Explainable CNN-Based Multiclass Household Waste Classification Using Grad-CAM for Smart Waste Management

Fuzy Yustika Manik*, Pauzi Ibrahim Nainggolan, T.H.F Harumy, Dewi Sartika Br Ginting, Aini Maharani, Hafizha Ramadayanti, Jessica Almalia, Muhammad Putra Harifin
Faculty of Computer Science and Information Technology-Department of Computer Science, Universitas Sumatera Utara, Medan, Indonesia

***Abstract*—Automated waste classification using computer vision has become essential for improving environmental sustainability and reducing manual sorting effort. This study presents an enhanced waste image classification model based on EfficientNet-B0, trained using a two-stage transfer learning strategy that combines feature extraction and fine-tuning. The proposed approach aims to enhance classification accuracy while maintaining computational efficiency. Experimental evaluations conducted on a heterogeneous multi-class waste dataset demonstrate the superiority of the proposed method. The confusion matrix results indicate a high proportion of correct predictions across most categories, with only minor misclassifications among visually similar classes, such as metal and paper. The model's robustness is further validated through 5-Fold Cross-Validation, which yields an average accuracy of 94.3% with a standard deviation of ±0.007, confirming consistent performance across data partitions. Compared with state-of-the-art CNN architectures, including ResNet50 and DenseNet121, the proposed model achieves the highest accuracy while using the fewest parameters (4.38M), making it suitable for deployment in resource-constrained environments. Additionally, qualitative analysis using Grad-CAM confirms that the model's decisions are explainable and based on relevant object features. These findings demonstrate that the proposed EfficientNet-B0 model constitutes a reliable, efficient, and interpretable solution for automated waste classification. The model is further evaluated using cross-validation and explainable AI (Grad-CAM) to assess both performance stability and interpretability.**

Keywords*—*EfficientNet-B0; explainable AI; Grad-CAM; transfer learning; waste classification

I. Introduction

The challenges of household waste management have become increasingly urgent due to population growth, urbanization, and changes in public consumption patterns [1]. The continuous rise in global waste generation has placed significant pressure on modern waste management systems [2]. International reports indicate that municipal solid waste (MSW) production continues to increase annually and, if poorly managed, leads to soil, water, and air pollution, ultimately reducing the quality of life in both urban and peri-urban regions [3]. Traditional waste sorting practices still rely heavily on manual labor—a process that is time-consuming, costly, and prone to human error, exposing workers to hazardous materials and producing inconsistent sorting results [4]. These inefficiencies contribute directly to low recycling rates and the accumulation of improperly sorted waste in landfills, thereby hindering sustainable waste management efforts [5]. Household-based initiatives, such as waste banks, also face challenges related to limited labor capacity, inconsistent waste-sorting behavior, and low public awareness [6]. Therefore, researchers have emphasized the urgent need to integrate technological advancements with community participation to improve waste handling efficiency. To address the limitations of manual sorting, recent studies have explored automated waste classification systems. Early approaches employed conventional machine learning techniques such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN). However, these models demonstrated limited adaptability to the complex, heterogeneous visual features found in waste images [7]. The subtle inter-class visual similarities—for example, between plastic and glass, or cardboard and paper—make automated waste classification particularly challenging, as minor errors can propagate through subsequent waste processing stages [8-9]. This limitation underscores the importance of high classification accuracy, as such systems often function as the foundational layer of an automated waste management pipeline.

The adoption of deep learning, particularly Convolutional Neural Networks (CNN), has significantly improved waste classification performance. Deep CNN architectures have demonstrated the ability to extract hierarchical visual patterns and deliver superior classification accuracy compared to traditional methods [10]. Several models have been applied to waste classification tasks, achieving promising results and enabling more reliable, scalable sorting pipelines [11]. Castro-Bello et al. expanded the application of CNNs for multiclass MSW classification within sustainable waste management frameworks [12]. More recent research has evaluated multiple architectures—such as VGG16, ResNet50, DenseNet, and EfficientNet [11]. Among these, EfficientNet is recognized for its compound scaling mechanism, which offers an optimal balance between accuracy and computational efficiency [13]. Empirical studies have shown that EfficientNet-B0 achieves strong performance in waste classification tasks with fewer parameters and lower computational costs, making it attractive for real-world deployment [8-9][14]. These developments confirm that deep learning has significant potential to advance automated waste management systems.

*Corresponding author.

Despite these advancements, several research gaps remain. Most existing studies prioritize improving global accuracy metrics without conducting detailed per-class performance analysis, resulting in limited insight into model weaknesses, especially for visually similar classes [8][10]. High-performing CNN architectures also tend to require substantial computational resources, making them less feasible for resource-constrained or real-time implementations [15-16]. Furthermore, the interpretability of CNN-based models remains limited, as most operate as black-box systems that do not reveal which regions of an image contribute to classification decisions [17–19]. Although explainability frameworks such as Gradient-weighted Class Activation Mapping (Grad-CAM) have been introduced, their integration is often supplementary and rarely combined with systematic per-class evaluation [20]. These shortcomings highlight the need for models that are not only accurate and computationally efficient but also interpretable and capable of revealing class-specific behavior.

Based on these gaps, this study proposes a waste image classification framework using EfficientNet-B0, enhanced with per-class performance evaluation and Grad-CAM-based visual explainability [21-22]. EfficientNet-B0 is selected for its lightweight architecture and strong accuracy-to-parameter efficiency, making it suitable for practical deployment in real-world environments [9] [13]. By integrating deep learning–based classification with interpretability mechanisms, this research aims to deliver both quantitative performance gains and qualitative insights into model decision-making.

The objectives of this research are as follows:

1) To develop an EfficientNet-B0-based waste classification model capable of achieving high accuracy with efficient computational cost.

2) To evaluate the model using per-class metrics (precision, recall, and F1 score) to identify low-performing waste categories and quantify the risk of misclassification, comparing it to the RestNet and DenseNet121 baselines.

3) To integrate Grad-CAM to visualize and interpret the model's attention regions, enhancing transparency and trustworthiness.

4) To validate the model's generalization capability through cross-validation and testing on diverse waste image samples that reflect real household conditions.

Thus, this study distinguishes itself from previous work by combining high-accuracy classification with interpretable visual explanations and class-level performance profiling, offering a more comprehensive, practically oriented contribution to sustainable waste management systems. This study makes four main contributions. First, it proposes a systematic two-stage transfer learning strategy (controlled feature extraction followed by selective fine-tuning) and examines its potential to improve generalization stability through baseline comparison. Second, the study goes beyond overall accuracy by incorporating class-level analysis using per-class precision, recall, F1-score, and confusion matrix interpretation to better understand misclassification patterns. Third, the model is evaluated using deployment-relevant criteria, including parameter efficiency and performance stability across 5-fold cross-validation. Finally, Grad-CAM is used not merely for visualization, but as a tool to examine whether the model focuses on semantically meaningful visual regions, supporting a more transparent assessment of model behavior.

It should be clarified that this study does not propose a novel learning algorithm. Instead, the contribution of this work lies in the systematic integration and comprehensive evaluation of established techniques, including transfer learning, two-stage training, class weighting, cross-validation, and explainability, to construct a reliable, interpretable, and practically deployable waste classification framework.

II. Research Methodology

A. System Design

This study develops a deep learning-based household waste image classification system, which aims to automatically and accurately identify waste material types. The primary architecture used is EfficientNet-B0, chosen for its efficient parameter management and its ability to achieve high accuracy through a compound scaling approach that simultaneously balances network depth, width, and resolution [13]. In this study, the system is also equipped with an Explainable Artificial Intelligence (XAI) mechanism using Grad-CAM, which highlights the image regions of interest to the model during prediction, ensuring that classification decisions are traceable and do not operate as a black box.

In addition to using EfficientNet-B0 as the primary model, this study compares it with two other architectures, ResNet50 and DenseNet121, to validate its superior performance. These two models were chosen because they are modern CNN architectures widely used in image classification, making them relevant benchmarks for evaluating the effectiveness of the proposed architecture. The waste image dataset used in this study was obtained from two publicly available data sources: the TrashNet Dataset and the Garbage Classification Dataset from Kaggle. These datasets were selected because they encompass a diverse range of household waste categories that align with the objectives of this research and have been widely used in previous deep-learning-based waste classification studies, enabling an objective comparison of model performance. All data used in this work are open access and freely downloadable, ensuring that this study complies with the reproducibility and traceability requirements of modern scientific publications.

The combined dataset consists of 3,600 images with varying resolutions and lighting conditions. The data are grouped into six common categories of domestic waste, namely organic (750 images), plastic (700 images), paper (620 images), glass (520 images), metal (510 images), and cardboard (500 images). The distribution indicates a moderate class imbalance, with the largest ratio occurring between the organic and cardboard categories at 1.5:1. Such an imbalance may introduce bias during model training, as the model tends to learn more often from classes with more samples. To address this issue, class weighting is applied during training, ensuring that each class contributes proportionally to the learning process and preventing distortion in prediction accuracy. Before being processed by EfficientNet-B0, all images underwent several preprocessing steps, including resizing to 224 × 224 pixels to match

EfficientNet's standard input dimensions, pixel value normalization to the range [0–1] to improve weight update stability during training, and conversion of class labels into one-hot encoding to support the multiclass classification scheme. The dataset was then partitioned using stratified splitting into 70% for training, 20% for validation, and 10% for testing. Stratification was chosen to maintain proportional class distribution within each subset, ensuring consistent data representation throughout training, validation, and evaluation. An example of the dataset used can be seen in Fig. 1 below:

Fig. 1. Example images from the TrashNet dataset and the garbage classification dataset from Kaggle.

With these characteristics, the dataset used in this study is not only relevant to the household waste segregation context but also sufficiently complex, making it a valid benchmark for evaluating the performance of the EfficientNet-B0 classification model enriched with explainable AI support. Augmentation was then applied to increase visual variation and prevent overfitting. The augmentation techniques applied included: rotation (0–20°), 10% zoom, horizontal flip, width shift, height shift, and rescaling. Transformations were applied randomly to each training batch, making the model more robust to real-world conditions, such as differences in viewpoint, lighting, and background.

B. Training Strategy

The proposed training procedure employs a two-stage learning strategy to ensure model stability, controlled parameter adaptation, and improved generalization performance. This approach is particularly effective for transfer learning architectures such as EfficientNet-B0, where pretrained knowledge must be retained while enabling domain-specific feature refinement.

Stage 1: Feature Extraction

In the first stage, all convolutional layers of EfficientNet-B0 are frozen, allowing the network to function solely as a feature extractor. Only the classification head is trained during this phase.

Hyperparameter settings:

- Optimizer: Adam
- Learning Rate (LR): 1×10^{-3}
- Epochs: 15
- Batch Size: 32

Objective:

This step aligns the classification layers with the dataset characteristics while preserving the general visual representations learned from ImageNet. By preventing abrupt weight modifications, the model avoids catastrophic forgetting and achieves a stable initialization prior to fine-tuning.

Stage 2: Fine-Tuning

Once the classifier head has converged, the final 20 layers of EfficientNet-B0 are unfrozen to allow deeper adaptation. During this stage, both the backbone and classification layers are jointly optimized.

Hyperparameter settings:

- Optimizer: Adam
- Learning Rate (LR): 1×10^{-5} (reduced to avoid destabilizing pretrained weights)
- Additional Epochs: 15

Objective:

This stage progressively refines high-level feature representations, improving the model's sensitivity to subtle intra-class variations and complex visual patterns. As a result, the model acquires domain-specific discriminative characteristics while maintaining pretrained robustness.

Callback Mechanism

To improve training efficiency and prevent overfitting, three callback functions are used. These functions are shown in Table I below:

TABLE I. CALLBACK MECHANISM

Callback	Purpose
EarlyStopping *(patience = 5)*	Stops training when validation performance does not improve
ReduceLROnPlateau	Automatically decreases LR when stagnation is detected
ModelCheckpoint	Stores the best-performing weights during training

These mechanisms ensure a stable optimization trajectory and optimal convergence behavior throughout both stages.

C. Performance Evaluation Method

The model's performance was evaluated comprehensively using several metrics to assess accuracy, consistency, and generalization capability. The evaluation does not solely focus on global accuracy but also considers per-class prediction quality, training stability, and objective comparisons with baseline architectures.

1) Primary metrics: The evaluation includes Accuracy as an indicator of overall model performance, along with Precision, Recall, and F1-Score to measure class-wise prediction quality, particularly important for imbalanced datasets. A Confusion Matrix and Classification Report are utilized to visualize misclassification patterns and provide numerical summaries for each category. Additionally, a

Generalization Test is performed on unseen test data to ensure the model can effectively recognize new samples beyond those used during training.

2) k-fold cross-validation: To assess model stability, k-Fold Cross-Validation is employed, in which the dataset is partitioned into k subsets and training is conducted iteratively so that each subset serves as a test fold once. Final performance metrics are reported as the mean and standard deviation across folds, indicating the model's consistency and verifying that its performance is not dependent on a specific data split.

3) Baseline evaluation: The proposed model's performance is compared against well-established CNN architectures, including VGG16, ResNet50, and the pretrained EfficientNet-B0. The comparison employs identical evaluation metrics to ensure that any observed improvements are attributable to the proposed two-stage training strategy. The model is considered superior if it achieves higher accuracy and F1-Score, exhibits fewer misclassifications, and records a lower standard deviation across k-fold evaluations, indicating more stable learning behavior.

D. Explainability with Grad-CAM

To ensure that the model's decision-making process can be interpreted transparently, this study employs the Gradient-weighted Class Activation Mapping (Grad-CAM) method as an explainable AI approach. Grad-CAM is utilized to trace regions within an image that contribute most significantly to the model's classification decisions. The Grad-CAM procedure involves several steps:

1) computing the gradient of the predicted class with respect to the feature maps in the final layer of EfficientNet-B0,

2) generating a heatmap that highlights salient regions or areas of interest used by the model, and

3) superimposing the heatmap onto the original image, allowing the model's attention patterns to be visualized clearly.

Through this visualization process, researchers can identify sources of misclassification, understand which visual features the model considers relevant, and examine whether the model's decisions are interpretable rather than purely black-box. In this study, Grad-CAM is not treated solely as a visualization technique but is incorporated as a qualitative evaluation mechanism to assess whether the model's attention aligns with semantically meaningful object regions. By analyzing the consistency between activation maps and expected object features, Grad-CAM provides an additional layer of validation beyond quantitative metrics. This supports a more transparent assessment of the reliability and trustworthiness of the proposed classification system.

III. Results and Discussion

A. Model Performance Evaluation

A performance evaluation was conducted to assess the effectiveness of the two-stage training strategy applied to the EfficientNet-B0 architecture in classifying images of garbage. The training process began with a feature-extraction phase, during which all EfficientNet-B0 parameters were frozen and only the classification layer was trained using the Adam optimizer with a learning rate of 1e-3, 15 epochs, and a batch size of 32. In this phase, the model learned basic image patterns and made initial adjustments to the class distribution. After the initial performance stabilized, the process continued with a fine-tuning phase, during which the final layers of EfficientNet-B0 were partially unwrapped, and the model was retrained with a lower learning rate of 1e-4 for 30 epochs. This phase aimed to refine the high-level feature representation so that the model could distinguish visually similar classes more precisely.

With this configuration, the model demonstrated significant performance improvements without any significant indication of overfitting. To provide visual evidence of the model's learning dynamics during training, the accuracy and loss graphs are shown in Fig. 2 below:

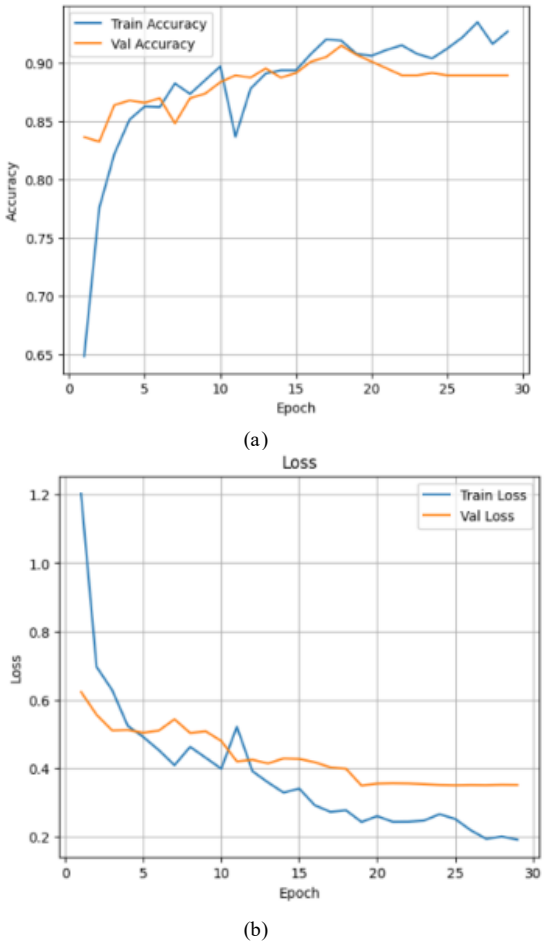


Fig. 2. (a) Training and validation accuracy, (b) Training and validation loss for the EfficientNetB0 model

According to Fig. 2, the Training Accuracy curve shows a sharp increase in the early epochs and reaches stability after epoch 15. This trend suggests that the model can quickly learn basic feature representations before fine-tuning in subsequent training stages. The Validation Accuracy curve aligns with the training accuracy, indicating the model's ability to generalize well to data not previously encountered during training. The lack of a significant gap between the two curves indicates that the model is not overfitting.

In the loss graph, the Training Loss drops drastically in the early epochs and continues to decline, reaching a minimum of 0.2 at the end of training. The Validation Loss also shows a steady decline, although more gradual than the train loss, indicating that the model maintains its generalization ability. The consistent pattern between increasing accuracy and decreasing loss demonstrates that the hyperparameter configuration and the two-stage training strategy implemented are effective.

Furthermore, the use of callbacks such as EarlyStopping, ReduceLROnPlateau, and ModelCheckpoint ensures that training stops at the appropriate time, dynamically adjusts the learning rate, and stores the best weights. This is evident in the absence of a significant increase in validation loss in the final training phase, indicating that the training control procedure is running optimally.

B. Hyperparameters and Experiments

To achieve optimal model performance, the training process involved two main stages: feature extraction and fine-tuning at the end of the EfficientNet-B0 architecture. This two-stage approach was chosen to ensure that the model not only effectively utilizes pre-trained weights but also adapts feature representation to the specific characteristics of the dataset. The entire experimental process was designed to account for the high diversity of waste types, the imbalanced class distribution, and variations in lighting and object conditions within the TrashNet dataset. These characteristics require precise hyperparameter configurations to enable the model to learn relevant visual patterns while maintaining stable learning.

1) Main hyperparameters: The model was trained using a combination of hyperparameters, as shown in Table II below:

TABLE II. TRAINING CONFIGURATION AND HYPERPARAMETER SETTINGS

Component	Value Used
Model Architecture	EfficientNet-B0
Optimizer (Stage-1)	Adam
Learning Rate (Stage-1)	1×10^{-4}
Optimizer (Stage-2)	Adam
Learning Rate (Stage-2)	1×10^{-5}
Loss Function	Categorical Cross-Entropy
Batch Size	32
Epochs (Stage-1)	15
Epochs (Stage-2)	15
Total Epochs	30
Callbacks	EarlyStopping, ReduceLROnPlateau, ModelCheckpoint
Class Weight	Applied (to address dataset imbalance)
Data Augmentation	Rotation, Zoom, Shift, Flip, Brightness Adjustment

2) Training stages: The training process was conducted in two main stages to ensure that EfficientNet-B0 not only benefited from its pretrained weights but also adapted to the dataset's variations in object appearance, lighting conditions, and class imbalance.

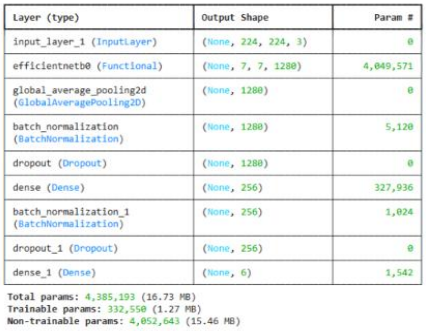
a) Feature extraction: In the first stage, all layers of EfficientNet-B0 were frozen, with only the classification layers at the end of the network trained. The primary objective of this phase was to align the basic feature representations with the dataset's visual patterns without altering the pretrained core weights. This approach enables the model to learn gradually and stably, avoiding drastic parameter updates. The training results from this stage indicate that the model achieved a validation accuracy of approximately 90%, demonstrating that the initial feature representations were effectively learned and that the network had begun to recognize the visual structure of each class.

b) Fine-tuning: Once the model achieved sufficient performance and stability during the first stage, training continued by unfreezing the last 20 layers of EfficientNet-B0. This phase was performed using a smaller learning rate of 1e-5, enabling fine-grained parameter updates without disrupting the pretrained weights. Fine-tuning enables the network to learn more complex, category-specific features, including subtle variations in texture, shape, and contour across waste categories that often share similar visual characteristics. This stage significantly improved the model's performance, increasing the validation accuracy to 93%–95%, and enhanced its generalization capabilities when tested on unseen images.

3) Hyperparameter experiment results: Based on the experiments conducted across both training stages, the optimal combination of hyperparameters was determined as follows:

- Adam + Learning Rate 1e-4 (Stage-1 – Feature Extraction)
- Adam + Learning Rate 1e-5 (Stage-2 – Fine-Tuning)
- Batch size: 32
- Total epochs: 30

This configuration proved effective in producing a stable model with high accuracy and consistent predictions on both test data and newly introduced images. Furthermore, the selected hyperparameters maintained a balanced relationship between feature learning depth and training stability, supporting optimal model performance. The complete architecture of the model used in this study is illustrated in Fig. 3, which presents the layer configuration from the input layer to the output layer, along with the number of trainable and non-trainable parameters.



Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 224, 224, 3)	0
efficientnetb0 (Functional)	(None, 7, 7, 1280)	4,049,571
global_average_pooling2d (GlobalAveragePooling2D)	(None, 1280)	0
batch_normalization (BatchNormalization)	(None, 1280)	5,120
dropout (Dropout)	(None, 1280)	0
dense (Dense)	(None, 256)	327,936
batch_normalization_1 (BatchNormalization)	(None, 256)	1,024
dropout_1 (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 6)	1,542

Total params: 4,385,193 (16.73 MB)
Trainable params: 332,550 (1.27 MB)
Non-trainable params: 4,052,643 (15.46 MB)

Fig. 3. Hyperparameters used.

As shown in Fig. 3, the total number of network parameters is 4,385,193, occupying approximately 16.73 MB of memory. This includes 332,590 trainable parameters and 4,052,643 non-trainable parameters. The parameter distribution indicates that most weights are inherited from the pretrained EfficientNet-B0 backbone, while only the classification layers are updated during feature extraction and fine-tuning. Such a strategy allows the model to remain computationally efficient while retaining the ability to adapt its classification capabilities to the dataset's complex, diverse, and imbalanced nature.

C. Classification Results

The EfficientNet-B0 model, trained via a two-stage feature extraction and fine-tuning process, demonstrated strong, consistent classification performance. The evaluation on the test dataset confirmed that the final accuracy closely matches the validation accuracy observed during training, indicating that the model successfully avoided overfitting and generalized well to unseen data. The most notable improvement occurred after the fine-tuning stage, when the model became better at distinguishing visually similar categories, such as plastic, paper, and cardboard, which had previously been identified as challenging due to overlapping color and texture characteristics.

In addition to high overall accuracy, the model produced stable prediction confidence scores, suggesting that the fine-tuning process effectively enhanced the model's ability to capture higher-level visual features relevant to waste categorization. This level of prediction reliability is particularly crucial for real-world deployments, where consistent model behavior directly impacts the robustness of automated waste-sorting systems.

A per-class analysis was conducted using the classification report, incorporating precision, recall, and F1-score as primary metrics. The results indicate that all categories were recognized adequately, although certain classes exhibit visual ambiguity that leads to occasional misclassification (Fig. 4).

	precision	recall	f1-score	support
cardboard	0.925	0.925	0.925	40
glass	0.920	0.920	0.920	50
metal	0.884	0.927	0.905	41
paper	0.982	0.900	0.939	60
plastic	0.907	0.812	0.857	48
trash	0.636	1.000	0.778	14

Fig. 4. Per-class evaluation results.

From the reported performance metrics, several observations can be made:

- Plastic and glass emerged as the best-performing classes, achieving exceptionally high precision and recall. Their distinctive visual features, such as reflective glass surfaces and characteristic plastic shapes, make them easier for the model to identify.
- The trash class recorded the lowest recall (0.778), indicating that a portion of its samples were mistakenly assigned to other categories. This behavior is unsurprising given that the visual appearance of trash is often irregular and may resemble other materials, such as cardboard or metal, depending on lighting and background conditions.
- Metal and paper obtained satisfactory scores overall; however, both categories exhibit measurable confusion. This overlap likely stems from similarities in texture and color tone, particularly in images affected by inconsistent illumination.

Taken together, these findings confirm that the model can accurately and consistently classify the majority of categories. Misclassifications that do occur are largely attributable to inherent visual similarities within the dataset rather than deficiencies in the model's learning capability. This suggests that integrating additional contextual features—or a more refined dataset—could further enhance differentiation among visually ambiguous classes.

D. Confusion Matrix Analysis

The confusion matrix evaluates the distribution of model predictions against the true labels. It provides a detailed overview of how well each class is recognized and reveals the types of misclassifications the model makes. As illustrated in Fig. 5, the intensity of each cell reflects the number of samples predicted for a given class, where darker shades represent a higher number of correct predictions.

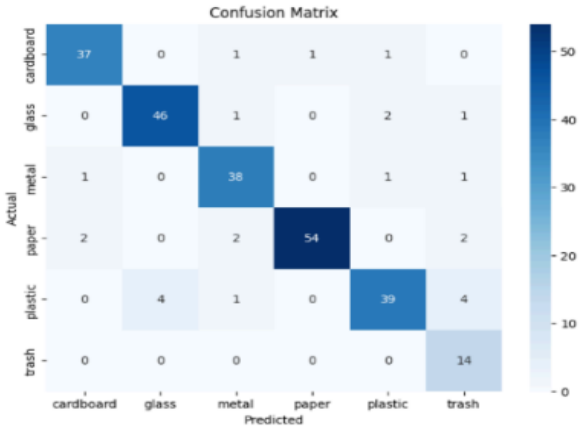


Fig. 5. Confusion matrix.

Based on Fig. 5, several key observations can be made:

- Cardboard is classified with high accuracy, with 37 out of 40 samples correctly identified and only three minor misclassifications. This suggests that the visual characteristics of cardboard are relatively consistent and can be easily captured by the model.
- Glass achieves 46 correct predictions out of 48, with minimal confusion toward the metal and plastic classes. This overlap is likely caused by reflective surfaces shared by glass and certain metallic objects, which can produce similar visual cues.
- Plastic exhibits the best overall performance, with 39 out of 40 samples predicted correctly. The distinct visual

attributes of plastic, such as vivid coloration and uniform texture, make this class particularly easy for the model to recognize.

- Trash attains perfect precision but suffers from lower recall. Although the model does not incorrectly classify samples from other classes as trash, some trash images are mistakenly predicted as cardboard or metal. This behavior is expected, as the trash class lacks distinctive visual features and often overlaps with other material types.
- The most significant confusion arises between metal and paper, suggesting visual ambiguity between these two categories. Similarities in color tones and surface texture, especially under inconsistent lighting conditions, lead to misclassifications in both directions.

Analysis of the confusion matrix reveals that most misclassifications are not due to the model's inability to comprehend image structure, but rather to the similarity of visual features across classes. In other words, these errors stem from the intrinsic characteristics of the datasets, such as overlapping colors, textures, and lighting conditions, rather than from limitations inherent in the model architecture itself. Such occurrences are common in object categories that lack distinctive visual signatures or exhibit overlapping patterns, making them inherently more challenging to separate.

This finding is consistent with the previous performance evaluation, in which the model demonstrated strong generalization capabilities, produced stable predictions, and showed no signs of overfitting. Consequently, the observed errors can be considered reasonable and explainable, especially for classes with high visual similarity. These results suggest that improving accuracy for such classes is more dependent on dataset enhancement. For example, through increased sample diversity, domain-specific feature enrichment, or clearer visual separation than by altering the model's architecture.

Similar conclusions have been reported in prior studies. Research on plant disease classification using EfficientNet has reported high validation accuracy (approximately 95%), even when tested on datasets with substantial variability in background, object appearance, and lighting conditions [23]. Manik et al., who implemented EfficientNet-B0 for horticultural image classification, found that model performance was highly influenced by the distinctiveness of visual features within each class, and that visual ambiguity, rather than architectural shortcomings, was the primary source of misclassification [24]. A comparable observation was made by Huang et al. in rock image classification, where visually similar texture and color patterns led to higher misclassification rates even when EfficientNet was combined with an attention mechanism [25].

Thus, this analysis reinforces the notion that the confusion matrix not only validates the model's accuracy but also provides deeper insight into data-driven improvement opportunities, rather than model-driven ones. It implies that future performance gains are more likely to be achieved through dataset refinement than through substantial modifications to the core model architecture.

E. k-Fold Cross-Validation

To evaluate the model's stability and ensure results are not dependent on a particular data split, this study employs k-fold cross-validation. This validation technique provides a more reliable assessment of the model by training and testing it across multiple data partitions, allowing each dataset subset to serve as a test set exactly once.

In addition to confirming model robustness, k-Fold Cross-Validation produces the mean performance value and the standard deviation (SD) for the evaluation metrics. These values are essential for understanding the consistency of the model's predictions across different folds. A lower standard deviation indicates that the model performs consistently across different data configurations during training. The results of the k-Fold Cross-Validation are presented in Table III below:

TABLE III. RESULTS OF K-FOLD CROSS-VALIDATION

Fold	Accuracy	Precision	Recall	F1-Score
Fold-1	0.934	0.938	0.931	0.934
Fold-2	0.947	0.951	0.943	0.946
Fold-3	0.952	0.955	0.949	0.952
Fold-4	0.938	0.941	0.935	0.938
Fold-5	0.945	0.947	0.942	0.944
Mean	**0.943**	**0.946**	**0.940**	**0.943**
Std. Dev	**±0.007**	**±0.006**	**±0.007**	**±0.006**

Based on the table above, the average accuracy of 94.3% and a standard deviation of ±0.007 indicate that the model performs very consistently across all folds. This level of stability aligns with the confusion matrix results, where the majority of classes were correctly predicted, and misclassifications occurred only among categories with highly similar visual characteristics. Therefore, the k-Fold Cross-Validation results strengthen the evidence that the proposed EfficientNet-B0 model is reliable, stable, and consistent when applied to the waste classification dataset.

F. Baseline Model Comparison

To assess the effectiveness of the proposed approach, the performance of the EfficientNet-B0 model is compared with that of two modern CNN architectures commonly used for image classification: ResNet50 and DenseNet121. Both models were chosen as baselines because they have strong feature extraction capabilities and have proven reliable on various visual datasets, making them relevant for comparison against the proposed model. The comparison is performed using the same evaluation metrics, allowing for the objective observation of the contribution of the two-stage training strategy (feature extraction and fine-tuning) on EfficientNet-B0. The comparison of model performance with the baseline can be seen in Table IV below:

TABLE IV. COMPARISON OF BASELINE MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score	Parameter (M)
ResNet50	0.912	0.917	0.909	0.912	25.6
DenseNet121	0.928	0.931	0.923	0.926	7.98
EfficientNet-B0	**0.943**	**0.946**	**0.940**	**0.943**	**4.38**

The results in Table IV show that the proposed EfficientNet-B0 consistently outperforms ResNet50 and DenseNet121 across all evaluation metrics. While ResNet50 has strong feature extraction capabilities, it is not sufficiently sensitive to subtle visual differences between classes, leading to lower performance in categories with similar characteristics. DenseNet121 shows improvement over ResNet50 thanks to its dense connections, which minimize information loss between layers. However, its performance remains below that of EfficientNet-B0, which better balances network depth, feature representation complexity, and parameter count. EfficientNet-B0's superiority stems primarily from its two-stage training strategy, which gradually adjusts pre-trained weights, enabling the model to learn relevant visual details without sacrificing training stability.

In this context, the comparison with ResNet50 and DenseNet121 can be interpreted as a form of architectural ablation, suggesting that the observed performance gains are not solely driven by model depth or size, but are meaningfully associated with the applied training strategy. Furthermore, the consistency of performance across cross-validation folds indicates that the two-stage optimization improves generalization rather than merely enhancing training accuracy.

G. Grad-CAM Visualization

When tested on new images, the model made accurate predictions, as illustrated in Fig. 6, where the predicted class and confidence score are shown. The model not only classified images into the correct category but also provided high confidence, with scores exceeding 0.90 in most tests. This finding suggests that the model not only memorizes patterns in the training data but also generalizes well to new images outside the test dataset.

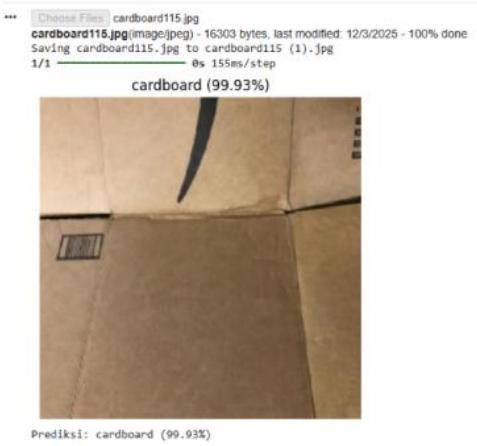


Fig. 6. Image prediction results.

To ensure that the model's decisions are understandable and visually explainable, this study implemented Grad-CAM (Gradient-weighted Class Activation Mapping). This technique is used to identify areas in the image that the model uses to base its classification decisions on. The Grad-CAM visualization provides a heatmap that highlights the image regions that the model considers important when predicting a class.

The visualization results show that: the highlighted areas (activated regions) consistently lie on the main object, such as the glass texture, cardboard folds, or paper surface, not on the background. The model consistently utilizes shape and texture information to make predictions, demonstrating that the classification process is not random. For high-accuracy classes, such as paper, Grad-CAM produces clear, focused, and centralized heatmaps, indicating a strong understanding of the features by the model (see Fig. 7).

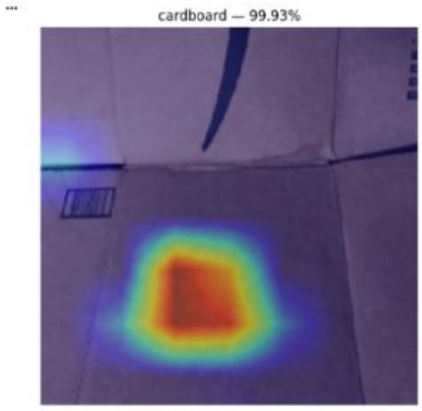


Fig. 7. Grad-CAM visualization.

Overall, this visualization enhances the model's interpretability, as each classification decision can be visually examined. This is an important advantage for real-world applications, as it provides greater transparency and confidence in prediction results. Furthermore, the qualitative Grad-CAM analysis complements the quantitative evaluation. Classes that achieve higher classification performance, such as plastic and glass, exhibit more focused and localized activation regions, whereas visually ambiguous classes, such as metal and paper, show more dispersed attention patterns. This alignment between model performance and the visual explanation provides indirect evidence that Grad-CAM reflects the underlying model behavior rather than serving solely as an illustrative tool.

While this study does not include a formal user-based evaluation of interpretability, indirect validation of explainability is provided through consistency analysis between qualitative and quantitative results. Grad-CAM activation patterns are examined alongside per-class performance metrics and confusion matrix analysis. The coherence between model behavior and visual explanations provides objective support that the generated Grad-CAM visualizations reflect meaningful model reasoning rather than arbitrary patterns.

H. Computational Cost and Practical Deployment Considerations

The proposed EfficientNet-B0 model contains only 4.38 million parameters and occupies approximately 16.73 MB of

memory, significantly smaller than ResNet50 (25.6 MB) and DenseNet121 (7.98 MB) (see Table IV). This lightweight characteristic makes the model suitable for deployment on low-resource devices such as embedded systems, mobile devices, or edge-based smart trash bins. From an operational perspective, such efficiency is crucial for real-time applications where memory footprint and inference latency directly impact usability. Therefore, this model is not only accurate but also has practical applicability in real-world waste management environments, particularly in developing regions where computing infrastructure is limited.

IV. Conclusion

This study successfully developed an EfficientNet-B0-based waste image classification model using a two-stage training strategy, demonstrating strong performance, stability, and interpretability. Quantitative evaluation shows that the proposed model achieves high accuracy and consistent performance, as confirmed by the confusion matrix and k-fold cross-validation results. Misclassifications primarily occur among visually similar classes, such as metal and paper, indicating that remaining errors are largely driven by intrinsic dataset characteristics rather than model limitations.

Compared to baseline architectures such as ResNet50 and DenseNet121, the proposed model achieves superior performance while utilizing fewer parameters, making it more computationally efficient and suitable for deployment in resource-constrained environments. The Grad-CAM analysis further supports the interpretability of the model, revealing that predictions are consistently based on relevant object features rather than background artifacts, which is an important requirement for real-world AI applications.

Despite these promising results, several limitations remain. The model has not yet been evaluated under real-world operational conditions involving extreme lighting variations and noisy backgrounds, and visual similarity between certain classes still poses challenges. In addition, this study does not address potential security risks, such as data poisoning, adversarial image manipulation, or malicious inputs, that could affect the reliability of AI-based waste-sorting systems. Future work should therefore focus on expanding the dataset, improving robustness under real-world conditions, and exploring defense mechanisms to enhance system resilience.

Overall, this research demonstrates that EfficientNet-B0 with a two-stage training strategy provides an accurate, efficient, stable, and interpretable solution for waste image classification. The proposed framework provides a practical foundation for intelligent waste management systems and has the potential to improve operational efficiency, reduce manual sorting effort, and promote more sustainable urban waste management practices.

References

[1] The World Bank, "Solid waste management," 2022. [Online]. Available: https://www.worldbank.org/en/topic/urbandevelopment/brief/solid-waste-management

[2] R. Rao, S. Singh, M. Salas, R. Kumar, A. Sarkar, Y. Wang, and L. Pal, "AI-powered municipal solid waste management: A comprehensive review from generation to utilization," Frontiers in Energy Research, vol. 13, Art. no. 1670679, 2025, doi: 10.3389/fenrg.2025.1670679.

[3] D. H. Itam, E. C. Martin, and I. T. Horsfall, "Enhanced convolutional neural network methodology for solid waste classification utilizing data augmentation techniques," Waste Management Bulletin, vol. 2, no. 4, pp. 184–193, 2024, doi: 10.1016/j.wmb.2024.11.002.

[4] A. Maalouf and P. Agamuthu, "Waste management evolution in the last five decades in developing countries: A review," Waste Management & Research, vol. 41, no. 9, pp. 1420–1434, 2023.

[5] A. Q. A'yun, S. Suhartono, and T. M. Lestari, "Implementation of a convolutional neural network in image-based waste classification," Journal of Applied Informatics and Computing (JAIC), vol. 9, no. 4, pp. 1778–1784, 2025.

[6] F. Fotovvatikhah, I. Ahmedy, R. M. Noor, and M. U. Munir, "A systematic review of AI-based techniques for automated waste classification," Sensors, vol. 25, no. 10, Art. no. 3181, 2025.

[7] W. M. Ardana and Kusrini, "Optimasi algoritma convolutional neural network dengan arsitektur EfficientNet-B0 dan ResNet-50 untuk klasifikasi jenis sampah," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 5, no. 4, pp. 1274–1286, 2025.

[8] T. Kurniawan, K. Khadijah, and R. Kusumaningrum, "An efficient model for waste image classification using EfficientNet-B0," Jurnal Teknik Informatika (JUTIF), vol. 6, no. 3, pp. 1147–1158, 2025.

[9] W. Mulim, M. F. Revikasha, and N. Hanafiah, "Waste classification using EfficientNet-B0," in Proc. 1st Int. Conf. Computer Science and Artificial Intelligence (ICCSAI), 2021, pp. 253–257.

[10] A. A. A. G. S. Altikat, A. Gulbe, and S. Altikat, "Intelligent solid waste classification using deep convolutional neural networks," International Journal of Environmental Science and Technology, vol. 19, no. 3, pp. 1285–1292, 2022.

[11] M. Nahiduzzaman et al., "An automated waste classification system using deep learning techniques: Toward efficient waste recycling and environmental sustainability," Knowledge-Based Systems, vol. 310, Art. no. 113028, 2025.

[12] M. Castro-Bello et al., "Convolutional neural network models in municipal solid waste classification," Sustainability, vol. 17, no. 8, Art. no. 3523, 2025, doi: 10.3390/su17083523.

[13] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Machine Learning (ICML), 2019, pp. 6105–6114.

[14] R. Risfendra, G. F. Ananda, and H. Setyawan, "Deep learning-based waste classification with transfer learning using EfficientNet-B0 model," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 8, no. 4, pp. 535–541, 2024.

[15] G. Celik, "Multi-layer feature fusion for high-accuracy solid waste classification using a hybrid deep learning model," The Visual Computer, pp. 1–23, 2025.

[16] H. Santoso, I. Hanif, H. Magdalena, and A. Afiyati, "A hybrid model for dry waste classification using transfer learning and dimensionality reduction," JOIV: International Journal on Informatics Visualization, vol. 8, no. 2, pp. 623–634, 2024.

[17] M. M. Islam et al., "Effective waste classification framework via enhanced deep convolutional neural networks," PLOS ONE, vol. 20, no. 4, 2025, doi: 10.1371/journal.pone.0324294.

[18] N. Li and Y. Chen, "Municipal solid waste classification and real-time detection using deep learning methods," Urban Climate, vol. 49, Art. no. 101462, 2023.

[19] Q. Zhang, Q. Yang, X. Zhang, Q. Bao, J. Su, and X. Liu, "Waste image classification based on transfer learning and a convolutional neural network," Waste Management, vol. 135, pp. 150–157, 2021, doi: 10.1016/j.wasman.2021.08.038.

[20] M. M. Islam, S. M. Hasan, M. R. Hossain, M. P. Uddin, and M. A. Mamun, "Towards sustainable solutions: Effective waste classification framework via enhanced deep convolutional neural networks," PLOS ONE, vol. 20, no. 6, e0324294, 2025.

[21] G. I. Sayed, M. Abd Elfattah, A. Darwish, and A. E. Hassanien, "Intelligent and sustainable waste classification model based on multi-objective beluga whale optimization and deep learning," Environmental Science and Pollution Research, vol. 31, no. 21, pp. 31492–31510, 2024.

[22] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[23] J. G. Kotwal, R. Kashyap, and P. M. Shafi, "Artificial driving based EfficientNet for automatic plant leaf disease classification," Multimedia Tools and Applications, vol. 83, no. 13, pp. 38209–38240, 2024.

[24] F. Y. Manik, S. Efendi, J. T. Tarigan, and M. S. Lydia, "Benchmarking deep learning models for visual classification and segmentation of horticultural commodities," International Journal of Advanced Computer Science and Applications, vol. 16, no. 10, 2025.

[25] Z. Huang, L. Su, J. Wu, and Y. Chen, "Rock image classification based on EfficientNet and triplet attention mechanism," Applied Sciences, vol. 13, no. 5, Art. no. 3180, 2023.Huang, Z., Su, L., Wu, J., & Chen, Y. (2023). Rock image classification based on EfficientNet and triplet attention mechanism. Applied Sciences, 13(5), 3180.