21], [22].

Class-wise evaluation further indicates that model errors are not random but structurally concentrated among visually similar material categories. Confusion primarily occurs between plastic, glass, and paper—classes sharing comparable reflectance and texture patterns. In contrast, semantically distant categories such as biological and metal remain well separated. This pattern implies that the remaining performance ceiling is driven more by inter-material visual ambiguity than by insufficient model depth. Enhancements may therefore require domain-specific feature enrichment, such as material-aware representations or multimodal sensing, rather than deeper convolutional stacks.

Importantly, multi-seed evaluation confirms that the observed ranking among backbones is statistically stable. Performance variance remains low across all configurations, and no architecture exhibits instability or convergence collapse. This consistency supports the robustness of the trade-off conclusions and indicates that the comparative hierarchy reflects structural architectural properties rather than stochastic training artifacts.

From a practical standpoint, the results suggest that MobileNetV3-Large under moderate augmentation offers the most favorable balance between predictive accuracy and computational efficiency. EfficientNet-B0 remains competitive when marginal accuracy gains are prioritized, whereas ResNet50 does not provide a compelling justification under latency-sensitive conditions. Consequently, backbone selection for real-world waste sorting systems should prioritize efficiency-aware architectures that maintain strong discriminative capability without excessive computational cost.

Overall, the study demonstrates that effective model selection in waste classification systems must consider the joint landscape of accuracy, stability, and inference efficiency. The findings challenge the implicit assumption that deeper networks necessarily yield superior deployment performance and instead highlight the relevance of architectural efficiency in practical computer vision applications.

It is also important to acknowledge two limitations of the present study. First, all experiments rely on a single waste image dataset; while the dataset spans eight common material categories, the generalizability of the accuracy–efficiency conclusions to other waste compositions, lighting conditions, or geographic contexts remains to be verified. Second, the reported inference latency values are specific to the hardware configuration used in this study (Section 2.4); absolute latency figures may differ on alternative deployment hardware such as embedded or mobile devices, although the relative ordering among backbones is expected to remain consistent given the substantial margins observed. Future work involving cross-dataset validation and multi-hardware latency profiling would further strengthen the generalizability of these findings.

Copyright © 2026 | Fauzan et al. | Licensee Universitas Islam Indonesia

4. Conclusion

This study investigated the comparative performance of ResNet50, EfficientNet-B0, and MobileNetV3-Large for eight-class waste image classification under controlled training conditions and consistent augmentation regimes. Rather than focusing solely on accuracy, the analysis incorporated inference latency, class-wise behavior, and multi-seed stability to provide a more deployment-oriented evaluation framework.

The findings indicate that increasing architectural depth does not automatically translate into meaningful performance gains for material-based waste categorization. While EfficientNet-B0 achieves the highest Macro-F1 under stronger augmentation, the performance difference relative to MobileNetV3-Large remains marginal. At the same time, MobileNetV3-Large maintains substantially lower inference latency, demonstrating that lightweight architectures can capture discriminative material features without excessive computational cost. In contrast, ResNet50 introduces significantly higher latency without providing a proportional predictive advantage.

Class-level analysis shows that most residual errors occur among visually similar recyclable materials such as plastic, glass, and paper, suggesting that the remaining performance gap is driven more by inter-material ambiguity than by insufficient model depth. Stability analysis further confirms that performance rankings remain consistent across random seeds, reinforcing the reliability of the comparative results.

These findings emphasize that backbone selection for practical waste classification systems should be guided by both predictive performance and computational constraints. Lightweight architectures emerge as strong candidates for real-time or edge-based smart waste sorting applications, where inference efficiency is critical. Future work may explore material-aware feature enhancement or multimodal approaches to further reduce inter-class ambiguity and improve robustness under real-world variability.

Reference

[1] M. Chhabra, B. Sharan, M. Elbarachi, and M. Kumar, “Intelligent waste classification approach based on improved multi-layered convolutional neural network,” Multimed. Tools Appl., vol. 83, no. 36, pp. 84095–84120, Nov. 2024, doi: 10.1007/s11042-024-18939-w.

[2] M. Nahiduzzaman et al., "An automated waste classification system using deep learning techniques: Toward efficient waste recycling and environmental sustainability," Knowledge-Based Syst., vol. 310, p. 113028, Feb. 2025, doi: 10.1016/j.knosys.2025.113028.

[3] M. Diqi, “Waste Classification using CNN Algorithm,” Int. Conf. Inf. Sci. Technol. Innov., vol. 1, no. 1, pp. 130–135, Feb. 2022, doi: 10.35842/icostec.v1i1.17.

[4] M. Malik et al., “Waste Classification for Sustainable Development Using Image Recognition with Deep Learning Neural Network Models,” Sustain., vol. 14, no. 12, Jun. 2022, doi: 10.3390/su14127222.

[5] W. Hurst et al., “Solid Waste Image Classification Using Deep Convolutional Neural Network,” Infrastructures 2022, Vol. 7, Page 47, vol. 7, no. 4, p. 47, Mar. 2022, doi: 10.3390/infrastructures7040047.

[6] J. J. C. Simbolon, Robet, and Hendri, “Household Waste Image Classification Using Deep Learning Model,” J. Artif. Intell. Eng. Appl., vol. 5, no. 1, pp. 1681–1690, Oct. 2025, doi: 10.59934/jaiea.v5i1.1690.

[7] L. S. Pieters, “Development of Automatic Waste Classification System using CNN-Based Deep Learning to Support Smart Waste Management,” INOVTEK Polbeng - Seri Inform., vol. 10, no. 1, pp. 214–224, Mar. 2025, doi: 10.35314/wst8mh87.

[8] D. R. Fauzi and G. A. H. D, “Comparison of CNN Models Using EfficientNetB0, MobileNetV2, and ResNet50 for Traffic Density with Transfer Learning,” J. Intell. Syst. Technol. Informatics, vol. 1, no. 1, pp. 22–30, Jun. 2025, doi: 10.64878/jistics.v1i1.6.

[9] X. Li and R. Grammenos, “Evaluation of practical edge computing CNN-based solutions for intelligent recycling bins,” IET Smart Cities, vol. 5, no. 3, pp. 194–209, Sep. 2023, doi: 10.1049/smc2.12057.

[10] Z. Qiao, “Advancing Recycling Efficiency: A Comparative Analysis of Deep Learning Models in Waste Classification,” Nov. 2024, Accessed: Mar. 01, 2026. [Online]. Available: http://arxiv.org/abs/2411.02779

[11] Y. Chen, Y. He, J. Lin, and S. Sun, “Garbage image recognition and classification based on CNN,” Appl. Comput. Eng., vol. 4, no. 1, pp. 416–421, May 2023, doi: 10.54254/2755-2721/4/20230507.

[12] E. D. Cherpanath, P. R. Fathima Nasreen, K. Pradeep, M. Menon, and V. S. Jayanthi, “Food Image Recognition and Calorie Prediction Using Faster R-CNN and Mask R-CNN,” 9th Int. Conf. Smart Comput. Commun. Intell. Technol. Appl. ICSCC 2023, pp. 83–89, 2023, doi: 10.1109/ICSCC59169.2023.10335053.

[13] M. S. Minu, V. Priya Kasa, H. Ketharaman, V. Mandepudi, and S. Krishnaa, “Waste Classification Using Machine Learning Models: A Comparative Study,” 2026, doi: 10.5220/0013927500004919.

[14] G. Ahmad et al., “Intelligent waste sorting for urban sustainability using deep learning,” Sci. Reports 2025 151, vol. 15, no. 1, pp. 27078-, Jul. 2025, doi: 10.1038/s41598-025-08461-w.

[15] M. Castro-Bello et al., “Convolutional Neural Network Models in Municipal Solid Waste Classification: Towards Sustainable Management,” Sustain. 2025, Vol. 17, Page 3523, vol. 17, no. 8, p. 3523, Apr. 2025, doi: 10.3390/su17083523.

[16] A. Thakur, S. K. Biswas, S. Majumdar, and D. M. Thounaojam, “A Comprehensive Review on Different CNN Architectures,” 3rd Int. Conf. Intell. Data Commun. Technol. Internet Things, IDCIoT 2025, pp. 1997–2004, 2025, doi: 10.1109/IDCIOT64235.2025.10914792.

[17] M. Rybczak and K. Kozakiewicz, “Deep Machine Learning of MobileNet, Efficient, and Inception Models,” Algorithms 2024, Vol. 17, Page 96, vol. 17, no. 3, p. 96, Feb. 2024, doi: 10.3390/a17030096.

[18] M. Ragazzi et al., “Waste Classification for Sustainable Development Using Image Recognition with Deep Learning Neural Network Models,” Sustain. 2022, Vol. 14, Page 7222, vol. 14, no. 12, p. 7222, Jun. 2022, doi: 10.3390/su14127222.

[19] C. J. Yi and C. F. Kim, “AI-Powered Waste Classification Using Convolutional Neural Networks (CNNs),” Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 10, pp. 67–75, Oct. 2024, doi: 10.14569/IJACSA.2024.0151009.

[20] C. Chen, N. A. Mat Isa, and X. Liu, “A review of convolutional neural network based methods for medical image classification,” Comput. Biol. Med., vol. 185, no. 2, p. 109507, Feb. 2025, doi: 10.1016/j.compbiomed.2024.109507.

[21] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, “Review of Image Classification Algorithms Based on Convolutional Neural Networks,” Remote Sens. 2021, Vol. 13, Page 4712, vol. 13, no. 22, p. 4712, Nov. 2021, doi: 10.3390/rs13224712.

Copyright © 2026 | Fauzan et al. | Licensee Universitas Islam Indonesia

[22] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," Artif. Intell. Rev. 2024 574, vol. 57, no. 4, pp. 99-, Mar. 2024, doi: 10.1007/s10462-024-10721-6.

Copyright © 2026 | Fauzan et al. | Licensee Universitas Islam Indonesia