JSCE (Journal of System and Computer Engineering)
Vol.7, No.3, July 2026, pp. 241-254
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v7i3.2694

Explainable Non-Organic Waste Classification: A Comparative Study of CNN with SHAP Interpretability Against Vision Transformer Approaches

Setio Basuki [1,a,*]; Lika Anjelina [2,b]; Alfian Wahyu Juhar Putra [3,c]; Yusuf Nur Muhammad [4,d]
[1,2,3,4] Department of Informatics, Universitas Muhammadiyah Malang, Malang, Indonesia
[a] setio_basuki@umm.ac.id; [b] likaanjelina@webmail.umm.ac.id; [c] alfianwahyujuharputra@webmail.umm.ac.id; [d] yusufnurm@webmail.umm.ac.id
** Corresponding author*

Abstract

This study develops a classification model based on Machine Learning (ML) and Computer Vision (CV) to automatically distinguish recyclable and non-recyclable waste. Waste management still faces the challenge of low manual sorting efficiency, thus reducing recycling potential. The dataset consists of 399 waste images divided with a ratio of 80:10:10 for training, validation, and testing. Three combinations of visual features were tested, namely Mean Color (RGB) + LBP + HOG, Histogram + HOG + Edge, and Mean Color (RGB) + Histogram + GLCM, each of which was evaluated using four conventional ML algorithms, namely SVM-RBF, SVM-Linear, Random Forest, and Gradient Boosting. Meanwhile, deep learning models namely CNN, ViT, and LoRA ViT were trained directly on raw images without manual feature extraction. Experimental results show that CNN achieved the highest testing accuracy of 82.50%, outperforming all conventional ML models that achieved a maximum accuracy of 75.00%, as well as ViT (72.50%) and LoRA ViT (70.00%). The application of SHAP-based Explainable AI (XAI) provides transparency to the model's decision-making process. These findings demonstrate that CNN with certain regularization settings are effective for distinguishing recyclable and non-recyclable waste, in supporting sustainable smart waste management systems.

Keywords—*Computer Vision, Explainable AI, Machine Learning, Waste Image Classification, Waste Management.*

1. Introduction

The explosive growth becomes a serious environmental problem in the countries that mostly rely on manual waste sorting mechanisms (Asmawati et al., 2025; Ibnul Rasidi et al., 2022; Kurniawan et al., 2023). This type of sorting leads to increase environmental pollution and reduced economical value from recyclable materials (Sapanli et al., 2023). Ineffective sorting prevents the implementation of smart waste management systems in realizing the sustainable circular economy (Andi Muhammad Akbar et al., 2025). Moreover, the manual sorting makes it susceptible to human errors (workload, subjectivity, lighting conditions, and viewing angle) (Fotovvatikhah et al., 2025; Kiyokawa et al., 2024; Kiyokawa et al., 2021). Following this, the heterogeneity of waste in terms of shape, color, and texture brings difficulties to maintaining consistent classification (S.Intam et al., 2024). Looking at this situation, it is required automatic waste sorting system operate quickly, accurately, and reliably while reducing dependence on human labor (Doly Ilham Saputra Huta Julu & Dewi Nurdiyah, 2025; Kiyokawa et al., 2021).

The integration of Artificial Intelligence (AI) with Computer Vision (CV) is now paving the way for the creation of more reliable and efficient automated waste sorting systems

(Fotovvatikhah et al., 2025). As part of AI, Machine Learning (ML) enables computers to recognize visual attributes of waste from color and texture to shape as a basis for distinguishing between recyclable and non-recyclable waste (Sunardi et al., 2023). Conventional ML methods, which work by manually extracting features from images and then passing them to classifiers such as SVM and Random Forest, are known to be computationally efficient while still being able to compete in terms of accuracy on small datasets (Sami et al., 2020). On the other hand, Deep Learning (DL) methods such as Convolutional Neural Network (CNN) and Vision Transformer (ViT) are capable of automatically learning feature representations directly from raw images (Huang et al., 2021; Zhao et al., 2024; Muslihati et al., 2024). Despite showing superior performance on large-scale data, the application of DL to limited datasets often leads to overfitting, and it requires high computation resources (Fotovvatikhah et al., 2025; Gurcan & Soylu, 2026). This situation makes comparative evaluation between approaches crucial to determine which method is most effective in scenarios with limited training data.

Recently, there are several studies conducted feature engineering combined with traditional ML algorithms that can still produce competitive performance, and at several cases outperforming DL models (Nisa et al., 2022; Boumaraf et al., 2021). The visual features such as Mean Color, Histogram, Local Binary Pattern (LBP), Histogram of Oriented Gradients (HOG), and Gray Level Co-occurrence Matrix (GLCM) have shown its competitiveness in performing waste object detection (Hallur & Gavade, 2025; Alhindi et al., 2018). Here, the concept of Explainable Artificial Intelligence (XAI) allows users to understand the reason behind models' decisions is required (Saeed & Omlin, 2023). The model interpretability can be obtained using SHapley Additive exPlanations (SHAP), which calculates the contribution of each feature to the model's predictions (Givisis et al., 2025).

In reviewing the current literature, there are several gaps. Nisa et al. (2022) and other researchers first use SVM to analyze the features based on GLCM for multi-class trash categorization. We found that no previous work has attempted to integrate Mean Color, Histogram and GLCM versus CNN and ViT architectures for trash classification. Secondly, despite the success of ViT in multi-class trash classification (Huang et al., 2021), the efficiency of parameter-efficient fine-tuning with Low-Rank Adaptation (LoRA) for waste classification with limited data ($n < 500$) remains unknown. Most of the existing ViT-based research fine-tune on datasets with more than 2,000 images (Huang et al., 2021; Okubo & Kimura, 2024). So, this work proposes a garbage classification system that identifies recyclable vs non-recyclable using visual cues based on conventional ML algorithms (SVM, Random Forest, Gradient Boosting) and DL-based models (CNN, Vision Transformer (ViT), LoRA-enhanced ViT). Moreover, accommodating the importance of XAI (Saeed & Omlin, 2023), SHAP-based analysis will be applied to interpret the decision made by DL-based model.

To be more specific, this study delivers several contributions as follows:

- We provide a benchmark comparing four conventional ML algorithms with three feature set combinations (12 combinations in total), CNN, ViT, and LoRA ViT for the classification of recyclable and non-recyclable waste on a very limited dataset (399 images).
- We investigated the effectiveness of LoRA-based parameter-efficient fine-tuning techniques on ViT architecture under data-limited conditions, where LoRA ViT achieved 70.00% accuracy compared to standard ViT of 72.50%, indicating that this adaptation technique requires more diverse data to perform optimally.
- We integrate SHAP-based Explainable AI (XAI) analysis on the best CNN model to identify the visual regions that most influence classification decisions, thereby improving the transparency and interpretability of the system.
- We present empirical evidence that a CNN built from scratch with the right regularization configuration is able to achieve the highest accuracy of 82.50%,

outperforming the best conventional ML model (75.00%) and the Transformer-based architecture under limited training data conditions.

2. Method

The overall flow of the research process is shown in Figure 1. The flow consists of (1) dataset preparation; (2) image preprocessing (resizing and normalization); (3) building classification model classical ML with feature extraction and deep learning; and (4) model evaluation accompanied by interpretability analysis using SHAP (the best model).

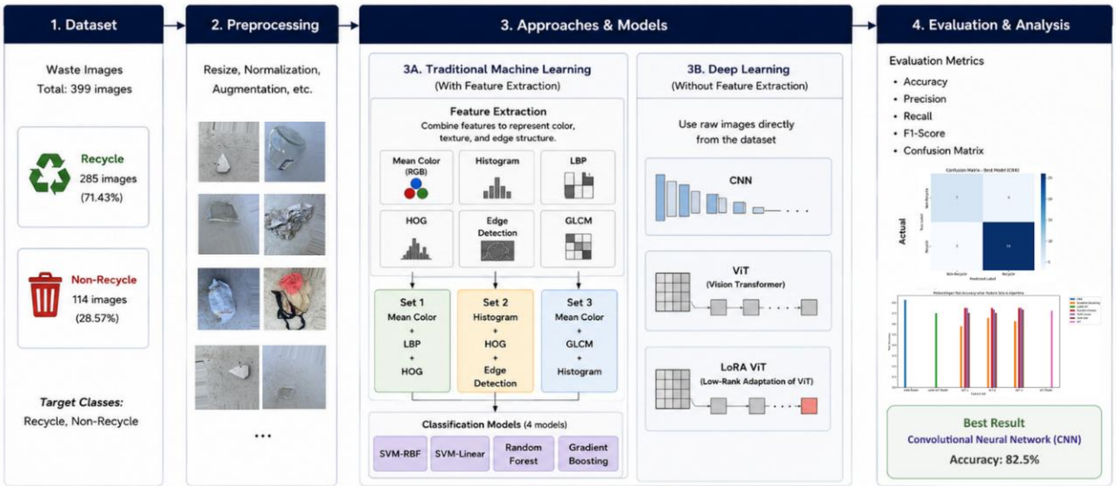

Figure 1. Flowchart of research on classification of recycled and non-recycled waste images

2.1 Data Acquisition

The dataset used in this study consists of 399 images of non-organic waste divided into two classes: recyclable and non-recyclable. Recycled waste refers to waste that can be reprocessed into raw materials, such as plastic, paper, metal, glass, etc and non-recycled waste refers to waste that cannot be reprocessed and generally ends up in landfills, such as rubber, medical devices, and styrofoam. Examples of images from both categories are shown in Figure 2.

Figure 2. Example of recyclable and non-recyclable images in our dataset.

The class imbalance exists in this dataset, as the number of images in the recycled class is 285 and the number of images in the non-recycled class is 114. These images were collected using two methods: scrapping from Google Images and collecting images in the field directly. This imbalance reflects real-world conditions as recycled waste such as plastic, paper, and metal is generally easier to find than non-recycled waste which has a more limited variety of forms and is more difficult to collect systematically.

This situation is similar to the usual issues in waste classification research, where lack of access to representative non-recyclable samples is a common hurdle (Fotovvatikhah et al., 2025). In this paper, we purposely avoided the use of resampling-based methods such as SMOTE or class weighting, as the introduction of synthetic samples into a relatively small image dataset can give rise to false visual patterns and introduce bias into the training process (Azhar et al., 2022). This was done to preserve the original data distribution and ensure that model evaluation would take place under realistic classification settings. However, the small size of the non-recylable class is still a shortcoming of this research that should be improved by collecting more representative data in the future.

These images subsequently underwent a quality screening process to ensure sufficient resolution and visual clarity for the subsequent feature extraction stage. For traditional ML models, each image was scaled to 224 × 224 pixels to standardize input dimensions before manual feature extraction. The size was selected since it is compatible with several feature extraction methods such as HOG, LBP, and GLCM that require a similar input size to provide a consistent feature representation across samples. Meanwhile, the deep learning models, such as CNN and ViT, were developed and trained from scratch without the pre-learned weights. This process can directly match the size of the input to the proposed architecture in this work without relying on the resolution criteria of specific pre-trained models.

2.2 Preprocessing and Normalization

Each image is further preprocessed after being resized. The image pixel values are in the range [0, 255] before normalization, which can cause instability in the training process due to the substantial variances in the intensity scales between the pixels. Then, the pixel values are normalized in the range of [0, 1] by dividing each pixel value by 255, which makes the value distribution more uniform and does not dominate the model learning process. Figure 3 shows the comparison of the distribution of pixel values before and after normalization on representative images from the dataset.

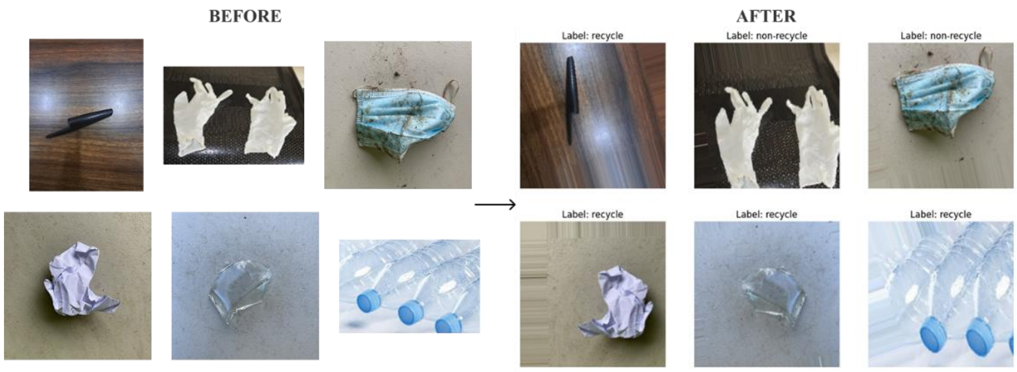

Figure 3. Images Visualization Before and After Preprocessing and Normalization

The dataset is split based on the model type. For deep learning models, the dataset is split at an 80:10:10 ratio for training, validation, and testing, respectively. For conventional ML models, an 80:20 ratio is applied for training and testing. This difference reflects their training requirements. Deep learning models need a validation set to monitor training, while conventional ML models do not.

2.3 Feature Extraction

In the feature extraction stage, each image is represented as a numeric vector that reflects its visual characteristics. Several types of features are used, including color, texture, and shape features. Color features include Mean Color (RGB), which captures mean color intensity, and a color Histogram, which describes the overall color distribution. Texture features are extracted

DOI : https://doi.org/10.61628/jsce.v7i3.2694

using LBP and GLCM, which capture surface texture patterns of waste objects. Shape and edge features are extracted using HOG and Edge Detection, which capture gradient and contour information. Three feature combinations were evaluated: (1) Set 1: Mean Color (RGB) + LBP + HOG, (2) Set 2: Histogram + HOG + Edge, (3) Set 3: Mean Color (RGB) + Histogram + GLCM. All extracted features are then scaled to a comparable range before classification.

2.4 Building Classification Model

The classification phase was done using different classical ML and deep learning methods. For traditional ML, we employed four algorithms: Support Vector Machine with RBF kernel (SVM-RBF), SVM with Linear kernel (SVM-Linear), Random Forest, and Gradient Boost. Each method was evaluated against the three feature sets (Set 1, Set 2 and Set 3) resulting in a total of 12 test scenarios. As a comparison, three deep learning models were trained directly on raw image data without any manual feature extraction, such as CNN, ViT and LoRA-ViT. The complete hyperparameter setup for each model is listed in Table 1. These values were picked according to the common approaches for image classification with limited datasets (Sami et al., 2020; Gurcan & Soylu, 2026).

Table 1. Model Hyperparameter Configuration

Model	Hyperparameter	Value
SVM-RBF	Kernel, C	RBF,1.0
SVM-Linear	Kernel, C	Linear, 1.0
Random Forest	n_estimators, max_depth	100. 10
Gradient Boosting	n_estimators, max_depth	100, 10
CNN	Epochs, Optimizer, Dropout, Learning Rate	10, Adam, 0.5, 0.001
ViT	Epochs, Patch Size, Num Heads, Transformer Layers	10, 16, 4, 8
LoRA-ViT	Epochs, Rank (r), LoRA Alpha, Target Modules	10, 16, 32, Query & Value

2.4.1 Support Vector Machine (SVM)

A Support Vector Machine (SVM) works by finding the optimal hyperplane that separates two data classes with the greatest margin. The SVM training process aims to minimize, as shown in equation 1 (Nisa et al., 2022):

$$min\ ½||w||^2 + C\Sigma\xi_i\ s.t.\ y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad (1)$$

where w is the weight vector, b is the bias, C is the regularization parameter, and ξ_i is the slack variable that accommodates non-perfectly separable data. For SVM-RBF, the Radial Basis Function kernel function is used to handle non-linear data, s shown in equation 2 (Razaque et al., 2021):

$$K(x_i, x^g) = exp(-\gamma||x_i - x^g||^2) \quad (2)$$

where $\gamma > 0$ is a parameter that controls the influence of each data point. Meanwhile, SVM-Linear uses a $K(x_i, x^g) = x_i^T x^g$ kernel for data that is linearly separable in the original feature space.

2.4.2 Random Forest

Random Forest constructs a random number of decision trees T and combines their predictions through majority voting. The final prediction is expressed in equation 3 (Sami et al., 2020):

$$\hat{y} = argmax\_c\ \Sigma^{t}{}_{=1}{}^{T\ T}(h^{t}(x) = c) \quad (3)$$

where $h^t(x)$ is the t-th tree prediction for input x, c is the class label, and $^T(\cdot)$ is the indicator function. The best split at each node is selected using Gini Impurity using equation 4 (Sami et al., 2020):

$$Gini(t) = 1 - \Sigma\_c\, p(c|t)^2 \quad (4)$$

where $p(c|t)$ is the proportion of samples of class c at node t. This ensemble approach reduces the risk of overfitting and improves model stability, especially on small datasets.

2.4.3 Gradient Boosting

Gradient Boosting builds a model incrementally, with each new model trained to correct the residual error of the previous model. The model at the m-th iteration is expressed in equation 5 (Friedman, 2001):

$$F_{\mathrm{m}}(x) = F_{\mathrm{m-1}}(x) + \eta \cdot h_{\mathrm{m}}(x) \qquad (5)$$

where η is the learning rate and $h_m(x)$ is a weak learner trained on pseudo-residuals, which are the negative gradients of the loss function relative to the previous model's predictions. For classification, the log-loss function (binary cross-entropy) is used as the optimized objective function.

2.4.4 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a deep learning architecture specifically designed to process grid-shaped data, such as digital images. CNN works by extracting hierarchical feature representations through convolutional layers, dimensionality reduction layers (pooling layers), and fully connected layers. CNN extracts feature representations automatically through the convolution operation stated in equation 6 (Zhao et al., 2024):

$$a^{\mathrm{g}}[l] = ReLU(\Sigma\_k\, w^{\mathrm{gk}}[l] * a^{\mathrm{k}}[l-1] + b^{\mathrm{g}}[l]) \qquad (6)$$

where $w^{gk}[l]$ is the convolution filter, $a^k[l-1]$ is the activation of the previous layer, $b^g[l]$ is the bias, and ReLU is the activation function. In the output layer, the softmax function is used to generate a probability distribution over the classes as shown in equation 7 (Gurcan & Soylu, 2026):

$$P(y = c \mid x) = exp(z\_c) \,/\, \Sigma\_k\, exp(z\_k) \qquad (7)$$

where z_c is the logit of class c of the last fully connected layer. In this study, the CNN was trained directly on 224 × 224 pixel waste images without using manual feature extraction.

2.4.5 Vision Transformer (ViT)

Vision Transformer (ViT) is a self-attention-based deep learning architecture adapted from the Transformer model in Natural Language Processing (NLP). ViT divides the image into N patches, which are then processed as a sequence of tokens by the Transformer encoder. The core of ViT is the Multi-Head Self-Attention (MHSA) mechanism using the formula in equation 8 (Huang et al., 2021):

$$Attention(Q, K, V) = softmax(QK^{\mathrm{T}} \,/\, \sqrt{d\_k})\, V \qquad (8)$$

where Q, K, and V are the query, key, and value matrices obtained from the linear projection of the input, and d_k is the key dimension as a normalization factor. This mechanism allows the model to capture global relationships between image parts, becoming a major advantage of ViT over local CNNs.

2.4.6 Vision Transformer dengan Low-Rank Adaptation (LoRA-ViT)

Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning technique using google/vit-base-patch16-224 model, where the pretrained model weights are retained and only low-rank decomposition matrices are added and trained. This approach is particularly relevant for small datasets because it reduces the risk of overfitting due to the much smaller number of optimized parameters compared to full fine-tuning. LoRA adds the weight change ΔW decomposed into a product of two low-rank matrices to the frozen pretrained weights W_0 as shown in equation 9 (Hu et al., 2021):

$$W = W_0 + \Delta W = W_0 + (\alpha/r)\, BA \qquad (9)$$

where $B \in \mathbb{R}$^{$d\times r$} and $A \in \mathbb{R}$^{$r\times k$} with rank $r \ll min(d,k)$, and α is the scaling factor. The number of trainable parameters is reduced from $d\times k$ to $r(d+k)$, thus minimizing the risk of overfitting on small datasets. In this study, LoRA is applied to the query and value projection matrices of each ViT self-attention block.

2.5 Evaluation and Expainable Artificial Intelligence

Model performance evaluation was performed using several standard classification metrics, namely accuracy, precision, recall, and F1-score. These metrics were used to assess the model's ability to classify recyclable and non-recyclable waste images in a balanced and comprehensive manner. The equations for each metric are written in equations 10-13:

$$Accuracy = (TP + TN)\,/\,(TP + TN + FP + FN) \qquad (10)$$
$$Precision = TP\,/\,(TP + FP) \qquad (11)$$
$$Recall = TP\,/\,(TP + FN) \qquad (12)$$
$$F1 - score = 2 \times (Precision \times Recall)\,/\,(Precision + Recall) \qquad (13)$$

In addition to these four metrics, this study also measures the level of model overfitting using the equation 14:

$$Overfitting = (Train\ Accuracy - Test\ Accuracy) \times 100\% \qquad (14)$$

Positive values indicate that the model overfits to the training data so that its generalization ability decreases, while negative values indicate that the Test Accuracy is higher than the Train Accuracy which is generally caused by the regularization effect during training. To improve model transparency and interpretability, an Explainable Artificial Intelligence (XAI) approach using SHapley Additive explanations (SHAP) was applied to the best-performing models. The SHAP value for each feature i is expressed in equation 15 (Givisis et al., 2025):

$$\varphi_i = \Sigma\_\{S \subseteq F\backslash\{i\}\}\ [|S|!\,(|F| - |S| - 1)!\,/\,|F|!]\ [f(S \smallint \{i\}) - f(S)] \qquad (15)$$

where F is the set of all features, S is the subset of features that does not include feature i, and $f(S)$ is the model prediction using subset S. *A* positive SHAP value indicates that the feature drives the prediction towards the positive class, while a negative value indicates the opposite. SHAP analysis is used to identify the visual features that contribute most to the prediction results, while providing insight into the main causes of misclassification to support further system development.

3. Results And Discussion

This section presents the experimental results and discusses feature extraction outcomes, classification model performance, and interpretability analysis.

3.1 Feature Extraction Result

This section presents the visualization results of feature extraction from the three feature set combinations used in this paper. Each image is represented as a numeric feature vector that reflects the visual characteristics of the waste object based on color, texture, and edge structure information.

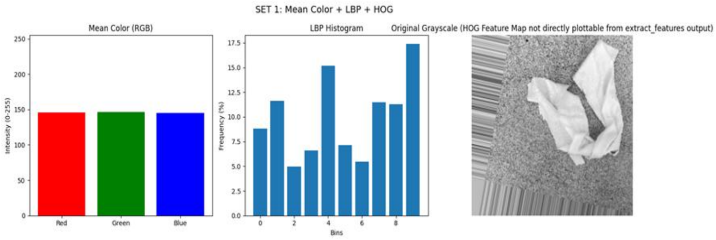

Figure 4. Preprocessing results using Mean color + LBP + HOG

Figure 4 shows the Mean Color (RGB) + LBP + HOG feature extraction results. This combination captures the color, surface texture, and edge structure of waste objects simultaneously.

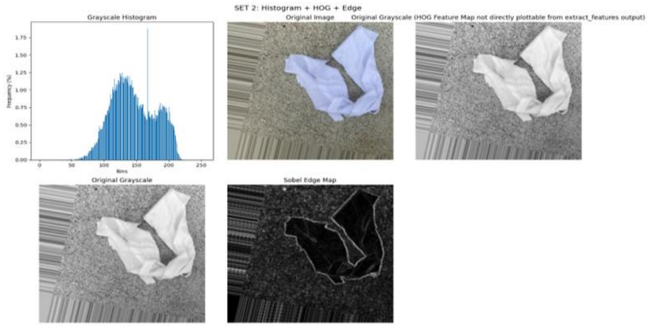

Figure 5. Preprocessing results using Histogram + HOG + Edge

Figure 5 shows the Histogram + HOG + Edge feature extraction results. This combination highlights the shape, contour, and visual intensity variations of waste objects.

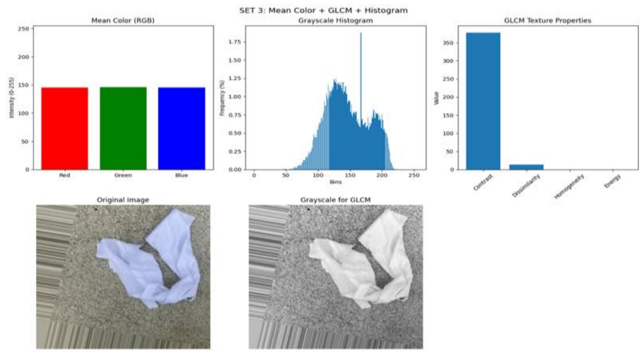

Figure 6. Preprocessing results using Mean Color + GLCM + Histogram

Figure 6 shows the Mean Color (RGB) + Histogram + GLCM feature extraction results. This feature set combines color average, color distribution, and texture information from the image surface. In general, each of the three feature sets captures different visual aspects of the waste images, so they can complement each other well.

3.2 Evaluation Results and Accuracy Comparison

Table 2 and Figure 7, CNN achieved the highest test accuracy of 82.50%, which is the best result among all models. CNN can learn visual features directly from raw images better than models that use feature engineering or feature extraction. Among classical ML settings, SVM-Linear Set 1 provided the best result with 75.00% accuracy, lower than CNN which shows that deep learning can still perform better even when the dataset is small. ViT achieved 72.50% and LoRA-ViT got 70.00%. Interestingly, LoRA produced a lower result than standard ViT. This indicates that LoRA requires more training data as well as its variety. Batch normalization and regularization techniques further contributed by preventing overfitting (Gurcan & Soylu, 2026; Zhao et al., 2024). Data augmentation and class weighting also played an important role in addressing data scarcity and class imbalance (Fotovvatikhah et al., 2025; Sunardi et al., 2023).

Table 2. Evaluation results and comparison of the performance of all classification models based on the combination of feature sets.

Feature Set	Algorithm	Train Acc (%)	Val Acc (%)	Test Acc (%)	Overfitting (%)	Time (s)
Set 1	SVM-RBF	93.75	-	70.31	23.44	0.25
Set 1	SVM-Linear	100.00	-	75.00	25.00	0.22
Set 1	Random Forest	100.00	-	75.00	25.00	1.31
Set 1	Gradient Boosting	100.00	-	57.81	42.19	179.52
Set 2	SVM-RBF	93.75	-	70.31	23.44	0.46
Set 2	SVM-Linear	100.00	-	73.44	26.56	0.22
Set 2	Random Forest	100.00	-	75.00	25.00	1.29
Set 2	Gradient Boosting	100.00	-	65.62	34.38	204.72
Set 3	SVM-RBF	82.29	-	73.44	8.85	0.00
Set 3	SVM-Linear	78.65	-	75.00	3.65	0.00
Set 3	Random Forest	100.00	-	75.00	25.00	0.21
Set 3	Gradient Boosting	100.00	-	62.50	37.50	1.39
CNN	CNN	73.04	82.50	82.50	-9.46	1.41
ViT	ViT	64.58	72.50	72.50	-7.92	1.83
LoRA	LoRA ViT	73.04	70.00	70.00	3.04	0.51

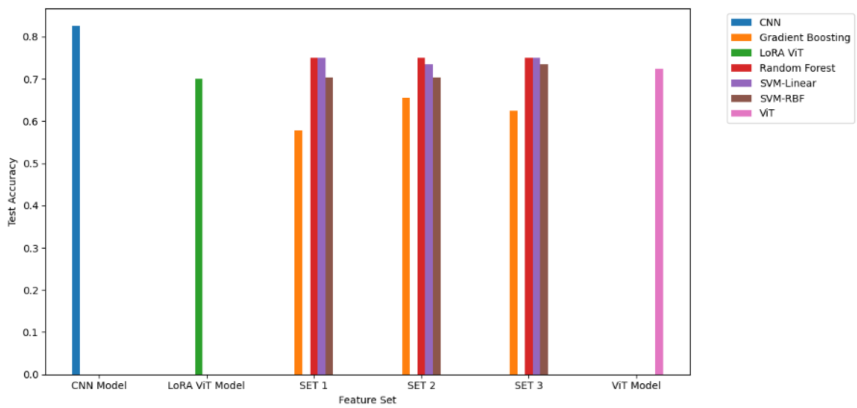

Figure 7. Testing Accuracy Comparison of Feature Sets Combined with Machine Learing Model.

Table 3 shows the classification report of the model produced by CNN. For the non-recyclable class, the model obtained a precision of 0.70, recall of 0.63, and F1-score of 0.66. These results reveal that the CNN struggled to recognize the visual pattern of this category. For the recyclable class, the report provided better results that are: precision 0.87, recall 0.90, and F1-score 0.88. The high recall indicated thata CNN is competitive for this category. However, this also indicated that the model would miss more non-recyclable samples, since the training data is dominated by the recyclable class. Overall, the model achieved 82% accuracy from a total of 64 test samples.

Table 3. Classification report (CNN)

	Precision	Recall	F1-Score	Support
Non-Recycle	0.70	0.63	0.66	11
Recycle	0.86	0.89	0.88	29
Accuracy	-	-	0.82	40
Macro Avg	0.78	0.76	0.77	40
Weighted Avg	0.82	0.82	0.82	40

A confusion matrix analysis was conducted to further evaluate the classification performance of the best models. Figure 8 shows the confusion matrix of the CNN model.

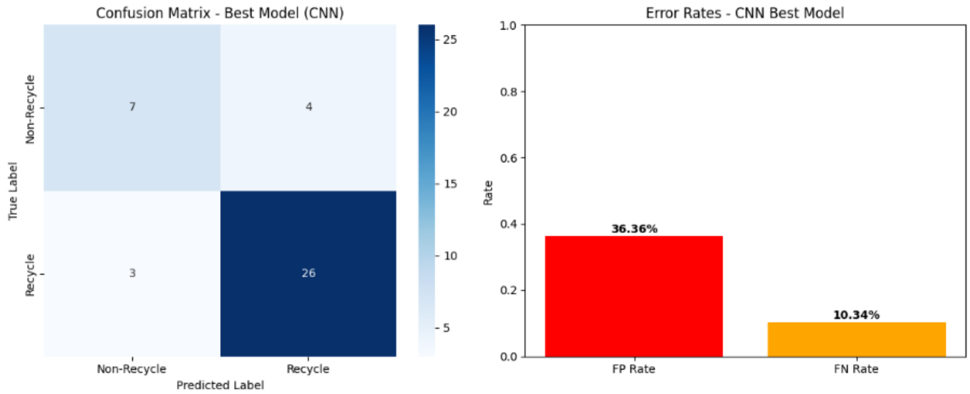

Figure 8. Confusion matrix best model (CNN)

Based on Figure 8, CNN managed to correctly classify 26 out of 29 recyclable waste samples. However, the model showed weakness in the non-recyclable class. Four out of 11 samples were misclassified as recyclable, giving a False Positive Rate of 36.36%. We identified two reasons why misclassification happened: First, some waste items from both classes look visually similar, and Second, the imbalanced imbalance dataset causes the model to lean toward predicting the majority class. On the other hand, the model performed better on the recyclable class. The False Negative Rate was only 10.34%, which is also supported by a recall value of 0.89. Overall, CNN achieved an accuracy of 82.50%. The precision was 0.86 for the recyclable class and 0.70 for the non-recyclable class. This explains that the model requires further treatmenr for handling the minority class.

SHAP analysis was applied to the CNN model to understand the reason behind model's decision. Figure 9 shows the SHAP Interpretation, which divides each waste image into 10 regions (superpixels) and measures each region's contribution to the classification decision. It is clearly seen that, red regions indicate positive contributions toward the recyclable class, while blue regions indicate negative contributions toward the non-recyclable class. The results show that the central waste object region provides the largest positive contribution to the recyclable prediction, while background regions tend to contribute negatively.

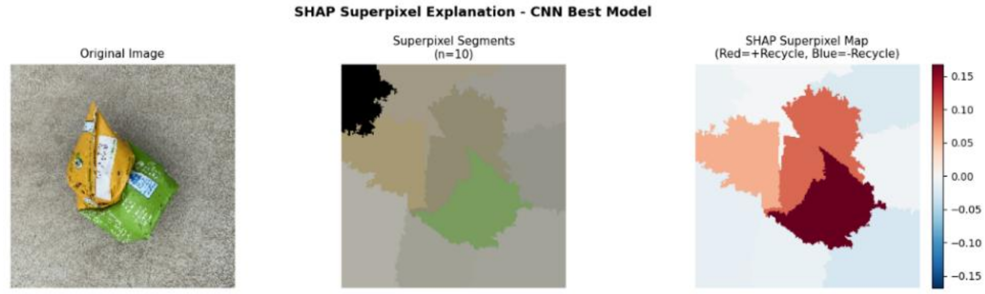

Figure 9. SHAP Superpixel Explanation on CNN Best Model

Figure 10 shows the SHAP bar plot indicating that Region 6 has the highest SHAP value (±0.170), followed by Region 3 (±0.097) and Region 4 (±0.062), indicating the strongest positive contributions of the recyclable class. On the other hand, Region 8 (-0.030) and Region 2 (-0.020) contribute negatively to the non-recyclable class. These findings suggest that the CNN model relies on object-level visual cues such as color, surface texture, and packaging shape, rather than background regions.

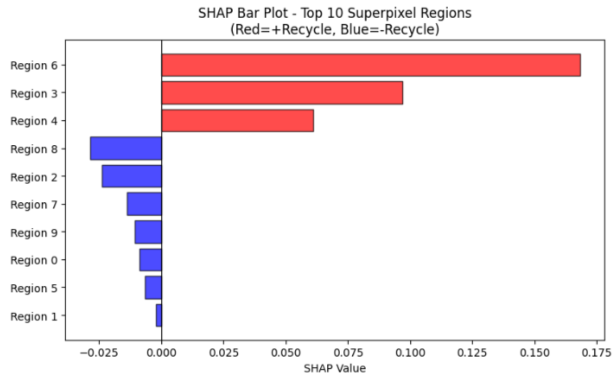

Figure 10. SHAP Contribution per Region (Superpixel) to CNN Model Prediction

To evaluate the contribution of this research more objectively, the results obtained were compared with several studies in the field of waste image classification, as presented in Table 4.

Table 4. Comparison with Previous Research

Study	Dataset Total	Classification Task	Method	Accuracy (%)	XAI
(Huang et al., 2021)	>2.500 images	Multi-class (6 classes)	ViT	96.98	No
(Nisa et al., 2022)	TrashNet	Multi-class	SVM+GLCM	78.87	No
(Muslihati et al., 2024)	Own Dataset	Organic vs Non-organic	CNN	96.43	No
This Study	399 Images	Recycle vs Non-recycle	CNN + SHAP	82.50	Yes

Based on Table 4, this study achieved an accuracy of 82.50% on a very limited dataset (399 images), lower than Huang et al. (2021) and Muslihati et al. (2024) who used much larger datasets with accuracies of 96.98% and 96.43%, respectively. However, the significant difference in dataset scale needs to be considered in interpreting this comparison. This study also outperforms Nisa et al. (2022) who used a similar SVM-based approach with manual features and achieved an accuracy of 78.87%. The main advantages of this study lie in the comprehensive comparative evaluation of six models simultaneously and the integration of SHAP-based Explainable AI (XAI) analysis (Givisis et al., 2025; Saeed & Omlin, 2023), which was not implemented in the comparative studies. This provides additional transparency and interpretability in waste classification model decision-making.

4. Conclusions

This study developed classification models to distinguish between recyclable and non-recyclable waste. The experiments show that CNN achieved the highest testing accuracy of 82.50%, outperforming all classical ML models. These results indicate that CNN using the proposed settings is able to understand visual representations of each class. The SHAP-based XAI provides transparent model interpretation of the most significant features which enhance the interpretability of the system. Through error analysis, the misclassifications were primarily caused by class imbalance, as well as image-level factors such as extreme lighting variation and wet or dirty waste conditions that alter visual appearance. For future work, we plan to focus on collecting larger and more varied instances, applying class balancing techniques through oversampling, class weighting, data augmentation, and exploring transfer learning approaches.

Acknowledgements

The authors thank to Universitas Muhammadiyah Malang and all contributors for providing facilities and resources to support this research.

References

Alhindi, T. J., Kalra, S., Ng, K. H., Afrin, A., & Tizhoosh, H. R. (2018). Comparing LBP, HOG and Deep Features for Classification of Histopathology Images. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–7. https://doi.org/10.1109/IJCNN.2018.8489329

Andi Muhammad Akbar, Muhammad Basri, & Wahyuddin Wahyuddin. (2025). Implementasi Machine Learning Menggunakan Algoritma Klasifikasi untuk Mendeteksi Jenis Sampah. *Jurnal Publikasi Manajemen Informatika*, *3*(3), 152–161. https://doi.org/10.55606/jupumi.v3i3.3751

Asmawati, A., Kurniawan, Y., & Fadilah, R. N. (2025). Efisiensi Pemilahan Sampah Daur Ulang Menggunakan Model CNN SSDMobileNet V2. *ICIT Journal*, *11*(2), 197–208. https://doi.org/10.33050/icit.v11i2.3939

Azhar, N. A., Mohd Pozi, M. S., Mohamed Din, A., & Jatowt, A. (2022). An Investigation of SMOTE based Methods for Imbalanced Datasets with Data Complexity Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/TKDE.2022.3179381

Boumaraf, S., Liu, X., Wan, Y., Zheng, Z., Ferkous, C., Ma, X., Li, Z., & Bardou, D. (2021). Conventional Machine Learning versus Deep Learning for Magnification Dependent Histopathological Breast Cancer Image Classification: A Comparative Study with Visual Explanation. *Diagnostics*, *11*(3), 528. https://doi.org/10.3390/diagnostics11030528

Doly Ilham Saputra Huta Julu, & Dewi Nurdiyah. (2025). KLASIFIKASI SAMPAH ORGANIK DAN NON ORGANIK MENGGUNAKAN TRANSFER LEARNING. *Jurnal Transformatika*, *23*(1), 12–29. https://doi.org/10.26623/transformatika.v23i1.12201

Fotovvatikhah, F., Ahmedy, I., Noor, R. M., & Munir, M. U. (2025). A Systematic Review of AI-Based Techniques for Automated Waste Classification. *Sensors*, *25*(10), 3181. https://doi.org/10.3390/s25103181

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Givisis, I., Kalatzis, D., Christakis, C., & Kiouvrekis, Y. (2025). Comparing Explainable AI Models: SHAP, LIME, and Their Role in Electric Field Strength Prediction over Urban Areas. *Electronics*, *14*(23), 4766. https://doi.org/10.3390/electronics14234766

Gurcan, F., & Soylu, A. (2026). Automated waste classification for smart recycling: A multi-class

CNN approach with transfer learning and pre-trained models. *Environmental Technology & Innovation*, *41*, 104673. https://doi.org/10.1016/j.eti.2025.104673

Hallur, S., & Gavade, A. (2025). Image feature extraction techniques: A comprehensive review. *Franklin Open*, *12*, 100366. https://doi.org/10.1016/j.fraope.2025.100366

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. http://arxiv.org/abs/2106.09685

Huang, K., Lei, H., Jiao, Z., & Zhong, Z. (2021). Recycling Waste Classification Using Vision Transformer on Portable Device. *Sustainability*, *13*(21), 11572. https://doi.org/10.3390/su132111572

Ibnul Rasidi, A., Pasaribu, Y. A. H., Ziqri, A., & Adhinata, F. D. (2022). Klasifikasi Sampah Organik dan Non-Organik Menggunakan Convolutional Neural Network. *Jurnal Teknik Informatika Dan Sistem Informasi*, *8*(1). https://doi.org/10.28932/jutisi.v8i1.4314

Kiyokawa, T., Katayama, H., Tatsuta, Y., Takamatsu, J., & Ogasawara, T. (2021). Robotic Waste Sorter With Agile Manipulation and Quickly Trainable Detector. *IEEE Access*, *9*, 124616–124631. https://doi.org/10.1109/ACCESS.2021.3110795

Kiyokawa, T., Takamatsu, J., & Koyanaka, S. (2024). Challenges for Future Robotic Sorters of Mixed Industrial Waste: A Survey. *IEEE Transactions on Automation Science and Engineering*, *21*(1), 1023–1040. https://doi.org/10.1109/TASE.2022.3221969

Kurniawan, R., Wintoro, P. B., Mulyani, Y., & Komarudin, M. (2023). IMPLEMENTASI ARSITEKTUR XCEPTION PADA MODEL MACHINE LEARNING KLASIFIKASI SAMPAH ANORGANIK. *Jurnal Informatika Dan Teknik Elektro Terapan*, *11*(2). https://doi.org/10.23960/jitet.v11i2.3034

Muslihati, M., Sahibu, S., & Taufik, I. (2024). Implementasi Algoritma Convolutional Neural Network untuk Klasifikasi Jenis Sampah Organik dan Non Organik. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(3), 840–852. https://doi.org/10.57152/malcom.v4i3.1346

Nisa, I. Z., Endah, S. N., Sasongko, P. S., Kusumaningrum, R., Khadijah, K., & Rismiyati, R. (2022). Klasifikasi Citra Sampah Menggunakan Support Vector Machine dengan Ekstraksi Fitur Gray Level Co-Occurrence Matrix dan Color Moments. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *9*(5), 921–930. https://doi.org/10.25126/jtiik.2022954868

Razaque, A., Ben Haj Frej, M., Almi'ani, M., Alotaibi, M., & Alotaibi, B. (2021). Improved Support Vector Machine Enabled Radial Basis Function and Linear Variants for Remote Sensing Image Classification. *Sensors*, *21*(13), 4431. https://doi.org/10.3390/s21134431

S.Intam, R. N. J., Raihan, A., Alfajri, M., Kaswar, A. B., Andayani, D. D., & Asnidar. (2024). Sistem Klasifikasi Jenis Sampah Berdasarkan Kombinasi Fitur Warnac Tekstur Menggunakan Artifical Neural Network Berbasis Pengolahan Citra Digital. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(2), 411–420. https://doi.org/10.25126/jtiik.20241128330

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*, 110273. https://doi.org/10.1016/j.knosys.2023.110273

Sami, K. N., Amin, Z. M. A., & Hassan, R. (2020). Waste Management Using Machine Learning and Deep Learning Algorithms. *International Journal on Perceptive and Cognitive Computing*, *6*(2), 97–106. https://doi.org/10.31436/ijpcc.v6i2.165

Sapanli, K., Putro, F. A. D., Arifin, S. D., Putra, A. H., Andamari, H. A., & Anggraini, U. (2023). Pengelolaan Sampah Rumah Tangga Berbasis Circular Economy di Tingkat Desa: Pendekatan Sistem Dinamik. *Jurnal Wilayah Dan Lingkungan*, *11*(2), 141–155. https://doi.org/10.14710/jwl.11.2.141-155

Sunardi, Yudhana, A., & Fahmi, M. (2023). Improving Waste Classification Using Convolutional Neural Networks: An Application of Machine Learning for Effective Environmental Management. *Revue d'Intelligence Artificielle*, *37*(4), 845–855. https://doi.org/10.18280/ria.370404

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, *57*(4), 99. https://doi.org/10.1007/s10462-024-10721-6