Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9844

Comparing Convolutional Neural Networks for Waste Image Classification with Visual Explanation and Label Audit

Teuku Muhammad Mirza Keumala[1], Mohd. Iqbal Muttaqin[2*], Nur Aynun Siregar[3]
[1,2,3]Universitas Bina Bangsa Getsempena, Indoensia
[1]mirza@bbg.ac.id, [2*]iqbalmuttaqin@bbg.ac.id, [3]aynun@bbg.ac.id

open access

***Corresponding Author**

Article History:
Submitted: 19-10-2026
Accepted: 28-10-2026
Published: 11-11-2026

Keywords:
Convolutional neural network; Explainable artificial intelligence; Label noise; MobileNetV2; Transfer learning; Waste classification.

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Automated waste classification from images can support smart sorting bins, but the choice of a convolutional neural network for low-cost hardware is rarely evaluated together with the quality of the training labels. This study aims to compare three pretrained architectures, namely ResNet18, MobileNetV2, and EfficientNet-B0, for classifying waste images into recyclable, electronic, and organic classes, and to examine why their errors occur. The models were trained through transfer learning under an identical protocol on 26,527 images from the Satria Data 2026 Big Data Challenge, using class-weighted cross-entropy and early stopping, and were evaluated on a stratified validation set of 2,651 images in terms of accuracy, macro F1-score, parameter count, model size, and inference time. Gradient-weighted class activation mapping was applied as an explainable artificial intelligence technique, and the 40 most confident misclassifications were audited manually to estimate label noise. MobileNetV2 achieved the highest accuracy (92.68%) and macro F1-score (0.9239) while also being the smallest model (2.23 million parameters, 8.71 megabytes) and the fastest (16.78 milliseconds per image on a laptop processor). All three architectures confused recyclable and organic waste most often. The visual explanations showed a reliance on surface texture, and the audit attributed 30% of the confident errors to probable label noise, mostly organic images predicted as recyclable. Lightweight networks are therefore a sound starting point for smart bins, and label auditing should accompany accuracy reporting in waste classification studies.

INTRODUCTION

Household waste keeps growing with urbanisation, and in many developing regions, including Indonesia, sorting at source is still done by hand or not at all. Mixing waste streams lowers the value of recyclable materials and makes it harder to isolate hazardous items such as discarded electronics, which need a separate disposal route (Shreya, Nimal Yughan, Katyal, & Ramesh, 2023). Image-based classification with convolutional neural networks (CNNs) has become the main route to automating this step, because networks pretrained on large image collections can be adapted to new waste categories through transfer learning with a modest amount of labelled data (Gürcan & Soylu, 2026; Zhuang et al., 2021)

Deploying such a classifier inside a smart bin raises two practical questions. The first concerns hardware. A bin is typically controlled by a single-board computer or a microcontroller, so the model must be accurate and small at the same time. Lightweight architectures such as MobileNetV2 were designed for this setting (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018), and several waste studies report that they match heavier networks (Kumala, Sari, & Rachmawanto, 2025; Mehedi et al., 2023). The second question concerns the data. Many public waste datasets were assembled by web search, and their labels are rarely verified (Nnamoko, Barrowclough, & Procter, 2022). Label errors of this kind can change which model appears best, even in carefully curated benchmarks (Northcutt, Athalye, & Mueller, 2021).

Existing waste classification studies answer these questions only in part. Most report the accuracy of a single architecture or compare two architectures trained with different settings (Girsang, Saputra, & Yanrie, 2023; Kumala et al., 2025; Oktayaessofa, Sari, Rachmawanto, & Yaacob, 2024), which makes it hard to attribute differences to the architecture itself. Computational cost is seldom reported next to accuracy. Visual explanation methods such as gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) are increasingly attached to waste classifiers (Alkılınç, Okay, Kök, & Özdemir, 2026; Manik et al., 2026), but mainly to show that the model looks at the object, not to diagnose systematic errors. To our knowledge, no waste classification study has quantified how much of its remaining error is caused by the labels rather than by the model.

This study addresses these gaps on a three-class dataset (Recyclable, Electronic, and Organic) released for the Satria Data 2026 Big Data Challenge (BDC2026). Three research questions guide the work: which of three widely used pretrained CNNs offers the best balance between accuracy and computational cost under an identical training protocol,

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9844

which visual cues drive the dominant errors, and how much of the high-confidence error can be attributed to label noise. The study makes three contributions. First, it provides a controlled comparison of ResNet18, MobileNetV2, and EfficientNet-B0 that reports accuracy, macro F1-score, parameter count, model size, and inference time on a common processor. Second, it uses Grad-CAM to link the persistent confusion between recyclable and organic waste to surface texture cues. Third, it introduces a confidence-ranked manual audit that estimates the share of label noise among the most confident errors and shows that this share differs sharply between error directions.

LITERATURE REVIEW

Early deep learning work on waste images trained networks from scratch. (Altikat, Gulbe, & Altikat, 2022) trained four- and five-layer CNNs on four waste types and reached at most 70% accuracy, which shows how quickly small models trained from scratch saturate. Transfer learning changed this picture. (Yong, Ma, Sun, & Du, 2023) reported that a pretrained MobileNetV2 improved accuracy by 15.42 percentage points over a conventional CNN, and (Poudel & Poudyal, 2022) fine-tuned seven ImageNet models on seven waste classes, with InceptionV3 achieving the highest accuracy. Later studies widened the comparison or combined models. (Liu et al., 2022) fused several transfer-learned networks for garbage recognition, (Gürcan & Soylu, 2026) compared pretrained CNNs on twelve household classes and obtained 96.95% with ConvNeXt, and (Alkılınç et al., 2026) reached 96% on TrashNet with a weighted ensemble. (Kumala et al., 2025) compared MobileNetV2 and VGG16 on ten classes and found the lighter model more accurate (95.60% against 89.13%). These studies agree that pretrained networks outperform models trained from scratch, but they differ in data, input size, and training settings, so their rankings of architectures are difficult to compare.

A large share of this work concerns the split between organic and recyclable waste. The Waste Classification Data published by Sekar on Kaggle, with about 25,000 web-collected images, is the most common source. (Nnamoko et al., 2022) restructured it by removing 372 images with an incompatible colour mode and reached 80.88% with a compact CNN, (Mehedi et al., 2023) obtained 96.00% with VGG16 and 95.51% with MobileNetV2 on the restructured version, and (Chhabra, Sharan, Elbarachi, & Kumar, 2024) proposed an improved deep CNN that reached 93.28% on 25,077 images and outperformed several transfer-learned baselines. Indonesian studies report similar binary results. (Aprisia Bahagia & Akbar, 2024) trained a CNN on 18,052 training and 4,512 test images and reached 95%, (Muslihati, Sahibu, & Taufik, 2024) reached 96.43% after six training trials, and (Oktayaessofa et al., 2024) obtained 98.47% with MobileNetV2 on 4,380 images. (Ibnul Rasidi, Pasaribu, Ziqri, & Adhinata, 2022), however, found a large gap between classes, with 96% accuracy for inorganic and 62% for organic waste. (Girsang et al., 2023) compared VGG16 and InceptionV3 on the same binary task and preferred VGG16. Electronic waste is absent from these binary datasets, although it requires its own handling (Shreya et al., 2023).

Explainability has recently become a common addition. Grad-CAM (Selvaraju et al., 2017) produces class-specific heat maps from the gradients of the last convolutional layer without modifying the network. (Manik et al., 2026) combined EfficientNet-B0 with Grad-CAM on six household classes and reported 94.3% mean accuracy in five-fold cross-validation, and (Alkılınç et al., 2026) used Grad-CAM to show that their ensemble attends to the object. In both cases the heat maps confirm correct behaviour rather than explain failures. (Shetty, Kallianpur, Fernandes, Rodrigues, & Padmanabha, 2025) also used Grad-CAM to interpret fine-tuned and hybrid CNN models on ten waste categories. Label quality, in contrast, has mostly been studied outside the waste domain. (Song, Kim, Park, Shin, & Lee, 2023) survey methods for training deep networks on noisy labels, (Goswami et al., 2023) provide a benchmark for comparing methods that detect mislabelled training data, and (Northcutt, Jiang, & Chuang, 2021) propose confident learning, which ranks examples by predicted probability to locate likely label errors. Applying such ideas to benchmark test sets, (Northcutt, Athalye, et al., 2021) estimated that at least 3.3% of the labels in ten popular datasets are wrong and that corrected labels can reverse model rankings. The present study connects these lines of work: it compares architectures under one protocol, measures their cost, and uses Grad-CAM together with a confidence-ranked audit to explain the errors that remain.

METHOD

Research Workflow

Figure 1 summarises the workflow. The same data split, preprocessing, and training recipe were applied to the three architectures, so that differences in the results can be attributed to the architectures. The best checkpoint of each model was evaluated for classification performance and computational cost, and the best model was then examined with Grad-CAM and a manual audit of its most confident errors.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9844

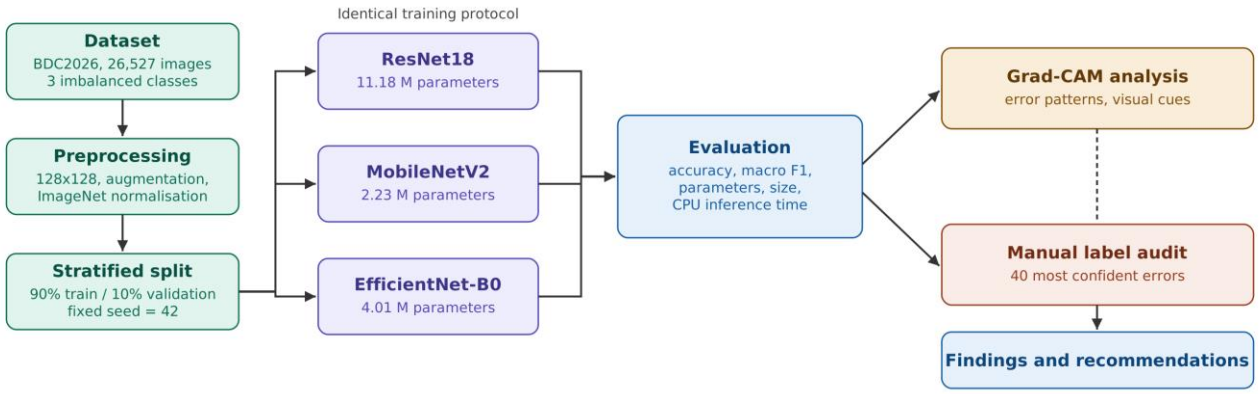


Figure 1. Research workflow of the comparison, explanation, and label audit

Dataset

The data are the official training images of the BDC2026 image classification task organised by Satria Data 2026. The training set contains 26,527 images in three classes: Recyclable (9,999), Electronic (3,961), and Organic (12,567). The competition test images are unlabelled, so all evaluation in this study uses a held-out validation split. The Recyclable class size equals, and the Organic class size nearly equals, the training split of the Waste Classification Data released by Sekar (9,999 recyclable and 12,565 organic images) and later restructured by (Nnamoko et al., 2022), which suggests that these two classes were at least partly drawn from that web-collected source. Fig. 2 shows examples of each class.



Figure 2. Sample training images of the Electronic, Organic, and Recyclable classes in the BDC2026 dataset

The data were split with a stratified 90:10 ratio and a fixed random seed (42), giving 23,876 training and 2,651 validation images (999 Recyclable, 396 Electronic, and 1,256 Organic). The same split was used in every experiment, including the Grad-CAM analysis and the label audit. Training images were resized and augmented as listed in Table 1, whereas validation images were only resized to 128 × 128 pixels and normalised with the ImageNet channel means and standard deviations.

Model Architectures

ResNet18 (He, Zhang, Ren, & Sun, 2016) uses residual connections to ease the training of deep networks. MobileNetV2 (Sandler et al., 2018) stacks depthwise separable convolutions in inverted residual blocks with linear bottlenecks, which reduces computation on mobile devices. EfficientNet-B0 (Tan & Le, 2019) is the base model of a family obtained by scaling network depth, width, and input resolution jointly. All three were initialised with ImageNet-1K weights and adapted through transfer learning (Zhuang et al., 2021). Partial fine-tuning was applied: all layers were frozen except the final feature stage, namely layer4 of ResNet18 and the last convolutional stage of the feature extractor of MobileNetV2 and EfficientNet-B0. The original classifier of each network was replaced by a dropout layer (rate 0.3) followed by a fully connected layer with three outputs. The output logits $z = (z_1, z_2, z_3)$ are converted into class probabilities by the softmax function in Equation (1), where $K = 3$ is the number of classes, and the predicted class is

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9844

the one with the highest probability.

$$p_c = \frac{e^{z_c}}{\sum_{k=1}^{K} e^{z_k}}, \qquad c = 1, \ldots, K \tag{1}$$

Training Protocol

Table 1 lists the training configuration, which was identical for the three architectures. Class imbalance was handled by class re-weighting (Zhang, Kang, Hooi, Yan, & Feng, 2023). The weight of class c was computed from the class counts of the full training set with Equation (2), where N is the number of images and n_c the number of images in class c, which gives weights of 0.884, 2.232, and 0.704 for Recyclable, Electronic, and Organic. The loss is the weighted cross-entropy in Equation (3), averaged over a mini-batch of B images as implemented in PyTorch, where y_i is the true label of image i.

$$w_c = \frac{N}{K\, n_c} \tag{2}$$

$$\mathcal{L}_{\mathrm{WCE}} = -\frac{\sum_{i=1}^{B} w_{y_i} \log p_{i,y_i}}{\sum_{i=1}^{B} w_{y_i}} \tag{3}$$

Table 1. Training Configuration Shared by the Three Architectures

Setting	Value
Input size	128 × 128 pixels
Augmentation (training only)	Resize to 160 × 160, random resized crop (scale 0.8–1.0), horizontal flip, colour jitter 0.2
Normalisation	ImageNet channel means and standard deviations
Pretrained weights	ImageNet-1K
Trainable layers	Final feature stage and new classifier
Classifier	Dropout (0.3) and fully connected layer with 3 outputs
Loss function	Class-weighted cross-entropy
Optimiser	Adam, learning rate 10^{-3}, weight decay 10^{-4}
Learning-rate schedule	Step decay every 7 epochs, factor 0.3
Batch size and maximum epochs	64 and 15
Early stopping	Patience of 4 epochs on validation accuracy

Parameters were updated with the Adam optimiser (Kingma & Ba, 2014). Weight decay was applied through the standard PyTorch implementation, in which the penalty $\lambda\theta$ is added to the gradient before the moment estimates are computed, as shown in Equations (4) to (6). In these equations η is the learning rate, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are the default decay rates of the moment estimates, and ε is a small constant for numerical stability.

$$g_t = \nabla_\theta \mathcal{L}(\theta_{t-1}) + \lambda\, \theta_{t-1} \tag{4}$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{5}$$

$$\theta_t = \theta_{t-1} - \eta\, \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}, \qquad \hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{6}$$

The learning rate followed the step schedule in Equation (7), where e is the zero-based epoch index, $s = 7$ is the step size, and $\gamma = 0.3$, so the rate fell from 10^{-3} to 3×10^{-4} after the seventh epoch.

$$\eta_e = \eta_0\, \gamma^{\lfloor e/s \rfloor} \tag{7}$$

Training ran for at most 15 epochs. Early stopping ended training when validation accuracy had not improved for four consecutive epochs, and the checkpoint with the highest validation accuracy was kept. All experiments were run with PyTorch on a laptop with an Intel Core i7-8565U processor (1.80 GHz) and 8 GB of memory. The training log reports that the computations ran on the processor, so the NVIDIA GeForce MX250 graphics card of the same machine was not used.

Evaluation Metrics

Performance was measured on the validation split. From the confusion matrix M, the true positives (TP_c), false positives (FP_c), and false negatives (FN_c) of each class were counted in a one-versus-rest manner, giving the precision P_c, recall R_c, and F1-score in Equation (8). Overall accuracy and the macro average of a per-class metric X_c, which weights each class equally regardless of its size, follow Equation (9). Computational cost was described by the number

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9844

of parameters, the size of the saved checkpoint, and the inference time per image, measured as the mean wall-clock time of a forward pass over 20 validation images processed one at a time with the best checkpoint.

$$P_c = \frac{TP_c}{TP_c+FP_c}, \qquad R_c = \frac{TP_c}{TP_c+FN_c}, \qquad F1_c = \frac{2P_cR_c}{P_c+R_c} \tag{8}$$

$$\text{Accuracy} = \frac{\sum_{c=1}^{K} M_{cc}}{\sum_{i=1}^{K}\sum_{j=1}^{K} M_{ij}}, \qquad \overline{X}_{\text{macro}} = \frac{1}{K}\sum_{c=1}^{K} X_c \tag{9}$$

Grad-CAM Analysis

Grad-CAM (Selvaraju et al., 2017) was applied to the final feature stage of the best model. For the predicted class c, the importance weight of feature map A^k is the spatial average of the gradient of the class logit y^c in Equation (10), where Z is the number of spatial positions, and the class activation map is the rectified weighted sum of the feature maps in Equation (11). Each map was normalised to [0, 1], resized to the input resolution, and overlaid on the image with 45% opacity. The analysis targeted misclassifications with a predicted probability of at least 0.90 in the two dominant error directions, Recyclable predicted as Organic and Organic predicted as Recyclable.

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{10}$$

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

Label Audit

The audit follows the idea behind confident learning (Northcutt, Jiang, et al., 2021): a confident disagreement between a model and a label points to a possible label error. All misclassified validation images were ranked by the probability of the predicted class, and the 40 most confident errors (probabilities between 0.90 and 1.00) were inspected visually, regardless of class pair. One rater assigned each image to one of three verdicts: model error, when the label is correct and the prediction is wrong; probable label error, when the prediction matches the visible object while the label does not; and ambiguous, when neither could be established with confidence. The verdicts were tabulated overall and by error direction.

RESULT

Model Comparison

Table 2 summarises the performance and cost of the three models on the validation split. MobileNetV2 obtained the highest accuracy (92.68%), macro precision (0.9170), macro recall (0.9319), and macro F1-score (0.9239). It was also the smallest model, with 2.23 million parameters and a checkpoint of 8.71 MB, about one fifth of the size of ResNet18, and the fastest, at 16.78 ms per image. ResNet18 ranked second in accuracy (92.34%) but was the largest model, whereas EfficientNet-B0 had the lowest accuracy (91.97%) and the longest inference time (25.60 ms). The accuracy of the three models spanned 0.71 percentage points.

Table 2. Performance and Computational Cost of the Three Architectures on the Validation Split

Architecture	Param (M)	Size (MB)	Epochs run	Best epoch	Accuracy (%)	Macro P	Macro R	Macro F1	Inference (ms)
ResNet18	11.18	42.71	6	2	92.34	0.9109	0.9220	0.9162	23.82
MobileNetV2	2.23	8.71	13	9	92.68	0.9170	0.9319	0.9239	16.78
EfficientNet-B0	4.01	15.56	15	15	91.97	0.9064	0.9273	0.9158	25.60

Figure 3 shows the training curves. ResNet18 reached its best validation accuracy at epoch 2 and stopped at epoch 6, before the first learning-rate reduction after epoch 7. MobileNetV2 and EfficientNet-B0 both improved visibly after the reduction. MobileNetV2 peaked at epoch 9 and stopped at epoch 13, whereas EfficientNet-B0 was still improving slightly when it reached the 15-epoch limit. In all three models the training loss kept decreasing while the validation loss levelled off, which indicates the onset of overfitting.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9844

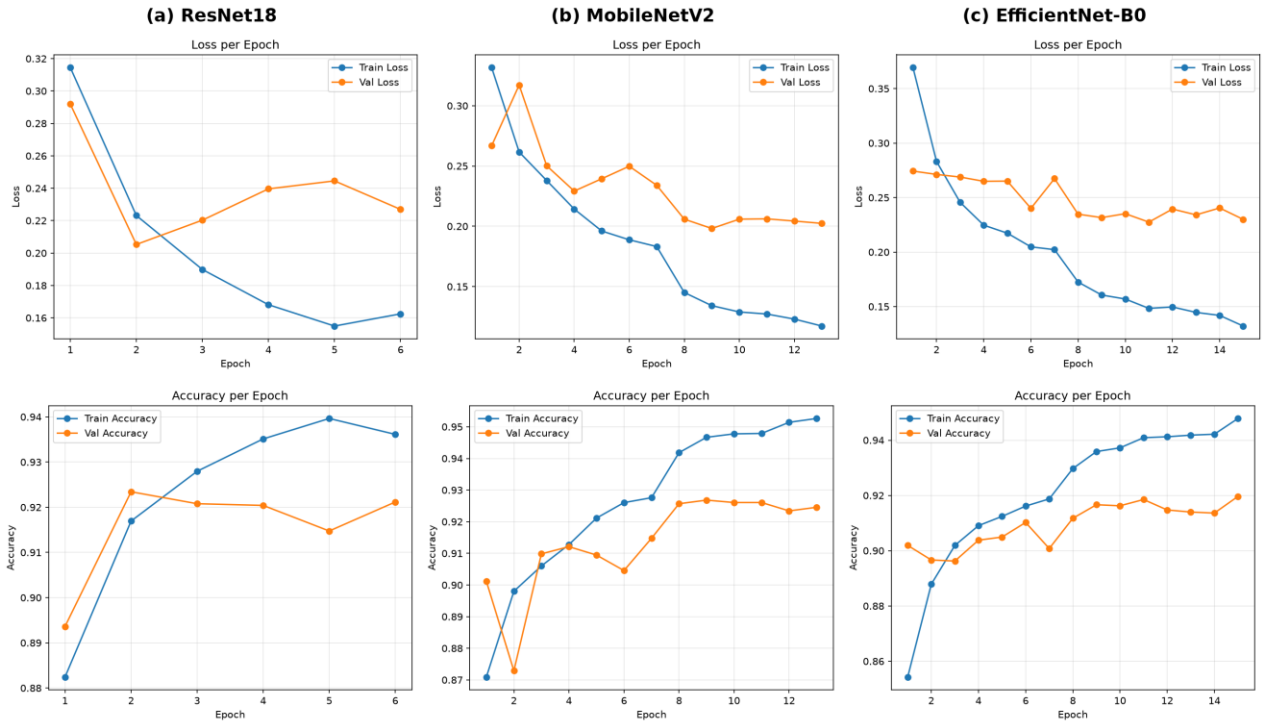


Figure 3. Training and validation loss (top) and accuracy (bottom) per epoch for (a) ResNet18, (b) MobileNetV2, and (c) EfficientNet-B0

Error Patterns

The confusion matrix of MobileNetV2 (Fig. 4) shows that 127 of its 194 errors (65.5%) were confusions between Recyclable and Organic: 57 Recyclable images were predicted as Organic and 70 Organic images as Recyclable. Electronic had the highest recall (95.45%) but the lowest precision (88.52%), mainly because 35 Recyclable images were predicted as Electronic, while Recyclable had the lowest recall (90.79%). Table 3 compares this confusion across the three architectures. The Recyclable–Organic pair accounted for 118 of the 203 errors of ResNet18 (58.1%), 127 of 194 for MobileNetV2 (65.5%), and 135 of 213 for EfficientNet-B0 (63.4%), so it was the largest source of error in every model. The direction differed slightly: ResNet18 and MobileNetV2 more often predicted Organic images as Recyclable (67 versus 51 and 70 versus 57 images), whereas EfficientNet-B0 was almost balanced (69 and 66 images). The remaining errors (85, 67, and 78 images, respectively) involved the Electronic class. Because the same pair dominates in all three models, the confusion is not specific to one architecture, which motivates the Grad-CAM analysis and the label audit below.

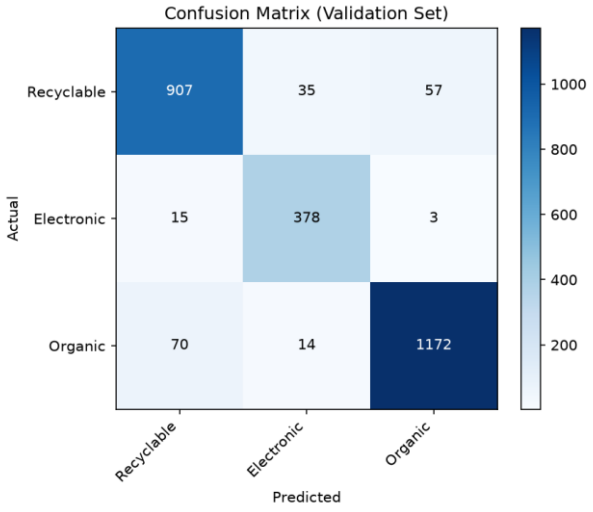


Figure 4. Confusion matrix of MobileNetV2 on the validation split (rows: true class, columns: predicted class)

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9844

Table 3. Recyclable–Organic Confusions per Architecture on the Validation Split

Architecture	Recyclable predicted as Organic	Organic predicted as Recyclable	Total
ResNet18	51	67	118
MobileNetV2	57	70	127
EfficientNet-B0	69	66	135

Figure 5 shows Grad-CAM maps for six confident errors in each direction. For Organic images predicted as Recyclable, the activations covered glossy, transparent, or geometric surfaces, and several images showed packaging or containers, such as a glass jug, a plastic wrapper, a perforated plastic basket, and a printed snack carton. For Recyclable images predicted as Organic, the activations covered fibrous, rough, or warm-coloured textures, and the images included a woven bag, a flower, moss-like surfaces, and vegetation.

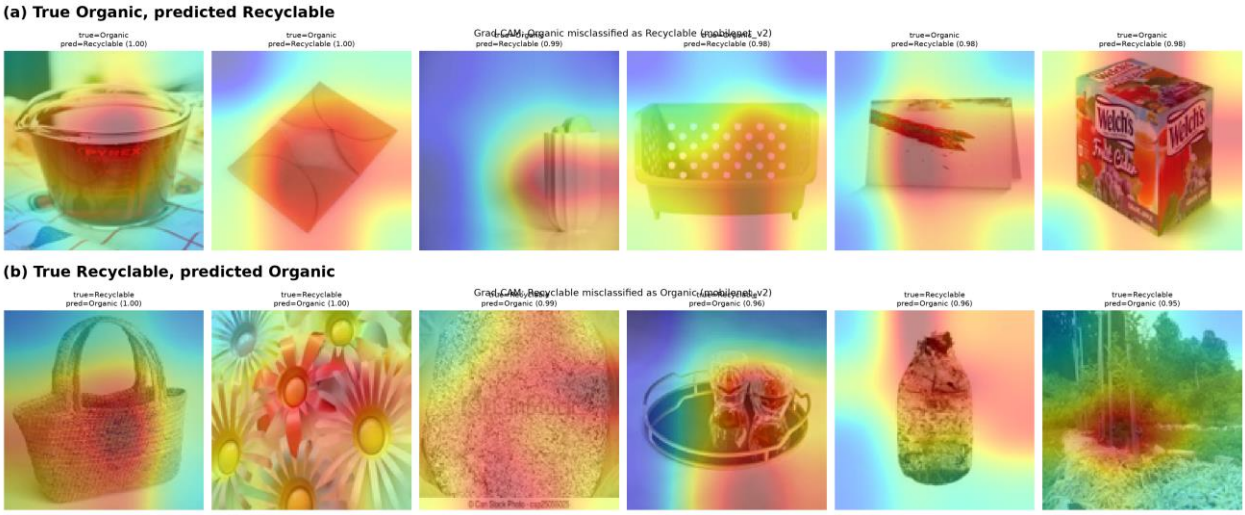


Figure 5. Grad-CAM maps of MobileNetV2 for confident errors: (a) Organic predicted as Recyclable and (b) Recyclable predicted as Organic

Label Audit

Table 4 reports the audit of the 40 most confident errors of MobileNetV2. Fourteen errors (35.0%) were model errors, 14 (35.0%) were ambiguous, and 12 (30.0%) were probable label errors. Table 5 breaks the verdicts down by error direction. For Organic images predicted as Recyclable, 8 of 17 cases (47.1%) were probable label errors and only 4 (23.5%) were clear model errors. For Recyclable images predicted as Organic the pattern was reversed: 6 of 14 cases (42.9%) were model errors and 2 (14.3%) were probable label errors. Recyclable images predicted as Electronic accounted for six of the audited errors.

Table 4. Audit Verdicts for the 40 Most Confident Errors of MobileNetV2

Verdict	Count	Percentage (%)
Model error (label correct)	14	35.0
Ambiguous	14	35.0
Probable label error	12	30.0
Total	40	100.0

Table 5. Audit Verdicts by Error Direction

True class → predicted class	n	Model error	Ambiguous	Probable label error
Organic → Recyclable	17	4 (23.5%)	5 (29.4%)	8 (47.1%)
Recyclable → Organic	14	6 (42.9%)	6 (42.9%)	2 (14.3%)
Recyclable → Electronic	6	2 (33.3%)	3 (50.0%)	1 (16.7%)
Organic → Electronic	2	1	0	1
Electronic → Recyclable	1	1	0	0

DISCUSSION

MobileNetV2 was the most accurate and the cheapest model at the same time, so on this dataset there was no trade-off between accuracy and cost. Its efficiency follows from its design: depthwise separable convolutions in inverted residual blocks need far fewer parameters and operations than the standard convolutions of ResNet18 (Sandler

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9844

et al., 2018), and on a processor without graphics acceleration this translates directly into shorter inference time. EfficientNet-B0 has fewer parameters than ResNet18 but was the slowest model, which suggests that its deeper sequence of blocks costs more time than its parameter count implies. The accuracy advantage of MobileNetV2 is, however, small (0.34 percentage points over ResNet18) and comes from a single run on one split, so it should not be read as a decisive ranking. The training curves also suggest that the shared protocol disadvantaged ResNet18. It stopped at epoch 6, before the learning-rate reduction that produced clear gains for the other two models (Fig. 3), so the early-stopping patience interacted with the step schedule. In addition, partial fine-tuning unfreezes a much larger final stage in ResNet18 than in the two lightweight models, which makes overfitting at the initial learning rate more likely.

The low precision for Electronic is consistent with class re-weighting. The Electronic weight (2.232) is more than three times the Organic weight, which pushes uncertain images towards the minority class; this side effect of re-weighting is discussed in the long-tailed learning literature (Zhang et al., 2023). For a smart bin, this behaviour may be acceptable, because sending a recyclable item to the electronic compartment is less harmful than sending an electronic item to the recyclable stream.

Table 6 places the results next to earlier studies. The figures are not directly comparable, because the datasets, numbers of classes, input sizes, and evaluation protocols differ. Even so, the accuracy obtained here lies within the range reported for transfer-learned models and well above the CNNs trained from scratch by (Altikat et al., 2022; Nnamoko et al., 2022). The closest references are the binary studies on the Sekar data, from which the Recyclable and Organic classes here appear to derive; they report between 93.28% and 96.00% (Aprisia Bahagia & Akbar, 2024; Chhabra et al., 2024; Mehedi et al., 2023). The lower accuracy in the present study plausibly reflects the additional Electronic class, the small input resolution of 128 × 128 pixels compared with the 224 × 224 pixels for which these backbones were pretrained, the partial fine-tuning, and label noise.

Table 6. Accuracy Reported by Related Waste Classification Studies

Study	Classes	Model	Accuracy (%)
Altikat et al. (2022)	4	Five-layer CNN (from scratch)	70.00
Nnamoko et al. (2022)	2	Compact CNN (from scratch)	80.88
Yong et al. (2023)	4	MobileNetV2	82.92
Mehedi et al. (2023)	2	VGG16 / MobileNetV2	96.00 / 95.51
Aprisia Bahagia and Akbar (2024)	2	CNN	95.00
Chhabra et al. (2024)	2	Improved deep CNN	93.28
Muslihati et al. (2024)	2	CNN	96.43
Oktayaessofa et al. (2024)	2	MobileNetV2	98.47
Kumala et al. (2025)	10	VGG16 / MobileNetV2	89.13 / 95.60
Manik et al. (2026)	6	EfficientNet-B0	94.30[a]
Gürcan and Soylu (2026)	12	ConvNeXt	96.95
Alkılınç et al. (2026)	6 (TrashNet)	Weighted ensemble	96.00
This study	3	ResNet18 / MobileNetV2 / EfficientNet-B0	92.34 / 92.68 / 91.97

Two-class studies separate organic from recyclable or non-organic waste. [a] Mean accuracy of five-fold cross-validation.

The Grad-CAM maps indicate that the dominant errors are driven by surface appearance rather than by object identity. With partial fine-tuning, most of the network keeps the features learned on ImageNet and only the last stage adapts to waste, so mid-level texture and colour features carry much of the decision. This explains the genuine errors in the Recyclable-to-Organic direction, where a woven bag or a moss-covered bottle resembles plant material. In the opposite direction the model often reacted to glossy or geometric surfaces of containers and packaging, which are, materially, recyclable, so many of these predictions are defensible.

The audit confirms that a substantial part of the remaining error is a property of the data rather than of the model. Thirty percent of the most confident errors were probable label errors, and the share rose to 47.1% for Organic images predicted as Recyclable, many of which show food packaging or containers. A plausible explanation is that web searches for food-related terms return images of packaged food, which were then filed as organic, a pattern that matches the web-collected origin of the source data (Nnamoko et al., 2022). The size of the effect matters. Relabelling the 12 probable label errors alone would raise the measured accuracy of MobileNetV2 from 92.68% to at most 93.13%, and to at most 93.66% if the 14 ambiguous cases were also resolved in favour of the model, while the 154 unaudited errors may contain further label errors. These shifts are of the same order as the 0.71 percentage-point spread between the three architectures, so, in line with (Northcutt, Athalye, et al., 2021), the ranking of models on this dataset is sensitive to label quality. The 30% figure should not be compared with the 3.3% average label-error rate reported in that study, because it refers only to the most confident errors and not to the whole validation set.

The efficiency results support MobileNetV2 as the first candidate for a smart bin, but they do not yet demonstrate deployment. Inference time was measured on a laptop processor rather than on a single-board computer, and an 8.71 MB floating-point model exceeds the memory of typical microcontrollers. A realistic design would run the classifier on

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9844

a single-board computer such as a Raspberry Pi, possibly after quantisation, and use a microcontroller only to drive the sorting actuators.

The study has several limitations. The audit was performed by one rater without an inter-rater reliability check, and it covered only the 40 most confident of 194 errors. Each model was trained once on one split, and the validation split served both for checkpoint selection and for reporting, which makes the reported accuracies optimistic; the competition test labels were not available for an independent estimate. The hyperparameters were shared rather than tuned per architecture, and partial fine-tuning unfroze different numbers of parameters in each network.

CONCLUSION

This study compared ResNet18, MobileNetV2, and EfficientNet-B0 for three-class waste image classification under an identical transfer-learning protocol. In answer to the first research question, MobileNetV2 offered the best balance: it was the most accurate model (92.68% accuracy and 0.9239 macro F1-score) and also the smallest and fastest, which makes it the most suitable starting point for a resource-constrained smart bin. In answer to the second, all three architectures confused recyclable and organic waste most often, and Grad-CAM showed that these errors are driven by surface texture and colour rather than by object identity. In answer to the third, 30% of the most confident errors were probable label errors, concentrated in organic images that show packaging, and correcting them could shift accuracy by an amount similar to the gap between the architectures. The main contribution is therefore methodological: waste classification studies should report computational cost and audit label quality alongside accuracy, because both can change which model is preferred.

Future work should clean the labels with semi-automatic methods such as confident learning, repeat the comparison over several seeds or cross-validation folds with per-architecture tuning and an independent labelled test set, and validate deployment on a physical prototype. In such a prototype, a quantised MobileNetV2 would run on a single-board computer with a camera while an Arduino microcontroller drives the sorting mechanism, so that latency, on-device preprocessing, and actuator control can be measured under real-time conditions.

ACKNOWLEDGMENT

The authors thank the organising committee of Satria Data 2026 for providing the Big Data Challenge dataset used in this study.

REFERENCES

Alkılınç, A., Okay, F. Y., Kök, İ., & Özdemir, S. (2026). Deep ensemble learning model for waste classification systems. *Sustainability*, *18*(1), 24. https://doi.org/10.3390/su18010024

Altikat, A., Gulbe, A., & Altikat, S. (2022). Intelligent solid waste classification using deep convolutional neural networks. *International Journal of Environmental Science and Technology*, *19*(3), 1285–1292. https://doi.org/10.1007/s13762-021-03179-4

Aprisia Bahagia, G., & Akbar, M. (2024). Klasifikasi sampah organik dan anorganik menggunakan metode convolutional neural network (CNN) [Classification of organic and inorganic waste using the convolutional neural network method]. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(5), 10349–10355. https://doi.org/10.36040/jati.v8i5.11038

Chhabra, M., Sharan, B., Elbarachi, M., & Kumar, M. (2024). Intelligent waste classification approach based on improved multi-layered convolutional neural network. *Multimedia Tools and Applications*, *83*(36), 84095–84120. https://doi.org/10.1007/s11042-024-18939-w

Girsang, A. S., Saputra, A. D., & Yanrie, V. (2023). Performance comparison between VGG16 and Inception V3 for organic waste and recyclable waste classification. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(2), 557–563.

Goswami, M., Sanil, V., Choudhry, A., Srinivasan, A., Udompanyawit, C., & Dubrawski, A. (2023). *AQuA: A benchmarking tool for label quality assessment*. arXiv. https://doi.org/10.48550/arXiv.2306.09467

Gürcan, F., & Soylu, A. (2026). Automated waste classification for smart recycling: A multi-class CNN approach with transfer learning and pre-trained models. *Environmental Technology & Innovation*, *41*, 104673. https://doi.org/10.1016/j.eti.2025.104673

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 770–778). Las Vegas, NV, USA: IEEE. https://doi.org/10.1109/CVPR.2016.90

Ibnul Rasidi, A., Pasaribu, Y. A. H., Ziqri, A., & Adhinata, F. D. (2022). Klasifikasi sampah organik dan non-organik menggunakan convolutional neural network [Classification of organic and non-organic waste using a convolutional neural network]. *Jurnal Teknik Informatika Dan Sistem Informasi*, *8*(1). https://doi.org/10.28932/jutisi.v8i1.4314

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9844

https://doi.org/10.48550/ARXIV.1412.6980

Kumala, R. A., Sari, C. A., & Rachmawanto, E. H. (2025). A comparison of MobileNetV2 and VGG16 architectures with transfer learning for multi-class image-based waste classification. *Journal of Applied Informatics and Computing*, *9*(4), 1610–1624. https://doi.org/10.30871/jaic.v9i4.9958

Liu, W., Ouyang, H., Liu, Q., Cai, S., Wang, C., Xie, J., & Hu, W. (2022). Image recognition for garbage classification based on transfer learning and model fusion. *Mathematical Problems in Engineering*, *2022*, 1–12. https://doi.org/10.1155/2022/4793555

Manik, F. Y., Nainggolan, P. I., Harumy, T. H. F., Ginting, D. S. B., Maharani, A., Ramadayanti, H., … Harifin, M. P. (2026). Explainable CNN-based multiclass household waste classification using Grad-CAM for smart waste management. *International Journal of Advanced Computer Science and Applications*, *17*(1), 580–589. https://doi.org/10.14569/IJACSA.2026.0170155

Mehedi, M. H. K., Arafin, I., Hasan, M., Rahman, F., Tasin, R., & Rasel, A. A. (2023). A transfer learning approach for efficient classification of waste materials. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, (pp. 636–640). Las Vegas, NV, USA: IEEE. https://doi.org/10.1109/CCWC57344.2023.10099127

Muslihati, M., Sahibu, S., & Taufik, I. (2024). Implementasi algoritma convolutional neural network untuk klasifikasi jenis sampah organik dan non organik [Implementation of the convolutional neural network algorithm for classifying types of organic and non-organic waste]. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(3), 840–852. https://doi.org/10.57152/malcom.v4i3.1346

Nnamoko, N., Barrowclough, J., & Procter, J. (2022). Solid waste image classification using deep convolutional neural network. *Infrastructures*, *7*(4), 47. https://doi.org/10.3390/infrastructures7040047

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Retrieved from https://openreview.net/forum?id=XccDXrDNLek

Northcutt, C. G., Jiang, L., & Chuang, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, *70*, 1373–1411. https://doi.org/10.1613/jair.1.12125

Oktayaessofa, E., Sari, C. A., Rachmawanto, E. H., & Yaacob, N. M. (2024). Classification of organic and non-organic waste with CNN-MobileNet-V2. *Jurnal Teknik Informatika (JUTIF)*, *5*(4), 1173–1180. https://doi.org/10.52436/1.jutif.2024.5.4.2165

Poudel, S., & Poudyal, P. (2022). Classification of waste materials using CNN based on transfer learning. IN *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, (pp. 29–33). Kolkata, India: ACM. https://doi.org/10.1145/3574318.3574345

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 4510–4520). Salt Lake City, UT: IEEE. https://doi.org/10.1109/CVPR.2018.00474

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, (pp. 618–626). Venice: IEEE. https://doi.org/10.1109/ICCV.2017.74

Shetty, S., Kallianpur, S., Fernandes, R., Rodrigues, A. P., & Padmanabha, V. (2025). ECO-HYBRID: Sustainable waste classification using transfer learning with hybrid and enhanced CNN models. *Sustainability*, *17*(19), 8761. https://doi.org/10.3390/su17198761

Shreya, M., Nimal Yughan, V., Katyal, J., & Ramesh, R. (2023). Technical solutions for waste classification and management: A mini-review. *Waste Management & Research: The Journal for a Sustainable Circular Economy*, *41*(4), 801–815. https://doi.org/10.1177/0734242X221135262

Song, H., Kim, M., Park, D., Shin, Y., & Lee, J.-G. (2023). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(11), 8135–8153. https://doi.org/10.1109/TNNLS.2022.3152527

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114. PMLR. Retrieved from https://proceedings.mlr.press/v97/tan19a.html

Yong, L., Ma, L., Sun, D., & Du, L. (2023). Application of MobileNetV2 to waste classification. *PLOS ONE*, *18*(3), e0282336. https://doi.org/10.1371/journal.pone.0282336

Zhang, Y., Kang, B., Hooi, B., Yan, S., & Feng, J. (2023). Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10795–10816. https://doi.org/10.1109/TPAMI.2023.3268118

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., … He, Q. (2021). A comprehensive survey on transfer learning. In *Proceedings of the IEEE*, *109*(1), (pp. 43–76). https://doi.org/10.1109/JPROC.2020.3004555

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.