JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

# Explainable Deep Learning Framework for Waste Image Classification Using MobileNetv2 with Grad-CAM and SHAP

**Andri Armaginda Siregar[1], Muhammad Al Adib[2*], Bill Raj[3], Rahmat Humala Putra Hasibuan[4], Jalaluddin Nasution[5], Andreas Jorghy Parapat[6], Rika Rosnelly[7]**

[1,2,3,4,5,6,7] Fakultas Teknik dan Ilmu Komputer, Ilmu Komputer, Universitas Potensi Utama, Kota Medan, Indonesia
Email: [1]andriarmagindasiregar@gmail.com, [2*]dibsowsen@gmail.com, [3]denilsdu10@gmail.com ,
[4]rahmathumala06@gmail.com , [5]jalaluddinnasution04@gmail.com, [6]parapatandreas@gmail.com,
[7]rikarosnelly@gmail.com
(*Email Corresponding Author: dibsowsen@gmail.com.)
*Received: December 23, 2025 | Revision: December 25, 2025 | Accepted: December 26, 2025*

**Abstract**

The increasing volume of waste resulting from urbanization and population growth poses significant challenges to waste management systems, particularly in the sorting stage. Deep learning approaches, especially Convolutional Neural Networks (CNNs), have been widely employed for waste image classification due to their ability to automatically extract complex visual features. However, a major limitation of these approaches lies in their limited interpretability, which may hinder user trust and real-world adoption. This study proposes an Explainable Deep Learning Framework for organic and inorganic waste image classification by integrating the MobileNetV2 architecture with Explainable Artificial Intelligence (XAI) methods, namely Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP). MobileNetV2 is utilized as a feature extractor due to its computational efficiency and suitability for deployment on resource-constrained devices. The dataset used in this study consists of a combination of a public benchmark dataset and field-acquired waste images, processed using a transfer learning approach. Model performance is evaluated using accuracy, precision, recall, and f1-score metrics. Experimental results demonstrate that the proposed model achieves a validation accuracy of 90.25% with balanced performance across both classes. Furthermore, interpretability analysis using Grad-CAM and SHAP reveals that the model focuses on semantically relevant visual features and provides explainable feature contributions. These findings confirm that integrating lightweight CNN architectures with XAI techniques can produce waste classification systems that are accurate, transparent, and accountable.

***Keywords***: Explainable AI,Waste Classification,Deep Learning,MobileNetV2,Grad-CAM,SHAP,Interpretability.

## 1. INTRODUCTION

The increasing volume of global waste resulting from urbanization, population growth, and shifts in societal consumption patterns has emerged as a critical challenge for modern waste management systems. The continuous accumulation of solid waste exerts substantial pressure on waste management infrastructure and directly contributes to environmental degradation, public health risks, and economic inefficiencies [1]. Within the waste management lifecycle, the classification and sorting stage represents a crucial process, as accuracy at this stage significantly determines the effectiveness of recycling and downstream treatment operations. Nevertheless, manual sorting processes continue to face inherent limitations, including strong dependence on human labor, low consistency in sorting outcomes, and high operational costs [2], [3].

With ongoing technological advancements, Artificial Intelligence (AI) has been increasingly adopted to automate waste sorting processes. In particular, deep learning approaches based on Convolutional Neural Networks (CNNs) have demonstrated superior performance in waste image classification due to their capability to automatically and hierarchically extract discriminative visual features. Numerous studies report that CNN architectures such as VGG, ResNet, Inception, EfficientNet, and Xception achieve high classification accuracy across various waste categories, including organic, inorganic, plastic, paper, metal, and glass [4], [5], [6], [7]. These results have established CNN-based approaches as the dominant paradigm in the development of automated waste sorting systems at both laboratory and industrial scales.

Despite these advances, findings from the Systematic Literature Review (SLR) conducted in this study indicate that the majority of existing CNN-based waste classification research estimated at more than 70% of reviewed studies primarily focuses on predictive performance metrics such as accuracy, precision, recall, and F1-score, while providing little or no analysis of model interpretability. As a consequence, many high-performing models remain black-box systems, where internal decision-making processes are difficult for users to understand or validate [8]. This lack of interpretability poses a significant barrier to real-world deployment, particularly in the waste management domain involving field operators, municipal authorities, and industrial stakeholders. Unexplained misclassifications may reduce user trust, hinder technology adoption, and raise concerns regarding system reliability and accountability [9].

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

To address this limitation, Explainable Artificial Intelligence (XAI) has been introduced to enhance the transparency and interpretability of deep learning models. XAI techniques enable visualization and interpretation of feature contributions to model predictions, thereby allowing users to understand the rationale underlying automated decisions. Among image-based explainability approaches, Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) are widely adopted. Grad-CAM generates spatial activation maps that highlight image regions most influential in determining class predictions, while SHAP provides quantitative explanations grounded in Shapley value theory to estimate feature contributions to model outputs [8], [10], [11].

However, the SLR also reveals that in most existing studies, XAI techniques are applied in a limited post-hoc manner and primarily used as supplementary visualization tools rather than as integral components of the evaluation framework. Grad-CAM and SHAP are often presented descriptively without systematic analysis of decision reliability or structured integration into the overall research methodology. Furthermore, many studies remain model-centric and have yet to propose an end-to-end explainable framework that cohesively integrates data preprocessing, model training, performance evaluation, and decision interpretation.

In addition to interpretability, the practical deployment of waste classification systems requires efficient model architectures that can operate under resource-constrained environments, such as edge devices, smart cameras, and Internet of Things (IoT)-based systems. This requirement is particularly relevant in developing countries, including Indonesia, where waste management systems often operate under limited computational infrastructure, high environmental variability, and strong reliance on manual sorting practices. In such contexts, lightweight and transparent AI models are not merely an optimization choice but a practical necessity to ensure usability, trust, and sustainability in real-world applications.

MobileNetV2 is a lightweight CNN architecture that employs depthwise separable convolutions and inverted residual blocks, enabling competitive classification performance with reduced computational complexity. These characteristics render MobileNetV2 particularly suitable for edge-based waste classification, where local inference is required to minimize latency and reliance on centralized servers [10], [11]. Nevertheless, SLR findings indicate that prior studies on MobileNetV2-based waste classification primarily emphasize computational efficiency and accuracy, while the interpretability of decisions produced by this lightweight architecture remains underexplored.

Moreover, binary waste classification distinguishing between organic and inorganic waste is frequently treated as a baseline problem and receives limited attention from an interpretability perspective. In real-world waste sorting systems, however, errors at the binary classification stage may propagate to subsequent processing steps and significantly affect overall system performance. Consequently, interpretability analysis in binary classification tasks remains critically important, particularly for ensuring reliability and accountability in operational environments.

Based on these identified research gaps, this study aims to develop and validate an explainable deep learning framework for organic and inorganic waste image classification, specifically designed for resource-constrained environments. The proposed framework integrates MobileNetV2 as a lightweight classification backbone with Grad-CAM and SHAP as complementary interpretability mechanisms. Unlike previous studies that focus primarily on predictive performance or employ XAI solely as a post-hoc tool, this research positions XAI as a core system-level component of the methodological framework, ensuring that model performance and decision transparency are jointly evaluated.

The primary contributions of this study are as follows: (1) the development of a validated explainable framework for waste classification suitable for deployment on resource-constrained devices; (2) the systematic integration of multi-level XAI techniques (Grad-CAM and SHAP) to enhance transparency and decision reliability at the system level; and (3) empirical evaluation using a combination of public benchmark datasets and real-world field data, providing practical insights for the deployment of transparent and accountable AI-based waste sorting systems in developing-country contexts.

## 2. RESEARCH METHODOLOGY

### 2.1 Research Workflow Overview

This study proposes an explainable deep learning framework for organic and inorganic waste image classification based on the MobileNetV2 architecture, enhanced with two interpretability approaches, namely Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP). Figure 1 illustrates the systematically designed methodological workflow to ensure that data acquisition, image preprocessing, model training, performance evaluation, and prediction interpretation are conducted in a structured and measurable manner.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

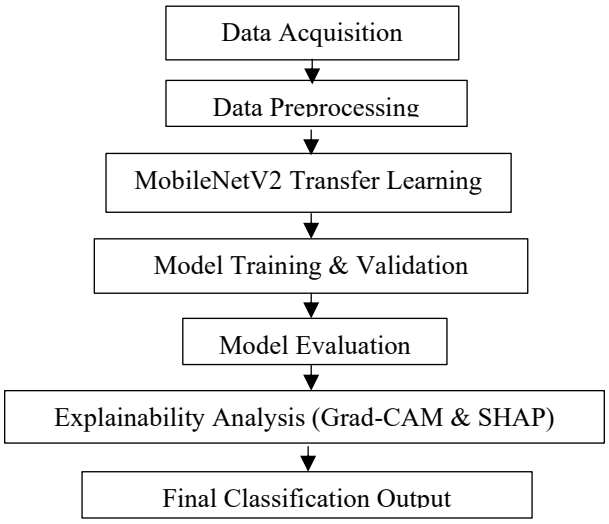


**Figure 1.** Overall methodological framework for CNN-based waste image classification

The research workflow begins with dataset compilation from two different sources: a public dataset obtained from the Kaggle platform and an internally collected dataset acquired through field image capture. This dual-source strategy aims to increase visual diversity and improve the model's generalization capability under real-world environmental conditions, as recommended in previous deep learning–based waste classification studies [4], [11], [12]. The compiled dataset subsequently undergoes data curation, normalization, and augmentation processes to enhance data quality and mitigate potential bias during model training.

Next, a transfer learning architecture is constructed using MobileNetV2 as the feature extractor. A pre-trained MobileNetV2 model on the ImageNet dataset is employed to accelerate convergence and reduce computational requirements, in accordance with best practices for lightweight CNN development in image classification tasks [10], [11]. Additional classification layers are trained using the waste dataset to adapt the learned feature representations to the target domain.

The final stage involves model performance evaluation using quantitative classification metrics, followed by interpretability analysis through Grad-CAM and SHAP. This multimodal interpretability approach is designed to ensure that model decisions are not only accurate but also explainable from both visual and quantitative perspectives, thereby enhancing transparency and user trust in the system [8], [9], [10].

### 2.2 Deep Learning Pipeline for Waste Classification

Figure 2 presents the end-to-end deep learning pipeline employed in this study for waste image classification. The pipeline encompasses the complete data processing chain, ranging from image acquisition to model interpretability visualization.

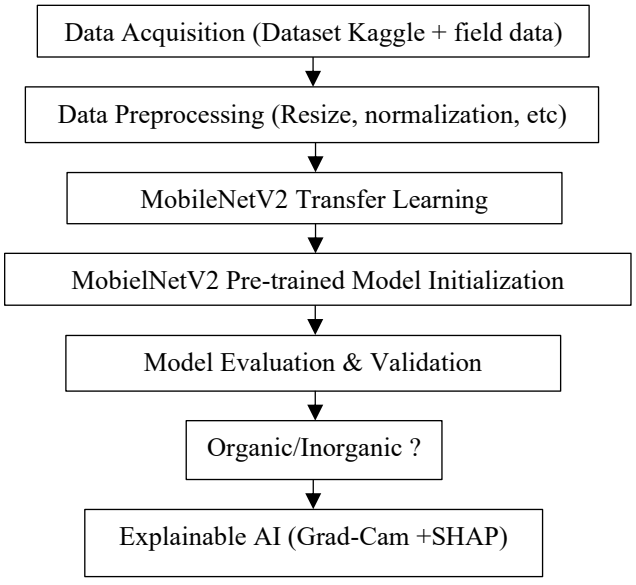


**Figure 2.** Deep learning pipeline

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

The initial stage is data acquisition, in which all image data are stored in Google Drive and accessed via the Google Colaboratory platform. The dataset is organized into a class-based directory structure compatible with TensorFlow's ImageDataGenerator, enabling batch-wise data loading and efficient on-the-fly preprocessing during training [5], [7].

The subsequent stage is the preprocessing pipeline, which includes image resizing, pixel intensity normalization, image format validation, and label encoding. These steps are performed to ensure uniform input dimensions and consistent pixel value ranges before images are processed by the MobileNetV2 network, thereby maintaining training stability and convergence.

To enhance data diversity and reduce the risk of overfitting, real-time data augmentation is applied during the training process. The augmentation techniques include rotation, translation, scaling, and horizontal flipping, which are designed to simulate real-world variations such as changes in viewing angle and illumination conditions [2], [4]

MobileNetV2 is then employed as the backbone feature extractor within the transfer learning scheme. The early convolutional layers are frozen to preserve pre-trained feature representations, while the newly added classification layers are trained using waste-domain-specific data. Upon completion of training, the model is evaluated using standard classification performance metrics, followed by an interpretability assessment using Grad-CAM and SHAP. This pipeline is designed to achieve a balance between predictive accuracy and model transparency.

### 2.3 Hardware and Software Environment

All experimental procedures in this study were conducted on the cloud-based Google Colaboratory platform, which provides access to NVIDIA Tesla T4 GPUs with 16 GB of memory. This environment was selected due to its capability to support hardware-accelerated computation, seamless integration with Google Drive, and high reproducibility of notebook-based experiments [5].

**Table 1.** Hardware and software specifications for experimental setup.

| Component | Specification |
|---|---|
| Execution Environment | Google Colab (Cloud-based Jupyter Environment) |
| Programming Language | Python 3.10 |
| Framework Deep Learning | TensorFlow 2.x (GPU enabled) |
| GPU Runtime | NVIDIA Tesla T4 / V100 |
| CUDA Toolkit | TensorFlow 2.x |
| cuDNN | Default Google Colab configuration |
| Model Evaluation Libraries | scikit-learn (classification report, confusion matrix) |
| Image Processing Library | OpenCV (cv2) |
| Interpretability Library | SHAP (SHAP Explainer, GradientExplainer) |
| Visualization Libraries | Matplotlib, Seaborn |
| Dataset Management | Google Drive API |

The hardware and software specifications used in this study are summarized in Table 1. The deep learning framework employed is TensorFlow version 2.x with GPU support, while Python 3.10 serves as the primary programming language. Model evaluation is performed using the scikit-learn library, whereas image preprocessing and manipulation are handled using OpenCV. For interpretability analysis, the SHAP library is utilized, including both SHAP Explainer and GradientExplainer modules, with Matplotlib and Seaborn employed for result visualization. The use of Google Colaboratory enables parallel execution, experimental scalability, and consistency of the computational environment without the need for manual CUDA and cuDNN configuration. These characteristics make the platform particularly suitable for deep learning–based research and facilitate reproducible experimentation.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

### 2.4 Dataset Description

The dataset used in this study is a combined dataset consisting of organic and inorganic waste images obtained from two primary sources: a publicly available dataset from the Kaggle platform and an independently collected dataset acquired through field image capture using a smartphone camera with a minimum resolution of 12 MP. This combination is intended to capture variations in visual appearance, texture, shape, and background that more accurately represent real-world conditions in domestic and public waste management scenarios [4], [11], [13]. Sample images from the dataset are illustrated in Figure 3.

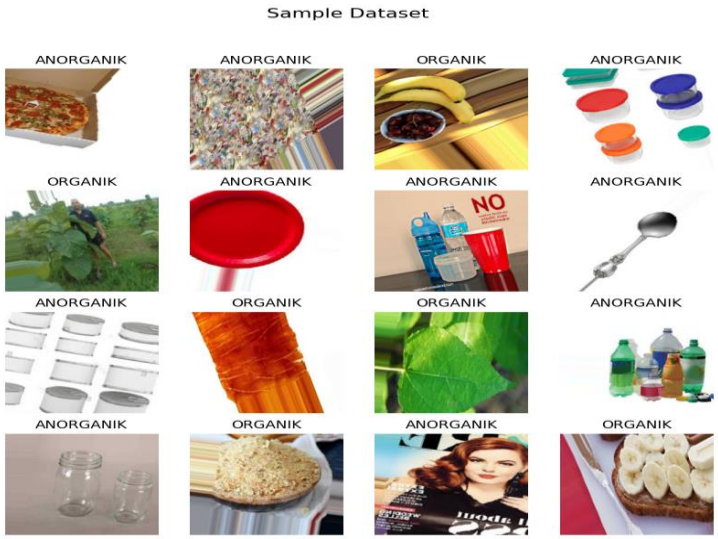


**Figure 3.** Sample Dataset

All images were collected across different time periods to minimize temporal bias. A manual inspection process was conducted to ensure image quality and labeling accuracy. Images exhibiting excessive noise, significant blur, or irrelevant objects were excluded to prevent degradation of model performance.

The dataset was subsequently organized into a hierarchical directory structure compatible with TensorFlow and strictly partitioned into two non-overlapping subsets: a training set and a testing set. The training set was used for model training and data augmentation, while the testing set was exclusively reserved for final model evaluation without any augmentation to ensure an unbiased assessment of model performance.

### 2.5 Dataset Distribution Analysis

At this stage, a dataset distribution analysis is conducted to examine the number of samples in each class. The distribution shown in Figure 4 indicates that the organic waste class contains a larger number of samples than the inorganic waste class. This class imbalance must be carefully considered, as it may affect model performance during the training process.

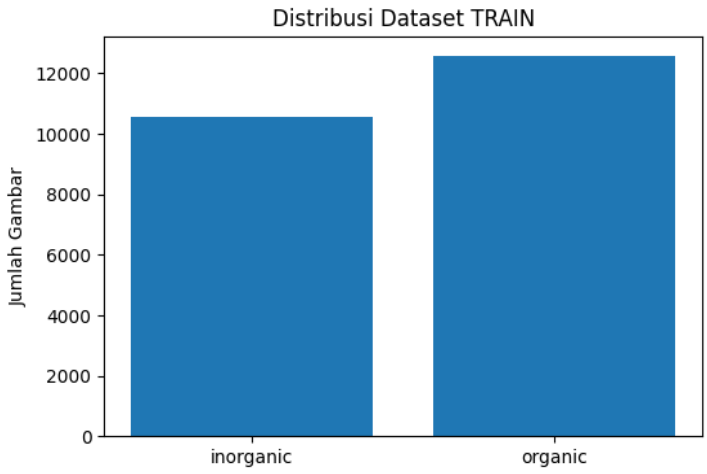


**Figure 4.** Training dataset distribution

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

The dataset distribution analysis is performed to identify the proportion of samples across classes. As illustrated in Figure 4, the organic class comprises a higher number of samples compared to the inorganic class, with approximately 12,600 images for organic waste and 10,600 images for inorganic waste. Such class imbalance has the potential to bias the learning process toward the majority class and degrade classification performance for the minority class. To mitigate this issue, a class weighting strategy is applied during model training, aiming to reduce class bias and promote balanced learning across both categories [7], [11].

## 3. RESULTS AND DISCUSSION

The results of this study are analyzed to evaluate the performance of the proposed MobileNetV2-based waste image classification model and to interpret the implications of the obtained findings. The analysis focuses on the impact of data augmentation on training stability, classification performance based on quantitative evaluation metrics, and the interpretability capability of the model through Explainable Artificial Intelligence (XAI) approaches. This evaluation strategy is consistent with current research practices in deep learning–based image classification systems, which emphasize the importance of balancing predictive accuracy and model transparency, particularly for real-world applications [9], [11], [12], [14]. Accordingly, the presented results not only demonstrate the classification effectiveness of the proposed model but also provide insights into the model's decision-making mechanisms from both visual and quantitative perspectives.

### 3.1 Data Augmentation

The distribution of samples in the training dataset was analyzed to identify the proportion of data in each class. As illustrated in Figure 5, the number of images in the organic waste class is higher than that in the inorganic waste class. Such distribution imbalance may lead to model bias toward the majority class, potentially degrading classification performance for the minority class, as commonly reported in deep learning–based image classification studies [7], [12].

```
Found 23148 images belonging to 2 classes.
Found 2513 images belonging to 2 classes.
```

**Figure 5.** Data augmentation

To mitigate the impact of dataset imbalance, a class weighting strategy was applied during the training process, such that misclassification errors in the minority class contributed more significantly to the loss function. This approach aims to encourage the model to learn feature representations more evenly across classes [11].

In addition, data augmentation was employed to enhance the visual diversity of the dataset and reduce the risk of overfitting, particularly given the high variability of environmental conditions and object orientations in real-world waste images [4], [8].The applied augmentation techniques include random rotations of up to 35°, horizontal and vertical translations of 0.25, zooming up to 0.3, shearing of 0.25, horizontal flipping, and the use of nearest fill mode to preserve pixel continuity. An example of the augmentation implementation is shown in Figure 6, while sample visualizations of the augmented dataset are presented in Figure 7.

```
train_gen = ImageDataGenerator(
    rescale=1./255,
    rotation_range=35,
    width_shift_range=0.25,
    height_shift_range=0.25,
    horizontal_flip=True,
    zoom_range=0.3,
    shear_range=0.25,
    fill_mode='nearest'
)
```

**Figure 6.** Data augmentation code

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***


**Figure 7.** Examples of augmented dataset

All augmentation processes were applied in real time to the training data using TensorFlow's ImageDataGenerator, while the testing data were not augmented to maintain the objectivity of model evaluation. This strategy is designed to simulate real-world variations such as changes in object orientation, background, and illumination commonly encountered in waste management environments, thereby improving the model's generalization capability to unseen data.

### 3.2 Data Preprocessing

The data preprocessing stage was conducted to ensure uniform input structure and maintain the stability of the MobileNetV2 training process. All images were first converted from the BGR color format to RGB, as the OpenCV library uses BGR by default, whereas the TensorFlow framework assumes RGB format to ensure consistent interpretation of color channels during feature extraction [10], [12].

Subsequently, all images were resized to 224 × 224 pixels, which corresponds to the standard input size required by the MobileNetV2 architecture, ensuring tensor dimensional compatibility within convolutional layers and alignment with pre-trained ImageNet weights [10], [11]. Pixel intensity normalization to the range [0, 1] was applied to stabilize gradient propagation, accelerate training convergence, and reduce the risk of exploding gradients, as recommended in best practices for CNN training [7], [12].

Class labels were then encoded into a binary format, with 0 representing the inorganic class and 1 representing the organic class. This encoding scheme was selected to align with the use of a sigmoid activation function and binary cross-entropy loss in the output layer, which are commonly adopted in CNN-based binary classification tasks [6]

```
# -------------------------
# Utility: preprocessing
# -------------------------
def preprocess_for_model(img_path_or_array, target_size=(224,224)):
    """Terima path string atau numpy array (H,W,3) -> kembalikan (1,H,W,3) float32 scaled."""
    if isinstance(img_path_or_array, str):
        img = cv2.imread(img_path_or_array)
        if img is None:
            raise FileNotFoundError(f"File tidak ditemukan: {img_path_or_array}")
        img = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)
    else:
        img = img_path_or_array.copy()
        # if BGR -> assume RGB from Gradio, but safe convert if suspicious
        if img.shape[2] == 3 and img.dtype == np.uint8:
            pass

    img_resized = cv2.resize(img, target_size)
    img_norm = img_resized.astype(np.float32) / 255.0
    return np.expand_dims(img_norm, axis=0), img_resized  # return batched array and rgb uint8 resized
```

**Figure 8.** Preprocessing code

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

Image loading and batch generation were performed using the flow_from_directory() function, which automatically converts the directory structure into training-ready image tensors, applies on-the-fly normalization, and generates standardized data batches without excessive memory consumption. The entire preprocessing pipeline was executed in real time during training, thereby reducing computational overhead, ensuring consistency across experiments, and enabling efficient integration with the TensorFlow data generator. The technical implementation of this stage is illustrated in Figure 8.

### 3.3 Model Architecture

The model architecture employed in this study is illustrated in Figure 9. The model is designed by utilizing MobileNetV2 as the feature extraction backbone, as this architecture is well known for its low computational complexity while maintaining competitive classification performance. These characteristics make MobileNetV2 particularly suitable for image classification systems deployed on resource-constrained devices, as reported in previous studies [11], [12].

To adapt the architecture to the binary waste classification task (organic and inorganic), a classification head is appended to the end of the network. MobileNetV2 is initialized using ImageNet pre-trained weights, and all convolutional layers are frozen during the initial training phase to preserve the general visual feature representations learned from large-scale data. This strategy aims to improve training efficiency and reduce the risk of overfitting, especially when working with relatively limited datasets [7], [10].

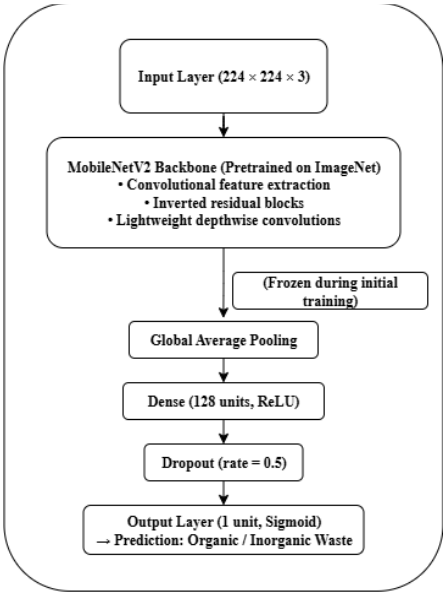


**Figure 9.** MobileNetV2 model architecture

A Global Average Pooling layer is applied to reduce the dimensionality of the extracted feature maps, thereby minimizing the number of parameters without discarding salient information. Subsequently, a fully connected (Dense) layer with 128 neurons and a ReLU activation function is added to learn nonlinear feature representations specific to the waste classification domain. To further enhance the model's generalization capability, a dropout layer with a rate of 0.5 is incorporated. Finally, an output layer consisting of a single neuron with a sigmoid activation function is used to generate binary predictions corresponding to organic or inorganic waste categories.

### 3.4 Training Parameters

The MobileNetV2 model is trained for 10 epochs using the Adam optimizer with an initial learning rate of 0.0001. The selection of the Adam optimizer is motivated by its adaptive learning rate mechanism, which has been shown to be effective for training CNN-based image classification models [2], [3], [15], [16]. During training, both accuracy and loss values for the training and validation datasets are recorded at each epoch to monitor model performance. The visualization of these training metrics is presented in Figure 10.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
*DOI: https://doi.org/10.62712/juktisi.v4i3.739*

```
Epoch 1/10
724/724 ━━━━━━━━━━━━━━━━━━━━ 7449s 10s/step - accuracy: 0.7052 - loss: 0.6150 - val_accuracy: 0.8830 - val_loss: 0.3002 - learning_rate: 1.0000e-04
Epoch 2/10
724/724 ━━━━━━━━━━━━━━━━━━━━ 1450s 2s/step - accuracy: 0.8526 - loss: 0.3523 - val_accuracy: 0.8961 - val_loss: 0.2747 - learning_rate: 1.0000e-04
Epoch 3/10
724/724 ━━━━━━━━━━━━━━━━━━━━ 1467s 2s/step - accuracy: 0.8746 - loss: 0.3134 - val_accuracy: 0.8973 - val_loss: 0.2660 - learning_rate: 1.0000e-04
Epoch 4/10
724/724 ━━━━━━━━━━━━━━━━━━━━ 1499s 2s/step - accuracy: 0.8845 - loss: 0.2899 - val_accuracy: 0.8993 - val_loss: 0.2584 - learning_rate: 1.0000e-04
Epoch 5/10
724/724 ━━━━━━━━━━━━━━━━━━━━ 1507s 2s/step - accuracy: 0.8870 - loss: 0.2875 - val_accuracy: 0.9025 - val_loss: 0.2518 - learning_rate: 1.0000e-04
Epoch 6/10
724/724 ━━━━━━━━━━━━━━━━━━━━ 1465s 2s/step - accuracy: 0.8884 - loss: 0.2767 - val_accuracy: 0.8965 - val_loss: 0.2549 - learning_rate: 1.0000e-04
Epoch 7/10
724/724 ━━━━━━━━━━━━━━━━━━━━ 1460s 2s/step - accuracy: 0.8984 - loss: 0.2608 - val_accuracy: 0.8993 - val_loss: 0.2496 - learning_rate: 1.0000e-04
Epoch 8/10
724/724 ━━━━━━━━━━━━━━━━━━━━ 1468s 2s/step - accuracy: 0.8983 - loss: 0.2557 - val_accuracy: 0.8973 - val_loss: 0.2551 - learning_rate: 1.0000e-04
Epoch 9/10
724/724 ━━━━━━━━━━━━━━━━━━━━ 1465s 2s/step - accuracy: 0.9010 - loss: 0.2507 - val_accuracy: 0.8973 - val_loss: 0.2548 - learning_rate: 1.0000e-04
Epoch 10/10
724/724 ━━━━━━━━━━━━━━━━━━━━ 0s 2s/step - accuracy: 0.9026 - loss: 0.2489
Epoch 10: ReduceLROnPlateau reducing learning rate to 2.9999999242136255e-05.
724/724 ━━━━━━━━━━━━━━━━━━━━ 1405s 2s/step - accuracy: 0.9026 - loss: 0.2489 - val_accuracy: 0.8977 - val_loss: 0.2546 - learning_rate: 1.0000e-04
Restoring model weights from the end of the best epoch: 7.
```

**Figure 10.** Training results across epochs

Based on the visualization, the model demonstrates a stable improvement in performance, as evidenced by increasing accuracy values and decreasing loss values across epochs for both training and validation data. In addition, the ReduceLROnPlateau callback automatically reduces the learning rate when performance improvement begins to plateau, while the ModelCheckpoint mechanism restores the best-performing model weights obtained at epoch 7.

**Table 2.** Summary of best model performance

| Metric | Value |
|---|---|
| Best epoch | 7 |
| Training accuracy | 0.898 |
| Validation accuracy | 0.9025 |
| Training loss | 0.2489 |
| Validation loss | 0.2518 |

As shown in Table 2, the model achieves its optimal performance at epoch 7, with a validation accuracy of 0.9025 and a validation loss of 0.2518. These results indicate that the model is able to learn effectively without exhibiting overfitting. The use of the ReduceLROnPlateau callback in conjunction with the ModelCheckpoint strategy contributes to training stability and ensures that the best-performing model weights are utilized during the evaluation phase.

### 3.5 Model Evaluation

Model evaluation is conducted to assess the capability of the MobileNetV2-based classifier in accurately predicting organic and inorganic waste categories on the test dataset. The evaluation is not limited to overall accuracy but is performed more comprehensively using learning curves, a confusion matrix, a classification report, and prediction visualization. This multi-metric evaluation approach is widely adopted in deep learning–based image classification research to provide a more representative assessment of model performance [7], [11], [12].

#### 3.5.1 Learning Curve Analysis

In this study, learning curve analysis is performed to examine the progression of model performance throughout the training process. The learning curves consist of accuracy and loss plots, which illustrate changes in model accuracy and error rates across epochs. These curves are useful for diagnosing underfitting, overfitting, or stable convergence behavior.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
*DOI: https://doi.org/10.62712/juktisi.v4i3.739*

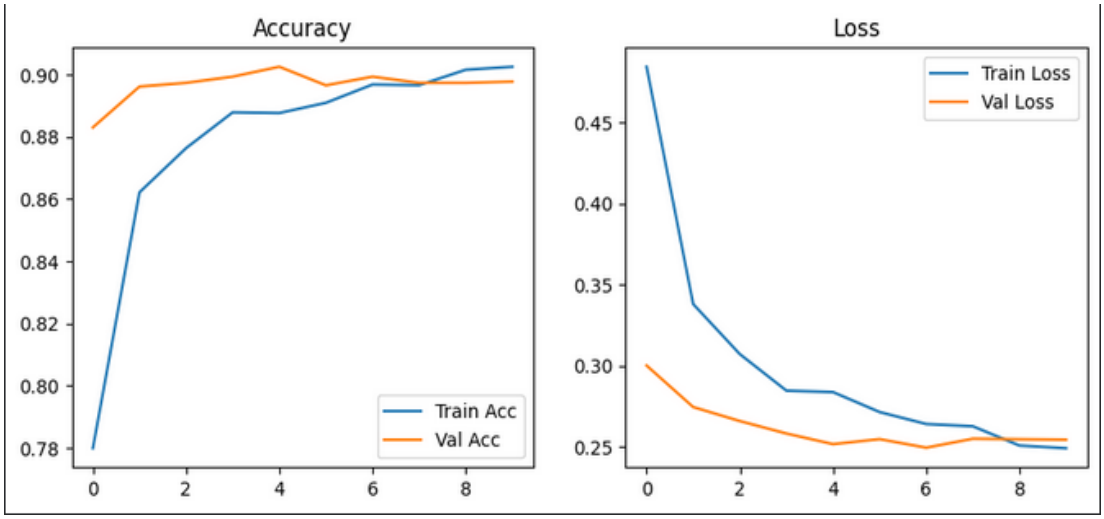


**Figure 11.** Learning Curve

As shown in Figure 11, the accuracy curves exhibit consistent improvement for both training and validation data until reaching a stable condition after several epochs. The slightly higher validation accuracy compared to training accuracy indicates the absence of overfitting and reflects good generalization capability. Meanwhile, the loss curves display a steady decline for both datasets, with validation loss values remaining comparable to training loss. This pattern confirms that the model effectively learns meaningful data patterns without memorizing the training samples. Overall, the learning curves demonstrate that the MobileNetV2 model operates under stable and optimal training conditions [11], [12].

### 3.5.2 Confusion Matrix Analysis of the MobileNetV2 Model

During the model evaluation stage, a confusion matrix is employed to assess the capability of the proposed MobileNetV2 model in distinguishing between inorganic and organic waste classes based on the generated predictions. The confusion matrix provides detailed information on the number of correct and incorrect predictions for each class, thereby offering a more comprehensive assessment of classification performance than overall accuracy alone. This analysis is particularly important for identifying class-specific recognition difficulties and dominant error patterns.

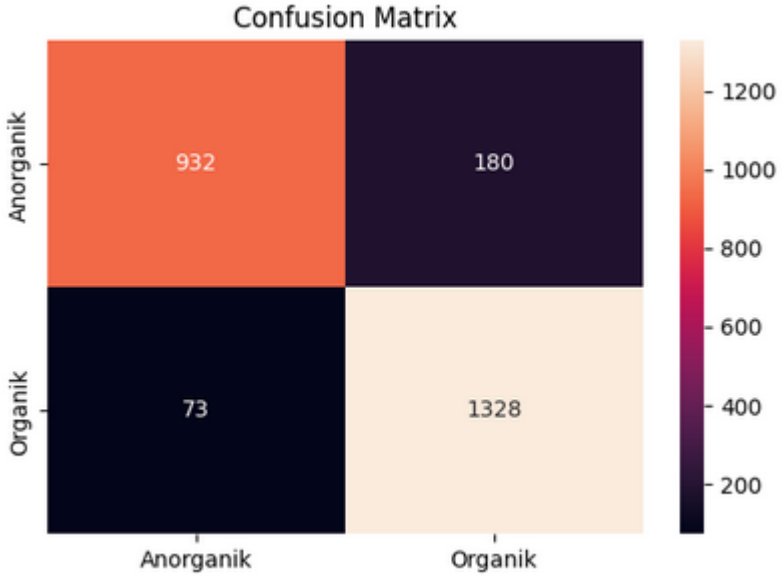


**Figure 12.** Confusion Matrix

As shown in Figure 12, the model correctly classifies 932 samples from the inorganic waste class, while 180 samples are misclassified as organic. For the organic waste class, the model correctly predicts 1,328 samples, with only 73 samples misclassified as inorganic. The dominance of values along the main diagonal of the matrix indicates strong classification performance for both classes. Overall, the number of misclassifications is relatively small compared to the total number of test samples, suggesting that MobileNetV2 exhibits effective class discrimination capability in the context of binary waste classification [7], [8].

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

### 3.5.3 Confusion Matrix Analysis of the MobileNetV2 Model

The evaluation results at this stage demonstrate stable performance across both classes. The precision, recall, and F1-score values for the inorganic and organic classes range from 0.88 to 0.95, indicating that the model is able to consistently capture relevant visual patterns. Detailed metric values are summarized in Table 3.

**Table 3.** Classification Report

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Inorganic | 0.93 | 0.84 | 0.88 | 1112 |
| Organic | 0.88 | 0.95 | 0.91 | 1401 |
| **Accuracy** | | | **0.90** | 2513 |
| **Macro Avg** | 0.90 | 0.89 | 0.90 | 2513 |
| **Weighted Avg** | 0.90 | 0.90 | 0.90 | 2513 |

For the inorganic class, the model achieves a precision of 0.93, a recall of 0.84, and an F1-score of 0.88. Meanwhile, the organic class attains a precision of 0.88, a recall of 0.95, and an F1-score of 0.91. The overall classification accuracy reaches 0.90, with both macro-average and weighted-average values remaining consistent across the primary metrics. These results indicate that the model not only achieves high overall accuracy but also maintains relatively balanced performance across both classes, despite differences in data distribution [7], [12].

### 3.5.4 Prediction Visualization

At this stage, prediction visualizations are conducted on selected test samples to examine the consistency of the model in classifying waste categories. This visualization presents a direct comparison between the ground truth labels and the model's predicted labels, allowing qualitative assessment of prediction quality through representative image examples.


**Figure 13.** Prediction visualization

Based on the visualization results, the model correctly classifies 7 out of 8 test images in accordance with their true labels. This observation indicates that the model demonstrates good generalization capability when applied to previously unseen data. The single misclassification case suggests the presence of visual feature similarities between certain waste categories, which remains a common challenge in waste image classification. Such limitations may be addressed in future work through increased data diversity or further refinement of the model architecture [4], [11].

### 3.5.4 PCA/t-SNE Visualization

Subsequently, feature space visualization is performed using Principal Component Analysis (PCA) to validate the feature extraction capability of the MobileNetV2 model. High-dimensional feature representations obtained from the final layer (global average pooling) are projected into a two-dimensional space to examine how effectively the model separates characteristics between waste classes. The resulting projection is illustrated in Figure 14.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

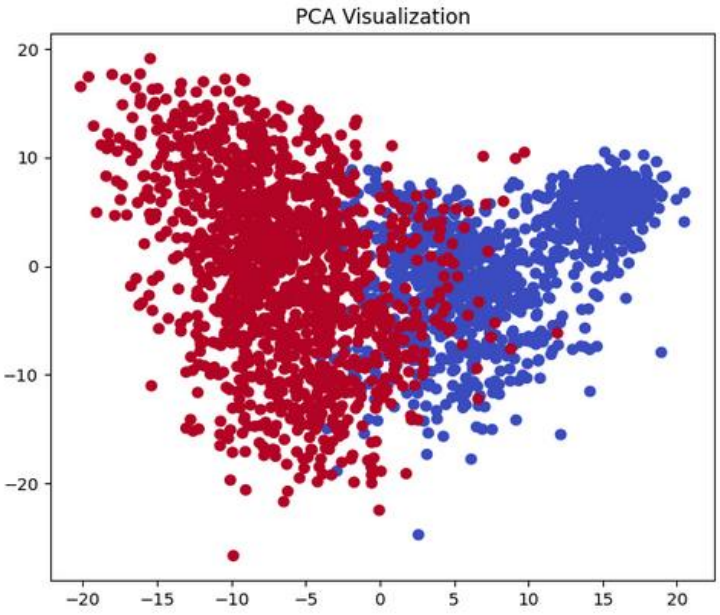


**Figure 14.** PCA Visualization

As shown in Figure 14, two primary clusters corresponding to the organic and inorganic waste classes are formed. The relatively clear separation of clusters on the left and right sides of the plot indicates that the model successfully learns discriminative visual features between the two classes. However, an overlapping region appears in the central area of the projection, suggesting the presence of samples with highly similar feature representations. This overlap explains some of the misclassification cases observed in the confusion matrix and highlights the model's limitations in distinguishing samples with very similar visual characteristics [7], [11], [17].

### 3.6 Grad-CAM & SHAP Visualization Overlay

In this stage, model interpretability is examined using Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) to observe the image regions and features contributing to the predictions generated by the MobileNetV2 model. These visualizations are presented as supporting results to complement the quantitative evaluation of the classification model.

Grad-CAM is applied to the final convolutional layer of MobileNetV2 to produce spatial activation maps indicating the regions with the highest contribution to the predicted class. The Grad-CAM overlays on the input images, together with the corresponding SHAP visualizations, are shown in Figure 15.

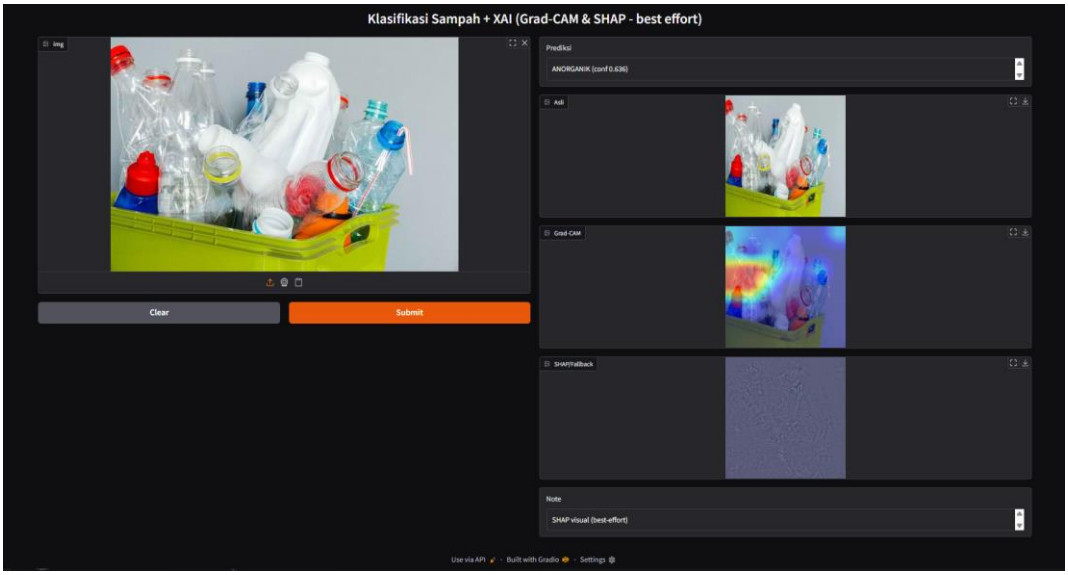


**Figure 15.** Grad-Cam & SHAP Visualization

As shown in Figure 15, the input image is classified as Inorganic with a prediction probability of 0.636. The Grad-CAM heatmap highlights regions associated with plastic bottle objects, particularly around the bottle caps and curved

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

surface areas. Regions with higher activation are represented by red and yellow intensities, while areas with lower contribution appear in blue.

The SHAP visualization illustrates pixel-level contribution patterns to the model's prediction. The resulting contribution map shows both positive and negative influences distributed around the primary object region. Although the SHAP representation appears more diffuse than the Grad-CAM heatmap, the areas with notable contribution remain concentrated around the same object regions identified by Grad-CAM.

Overall, the Grad-CAM and SHAP results indicate that the model's attention and feature contributions are primarily localized around the waste object present in the image. These visualizations are reported as observational results and are used to support the analysis of the model's prediction behavior.

## 4. CONCLUSION

This study successfully developed an Explainable Deep Learning Framework for organic and inorganic waste image classification by integrating the MobileNetV2 architecture with Grad-CAM and SHAP interpretability techniques, and the evaluation results show that the proposed model demonstrates strong performance with a validation accuracy of 90.25% as well as balanced precision, recall, and F1-score values across both classes, confirming that MobileNetV2 is capable of effectively extracting discriminative visual features even when applied to datasets with high variability. Furthermore, the integration of Grad-CAM and SHAP provides significant added value by enabling both visual and quantitative explanations of model decisions, where Grad-CAM consistently highlights semantically relevant image regions contributing to classification outcomes, while SHAP offers deeper insights into feature contributions. The combined use of these techniques demonstrates that the proposed model is not only accurate but also transparent and accountable, thereby enhancing user trust in the context of automated waste sorting system deployment. Overall, this research indicates that lightweight architectures such as MobileNetV2, when combined with Explainable Artificial Intelligence (XAI) approaches, offer an effective solution for automated waste sorting, particularly on resource-constrained devices, and future work may extend this framework by incorporating additional waste categories, increasing the diversity of field-collected datasets, integrating IoT-based sensing systems, or deploying the model in real-world environments to evaluate performance under operational conditions.

## REFERENCES

[1] Md. G. R. Alam, A. Al Nakib, Md. N. Talukder, C. Majumder, S. Biswas, and J. Hassan, "Deep learning-based waste classification system for efficient waste management," Thesis, BRAC UNIVERSITY, 2021.

[2] G. Thung and M. Yang, "Classification of Trash for Recyclability Status," *Environmental Science, Computer Science*, 2016.

[3] O. Adedeji and Z. Wang, "Intelligent Waste Classification System Using Deep Learning Convolutional Neural Network," *Procedia Manuf*, vol. 35, pp. 607–612, 2019, doi: 10.1016/j.promfg.2019.05.086.

[4] Doly Ilham Saputra Huta Julu and Dewi Nurdiyah, "KLASIFIKASI SAMPAH ORGANIK DAN NON ORGANIK MENGGUNAKAN TRANSFER LEARNING," *Jurnal Transformatika*, vol. 23, no. 1, pp. 12–29, Jul. 2025, doi: 10.26623/transformatika.v23i1.12201.

[5] Md. M. Hossen *et al.*, "A Reliable and Robust Deep Learning Model for Effective Recyclable Waste Classification," *IEEE Access*, vol. 12, pp. 13809–13821, 2024, doi: 10.1109/ACCESS.2024.3354774.

[6] R. Kurniawan, P. B. Wintoro, Y. Mulyani, and M. Komarudin, "IMPLEMENTASI ARSITEKTUR XCEPTION PADA MODEL MACHINE LEARNING KLASIFIKASI SAMPAH ANORGANIK," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 11, no. 2, Apr. 2023, doi: 10.23960/jitet.v11i2.3034.

[7] M. E. Purba, A. Z. Situmorang, G. L. Br Ginting, M. W. P. Lubis, and F. M. Sinaga, "Klasifikasi Sampah Organik dan Anorganik Menggunakan Algoritma CNN," *Jurnal Sifo Mikroskil*, vol. 26, no. 1, pp. 37–54, Apr. 2025, doi: 10.55601/jsm.v26i1.1510.

[8] G. Ahmad *et al.*, "Intelligent waste sorting for urban sustainability using deep learning," *Sci Rep*, vol. 15, no. 1, p. 27078, Jul. 2025, doi: 10.1038/s41598-025-08461-w.

[9] Z. Zhao, L. Alzubaidi, J. Zhang, Y. Duan, U. Naseem, and Y. Gu, "Robust and explainable framework to address data scarcity in diagnostic imaging," *Comput Biol Med*, vol. 197, p. 111052, Oct. 2025, doi: 10.1016/j.compbiomed.2025.111052.

[10] M. A. I. Aminudin, M. N. Abdullah, F. Mustapha, K. K. Eng, M. Mustapha, and A. Mustapha, "Explainable Deep Learning Framework for Binary Corrosion Image Classification Using Grad-CAM," *Sensors*, vol. 25, no. 22, p. 7070, Nov. 2025, doi: 10.3390/s25227070.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 1697-1710
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
***DOI: https://doi.org/10.62712/juktisi.v4i3.739***

[11] M. Naznine *et al.*, "PLDs-CNN-ridge-ELM: Interpretable lightweight waste classification framework," *Eng Appl Artif Intell*, vol. 162, p. 112522, Dec. 2025, doi: 10.1016/j.engappai.2025.112522.

[12] Md. M. Hossen *et al.*, "A Reliable and Robust Deep Learning Model for Effective Recyclable Waste Classification," *IEEE Access*, vol. 12, pp. 13809–13821, 2024, doi: 10.1109/ACCESS.2024.3354774.

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

[14] T. Hulsen, "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare," *AI*, vol. 4, no. 3, pp. 652–666, Aug. 2023, doi: 10.3390/ai4030034.

[15] G. Ahmad *et al.*, "Intelligent waste sorting for urban sustainability using deep learning," *Sci Rep*, vol. 15, no. 1, p. 27078, Jul. 2025, doi: 10.1038/s41598-025-08461-w.

[16] T. Hulsen, "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare," *AI*, vol. 4, no. 3, pp. 652–666, Aug. 2023, doi: 10.3390/ai4030034.

[17] K. Muchtar, N. T. Anshari, C. Chairuman, K. Alhabibie, and K. Munadi, "Rancang Bangun Purwarupa Pemilah Sampah Pintar Berbasis Deep Learning," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 3, p. 655, Jun. 2022, doi: 10.25126/jtiik.2022934976.