Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 5 (2026) 1404 - 1411 e-ISSN: 2580-0760

Hybrid DenseNet201-ViT Architecture for Robust Solid Waste Classification

Lailia Rahmawati[1], Catur Supriyanto[2]
[1]Informatics Engineering, Universitas Darul 'Ulum, Jombang, Indonesia
[2]Department of Informatics Engineering, Universitas Dian Nuswantoro, Semarang, Indonesia
*liaundarjombang@gmail.com

Abstract

Automatic sorting of solid waste is difficult because discarded items are often badly deformed and some material classes, such as transparent plastic and glass, look almost identical. Conventional Convolutional Neural Networks (CNNs) mainly learn local patterns; as a result, they tend to overlook global context and to overfit when only a small and unevenly distributed dataset is available. This study develops a hybrid deep learning model that pairs the dense local feature learning of DenseNet201 with the global context modeling of a Vision Transformer (ViT) in order to make waste classification more robust. An ImageNet-pretrained DenseNet201 serves as the backbone, and its high-level feature maps are converted into a sequence of patch tokens that is processed by a custom ViT encoder. Experiments were conducted on a public garbage image dataset with six unevenly represented classes, namely cardboard, glass, metal, paper, plastic, and trash. On-the-fly data augmentation and early stopping were employed during training to limit overfitting. The hybrid model reached an overall accuracy of 87% and a macro F1-score of 0.85, exceeding the CNN baselines trained under identical settings. It was especially accurate on fibrous materials, with precision values of 0.95 for cardboard and 0.89 for paper. These results indicate that coupling self-attention with densely connected convolutional layers alleviates the main weaknesses of pure CNNs and provides an accurate, interpretable, and computationally efficient basis for practical automated waste sorting systems.

Keywords: CNN; deep learning; DenseNet201; hybrid architecture; waste classification

How to Cite: L. . Rahmawati and C. . Supriyanto, "Hybrid DenseNet201-ViT Architecture for Robust Solid Waste Classification", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 5, pp. 1404–1411, Oct. 2026, doi: 10.29207/resti.v10i5.7669.
Permalink/DOI: https://doi.org/10.29207/resti.v10i5.7669

Received: March 25, 2026
Accepted: September 19, 2026
Available Online: October 12, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Global solid waste generation continues to rise, creating a need for waste management systems that are efficient, sustainable, and automated [1]. Sorting waste by hand demands considerable labor and time, and its outcome depends heavily on human accuracy [2]. For this reason, computer vision based on deep learning has become an attractive way to recognize and categorize garbage automatically. Beyond speeding up recycling, automated sorting also lowers the environmental and health risks that arise when waste is disposed of improperly [3].

Convolutional Neural Networks (CNNs) have been widely adopted for this image recognition problem. Earlier works relied on standard pre-trained models, including VGG-16, ResNet, and DenseNet, and obtained moderate results under controlled conditions. A closer look at these studies, however, shows that they differ considerably in their datasets and training settings. Kolla et al. [4], for example, fine-tuned a pre-trained ResNet50 with data augmentation on 2,527 images covering cardboard, plastic, paper, metal, glass, and trash, and their accuracy leveled off at 81.00%. Nnamoko et al. [5] built a five-layer CNN for a two-class task (organic versus recyclable) using augmented data and 50 training epochs, reaching 80.88%. Liu et al. [6] applied transfer learning with VGG-16 to 12,873 images from four material classes (glass, metal, paper, plastic) and obtained at most 84.60%. Comparable ceilings were reported for an IoT-based CNN system targeting household paper, plastic, and glass [7] and for a lightweight MobileNetV2 model trained on four domestic waste categories [8], namely 84.67% and 82.92%, respectively. These figures quantitatively illustrate the shortcomings of conventional

architectures. Their common weakness is methodological: each model depends only on local receptive fields and therefore emphasizes isolated textures rather than the overall spatial layout of an object. This bias is problematic in practice, because waste items are frequently crushed, folded, or torn, which substantially changes their original shape. Without a component that models global relationships, CNN-based methods also find it hard to separate visually similar materials, for instance transparent plastic containers whose reflections resemble those of glass bottles [9], [10]. Moreover, standard CNNs easily overfit and generalize poorly when trained on small and imbalanced datasets, a situation that is typical for minority categories such as residual trash [11].

The Vision Transformer (ViT) was proposed to overcome the restricted view of local feature extraction; through self-attention, it relates image patches to one another and thus captures long-range dependencies and global context. Training a ViT from scratch, however, requires very large amounts of data, which is unrealistic for the relatively small garbage datasets currently available. A hybrid design therefore offers a more practical and efficient alternative. This choice is supported by recent evidence that coupling a CNN backbone with Transformer layers yields better results than either network alone, since the combination exploits both dense local features and global sequence modeling [12], [13]. DenseNet201 was chosen as the feature extractor instead of other common backbones such as VGG or ResNet because of its dense connectivity. Prior studies report that this connectivity improves the flow of information across layers and retains fine, low-level textures at high resolution, which is essential for recognizing the complex deformations found in solid waste [14], [15]. When the projected DenseNet201 feature maps are supplied to a ViT encoder, the network can interpret fine local detail and the global structure of an object at the same time. In this way, the feature representation is enriched without requiring a huge dataset for the model to converge.

Accordingly, this study aims to design, implement, and critically assess a hybrid deep learning model in which DenseNet201 extracts features and a Vision Transformer encodes them as a sequence, with the goal of classifying solid waste accurately. The work targets an open problem in the literature, namely the difficulty of handling deformed objects and visually similar classes in strongly imbalanced data. Two contributions are offered. First, a robust hybrid CNN–Transformer framework is developed specifically for challenging waste images. Second, its advantage over conventional baselines is verified experimentally, with augmentation pipelines and early stopping used to control overfitting [16], [17]. Overall, the proposed framework is expected to serve as an interpretable and computationally feasible tool that makes automated waste sorting more dependable.

2. Methods

The proposed approach combines a Dense Convolutional Network (DenseNet201) and a Vision Transformer (ViT) in a single hybrid model that operates directly on raw images. Figure 1 summarizes the overall research procedure, which consists of dataset preparation, spatial feature extraction, global sequence encoding, and the configuration of the classification stage.

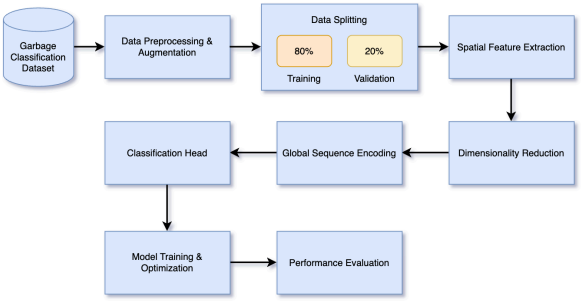


Figure 1. Research workflow of the proposed hybrid DenseNet201-ViT approach.

As Figure 1 shows, the procedure runs through a chain of connected stages, from collecting the raw data to the final evaluation. In the first stage, the public Garbage Classification dataset [18] is obtained; it contains 2,527 images grouped into six waste classes: cardboard, glass, metal, paper, plastic, and residual trash. In the preprocessing and augmentation stage, every image is resized to 224 × 224 pixels so that it fits the network input. To make the model more robust and to imitate the physical deformations encountered in practice, random horizontal and vertical flips as well as random rotation and zoom changes of 20% each are applied on the fly. The data are then split into an 80% training portion (2,022 images), used for learning, and a 20% validation portion (505 images), reserved for an unbiased assessment of how well the model generalizes [19].

In the feature extraction stage, the augmented training images pass through a DenseNet201 backbone that was pre-trained on ImageNet and whose weights are kept frozen. Owing to its dense connections, this backbone produces rich high-level spatial and texture descriptors of the waste images [20]. The dimensionality is then reduced to keep the subsequent Transformer computations efficient: a 1x1 convolution projects the feature maps from 1920 to 256 channels, after which the spatial grid is flattened into 49 visual patch tokens. In the global sequence encoding stage, these tokens are processed by the ViT encoder blocks, whose Multi-Head Self-Attention (MHSA) layers model the contextual relations and long-range dependencies among different regions of a deformed waste object [21].

Next, the classification head aggregates the encoded sequence with a one-dimensional Global Average Pooling layer, applies Dropout (rate 0.3) for regularization, and passes the result to a 128-unit Dense layer with GELU activation. A final Softmax layer

produces the probabilities of the six waste classes. The hybrid network is trained with the Adam optimizer. Two callbacks are employed to avoid overfitting and to support stable convergence on the imbalanced data: ReduceLROnPlateau, which adapts the learning rate, and EarlyStopping, which ends training automatically [22].

In the last stage, the best trained model is assessed on the held-out validation subset. Its predictive performance is quantified by Accuracy, Precision, Recall, and F1-Score, and a Confusion Matrix is used to examine the errors of each class in detail.

2.1 Dataset Preparation and Advanced Augmentation

This study employs a public solid waste image dataset consisting of 2,527 color images [18]. The images are spread unevenly across six classes: cardboard (403), glass (501), metal (410), paper (594), plastic (482), and residual trash (137). Figure 2 shows example images that reflect the appearance and typical deformations of each class. Every image was resized to 224 × 224 pixels with three RGB channels to match the input size required by the network.

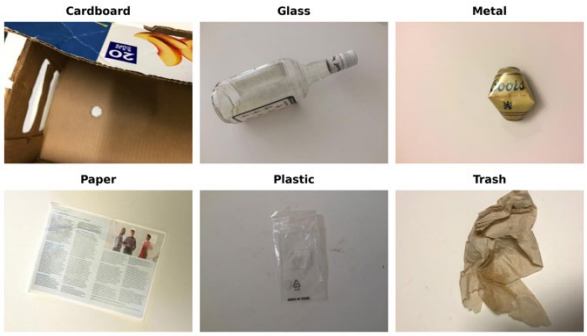


Figure 2. Example images from each of the six solid waste classes.

A stratified split was used to divide the data into a training subset of 80% (2,022 images) and a validation subset of 20% (505 images), so that every class keeps the same proportion in both subsets and generalization can be measured on images the model has never seen [23]. As a result, the validation subset contains 81 cardboard, 100 glass, 82 metal, 119 paper, 96 plastic, and 27 residual trash images. Importantly, the split was performed before any augmentation, which rules out data leakage between the two subsets.

To reduce the influence of class imbalance and improve generalization, an augmentation pipeline was built into the network itself. This augmentation is applied uniformly to every training batch regardless of class, and no class-specific oversampling was used. The augmentation layer performs random horizontal and vertical flipping together with random rotation, zoom, and contrast changes, each with a factor of 20%. Acting as a regularizer, these transformations encourage the network to learn features that remain stable under severe deformation, while the validation metrics remain free from contamination.

2.2 Spatial Feature Extraction using DenseNet201

In the proposed model, the first stage is responsible for extracting spatial features. DenseNet201 pre-trained on ImageNet was adopted for this purpose because of its efficient dense connectivity [24]. In contrast to conventional CNNs, DenseNet links every layer to all subsequent layers in a feed-forward manner, which improves information flow and reduces the vanishing gradient problem. Formally, let x_l denote the output of the l^{th} layer. Within a dense block, each layer takes as input the feature maps produced by all of its preceding layers, as expressed in Equation 1.

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (1)$$

Here, $[x_0, x_1, \dots, x_{l-1}]$ is the concatenation of the feature maps generated by layers 0 through $l-1$, while $H_l(\cdot)$ is a composite function that applies Batch Normalization, a Rectified Linear Unit (ReLU) activation, and a convolution in sequence [25].

The original classification layers of DenseNet201 were discarded, and the pre-trained weights were kept fixed. The last convolutional block of the backbone outputs a high-level feature map of size $7 \times 7 \times 1920$. Because this size would make the following Transformer block computationally expensive, a 1×1 convolution was used to project the 1920 channels down to 256, producing a compact spatial representation of $7 \times 7 \times 256$.

2.3. Vision Transformer (ViT) Sequence Encoding

Although DenseNet201 is very effective at capturing local texture, it has limited ability to model global spatial relationships. Therefore, the $7 \times 7 \times 256$ feature map was flattened into a patch sequence that acts as the input of the ViT encoder. Reshaping the grid yields 49 patches in sequence, and each patch is described by a 256-dimensional vector [26].

The central component of the ViT is Multi-Head Self-Attention (MHSA), which lets the model assign global importance weights to different waste features; for example, it can learn that the reflection on the neck of a bottle is related to the curvature of the bottle base [27]. Self-attention produces its output sequence from three projections, namely queries (Q), keys (K), and values (V), according to Equation 2.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

d_k denotes the key dimension, which serves as a scaling factor that keeps the gradients stable. The custom Transformer block designed in this study comprises layer normalization, an MHSA layer with 4 attention heads, dropout with a rate of 0.2, and a feed-forward network whose hidden layer has 512 units and uses the Gaussian Error Linear Unit (GELU) activation.

2.4 Proposed Hybrid Architecture and Optimization

Figure 3 depicts the complete configuration of the proposed hybrid DenseNet201-ViT model. The design

couples a Convolutional Neural Network (CNN) with Vision Transformer (ViT) layers so that raw images are processed in one continuous pipeline.

This pipeline is composed of several stages executed one after another. Initially, the preprocessed images enter the pre-trained DenseNet201 backbone, which captures the intricate local spatial textures of waste items. A 1x1 convolutional projection then links the CNN backbone to the Transformer by compressing the feature maps from 1920 to 256 channels. Finally, the spatial grid is flattened into a series of 49 visual patch tokens [28], [29].

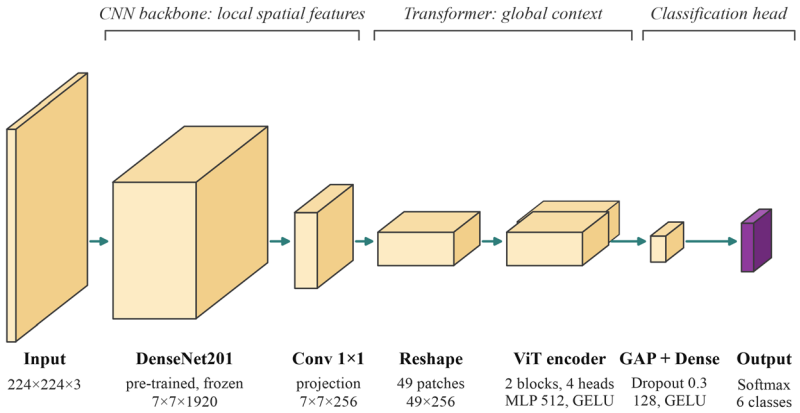


Figure 3. Architecture of the proposed hybrid DenseNet201-ViT model.

The resulting token sequence is fed to two stacked ViT encoder blocks. Through MHSA, these blocks learn global context and long-range relations among the different structural parts of a waste item. The output of the second encoder block is then summarized by a one-dimensional Global Average Pooling layer, which condenses the sequence into a single feature vector. A dropout layer with a rate of 0.3 provides additional regularization and is followed by a 128-unit dense layer with GELU activation. In the last layer, a dense layer with softmax activation outputs the class probabilities for the six waste categories [30], [31].

All experiments were run with TensorFlow on Google Colaboratory using an NVIDIA T4 GPU. For reproducibility, the random seed of the data pipeline was fixed at 42 and the batch size was set to 32. The hybrid architecture contains 21,479,110 parameters in total, of which 3,157,126 are trainable; these belong to the 1x1 convolutional projection, the ViT encoder blocks, and the classification head. The remaining 18,321,984 parameters of the DenseNet201 backbone were fully frozen. Freezing the whole backbone, rather than fine-tuning part of it, was an intentional choice, because updating such a large convolutional network with only 2,022 training images would quickly lead to overfitting. With frozen weights, the backbone supplies reliable ImageNet-derived low-level textures, while the approximately 3.1 million trainable parameters, most of which reside in the Vision Transformer, concentrate on learning the global structural relationships within waste objects.

Optimization was performed with Adaptive Moment Estimation (Adam) at an initial learning rate of 0.001, using Sparse Categorical Crossentropy as the loss. An adaptive training scheme was adopted to promote stable convergence and to discourage the network from memorizing the training data [32]. Specifically, ReduceLROnPlateau halved the learning rate whenever the validation loss did not improve for 5 epochs, and EarlyStopping tracked the validation loss with a patience of 10 epochs, stopping training early and restoring the best weights to limit overfitting.

3. Results and Discussions

This section reports the evaluation of the proposed hybrid DenseNet201–Vision Transformer (ViT) model, then discusses its behavior and limitations and compares it with baseline and previously published methods. All tests were performed on the 505-image validation subset to assess generalization, classification accuracy, and robustness to class imbalance and to visual similarity between classes.

3.1 Training Dynamics and Optimization Analysis

The hybrid network was trained for up to 50 epochs with the Adam optimizer [33]. The training behavior plotted in Figure 4 and Figure 5, however, shows how important the adaptive learning rate and early stopping were [34], [35]. In the first epochs, training accuracy rose sharply and consistently, suggesting that the DenseNet201 backbone quickly supplied useful local features from which the ViT encoder could build global context. The lowest validation loss, 0.4441, was obtained at epoch 10, together with a validation accuracy of 86.93%.

After epoch 10, a clear gap between training and validation performance started to appear. Training accuracy kept increasing toward 97.58% and training loss approached zero, whereas validation accuracy leveled off and validation loss fluctuated slightly with a rising tendency. Such divergence between the two sets of curves is a well-known sign of the onset of overfitting, in which a complex hybrid model starts to memorize training samples instead of learning features

that generalize. EarlyStopping prevented this from progressing: training was halted automatically at epoch 20, and the weights from epoch 10 were restored. Together with ReduceLROnPlateau, this strategy avoided the severe overfitting that commonly affects standard CNNs trained on small datasets [36], [37].

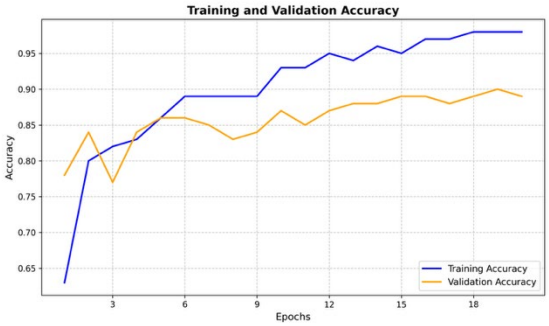


Figure 4. Training and validation accuracy curves.



Figure 5. Training and validation loss curves.

3.2 Classification Performance and Architectural Efficacy

On the unseen validation data, the hybrid model delivered strong classification results [38]. Table 1 lists the precision, recall, and F1-score obtained for every waste class, together with the support, i.e., the number of validation samples per class, so that the class distribution is fully transparent. Across the 505 validation samples, the model reached an overall accuracy of 87.0%. To quantify the statistical reliability of this figure, a 95% Confidence Interval (CI) was computed, giving 87.0% ± 2.9% (84.1% to 89.9%). The macro-averaged and weighted-averaged F1-scores were 0.85 and 0.87, respectively. Because these two averages are close, the model performs consistently across classes; it is not strongly biased toward majority classes such as paper and still achieves acceptable results on minority classes.

Table 1. Hybrid model classification report

Class	Precision	Recall	F1-Score	Support
Cardboard	0.95	0.86	0.90	81
Glass	0.89	0.83	0.86	100
Metal	0.87	0.90	0.88	82
Paper	0.89	0.95	0.92	119
Plastic	0.80	0.85	0.83	96
Trash	0.77	0.68	0.72	27
Accuracy	-	-	0.87	505
Macro Avg	0.84	0.84	0.85	505
Weighted Avg	0.87	0.87	0.87	505

Architecturally, combining a Vision Transformer with a densely connected convolutional backbone proved effective in overcoming the spatial constraints of conventional deep learning models [39]. Pure CNNs depend strongly on local receptive fields, so when a waste item is badly deformed, its local geometric cues are broken and the object is often misclassified [40]. The hybrid model counters this degradation by treating the high-level DenseNet201 feature maps as a sequence of patches. MHSA in the ViT encoder then relates these deformed patches to each other across the whole image, which allows the object to be identified from its global structure rather than from local cues alone [41].

The benefit of this combination is most evident for fibrous materials. Paper obtained the best scores, with a precision of 0.89, a recall of 0.95, and an F1-score of 0.92, while cardboard reached a precision of 0.95, meaning that 95% of the cardboard predictions were correct. The dense connections of DenseNet201 capture the fine, low-level textures of fibers and corrugated layers, and the self-attention of the ViT links these textures globally to form a clear decision boundary between classes.

3.3 Confusion Matrix and Error Analysis

Figure 6 presents the confusion matrix computed on the validation set, which visualizes the classification behavior and reveals where the predictions of the network go wrong.

Although overall accuracy is high, the matrix exposes specific limits of visual recognition. Of the 100 glass samples, 8 were predicted as plastic, and 4 plastic samples were in turn labeled as glass. This two-way confusion reflects an inherent weakness of standard RGB optical sensing: both materials transmit and reflect light in an almost identical way, so separating synthetic polymers from silicate glass without additional data modalities remains difficult for any vision-based method.

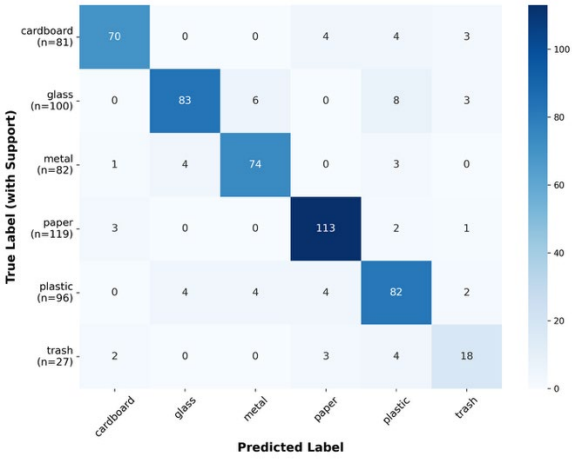


Figure 6. Confusion matrix of the hybrid model.

The weakest results were obtained for the residual trash class. As the matrix shows, 4 of the 27 trash samples were classified as plastic, 3 as paper, and 2 as

cardboard. Residual trash usually consists of crumpled or strongly deformed mixtures of exactly these materials, so the local features extracted by DenseNet sometimes activate patterns linked to those materials. This high intra-class variability, together with the strong class imbalance (27 trash images versus 119 paper images in the validation set), limited how well the Transformer could generalize. A main limitation of the present study is therefore its dependence on standard optical images alone to separate materials with nearly identical textures. Future work should consider multi-modal sensing, for example hyperspectral or near-infrared imaging, which can supply the chemical signatures needed to resolve such visual ambiguities.

3.4 Comparative Analysis with Existing Architectures

To place the performance of the proposed framework in context, an ablation study and a comparison with conventional baseline models were carried out. For a fair comparison, the baselines reported in Tables 2 and 3 (standard DenseNet201 and DenseNet201 + ECA) were re-implemented and trained under the same experimental settings as the hybrid model. All models therefore shared the same 80/20 stratified split, the same random seed (seed=42), the same augmentation pipeline, and the same EarlyStopping and ReduceLROnPlateau settings.

Table 2. Class-wise F1-score comparison.

Class	Baseline (DenseNet201)	DenseNet201 + ECA	Proposed Hybrid (DenseNet201–ViT)
Cardboard	0.84	0.81	0.90
Glass	0.68	0.71	0.86
Metal	0.73	0.72	0.88
Paper	0.84	0.87	0.92
Plastic	0.69	0.69	0.83
Trash	0.63	0.60	0.72

Table 3. Comparison of overall performance metrics.

Model	Accuracy	Macro F1-Score	Weighted F1-Score
Baseline (DenseNet201)	0.75	0.73	0.75
DenseNet201 + ECA	0.76	0.74	0.76
Proposed Hybrid (DenseNet201–ViT)	0.87	0.85	0.87

Table 4. Comparison with existing methods.

Reference	Methodology	Dataset Description	Accuracy
Nnamoko et al. [5]	5-Layer CNN	Organic and Recyclable	80.88%
Liu et al. [6]	VGG-16	Metal, Plastic, Paper and Glass	84.60%
Kolla et al. [4]	ResNet50	TrashNet	81.00%
Rijah et al. [7]	VGG-16	Plastic, Paper, Glass	84.67%
Yong et al. [8]	MobileNetV2	Recyclable, Kitchen, etc.	82.92%
Proposed Study	Hybrid DenseNet201-ViT	Garbage Classification	87.0%

Compared with earlier work on automated garbage classification, the proposed hybrid model performs better. Table 4 shows that recent studies mostly used standard pre-trained CNNs, for example VGG-16 [6], [7] and ResNet50 [4], or custom convolutional networks [5], with reported accuracies between 80.88% and 84.67%. A lightweight MobileNetV2 model likewise did not exceed 82.92% [8]. Since these studies used different datasets and class definitions, this comparison should be regarded as indicative rather than strictly direct.

Examining the baseline models reveals a recurring performance limit. Purely convolutional networks had considerable difficulty with heterogeneous and shapeless categories, most notably residual trash (F1 = 0.63) and plastic (F1 = 0.69), as the baseline DenseNet201 results in Table 2 show. Adding a lightweight Efficient Channel Attention (ECA) module improved performance only slightly, raising overall accuracy to 76% (Table 3), and did not resolve the difficult classes. Because these networks rely only on local receptive fields, they tend to overfit strongly over longer training as they start to memorize the training data.

By contrast, the hybrid approach proposed here raises classification accuracy to 87.0%. Coupling the DenseNet201 backbone with the self-attention of the Vision Transformer closes the gap between local texture recognition and global context understanding. Table 2 shows that the hybrid model reached F1-scores of 0.83 for plastic and 0.72 for residual trash, and Table 3 shows that its overall accuracy is more than 11 percentage points higher than that of both baselines. These gains confirm that global attention effectively compensates for the weaknesses of standard CNNs. In practical terms, recycling facilities could reduce manual sorting bottlenecks by adopting attention-based deep learning. Future studies are encouraged to reduce the computational cost of the hybrid architecture so that it can run in real time on edge devices within smart-city waste management systems.

4. Conclusions

This study shows that combining global attention with densely connected convolutional networks addresses the spatial and contextual limitations of conventional deep learning methods for solid waste classification. The proposed hybrid framework, which joins a pre-trained DenseNet201 backbone with a Vision Transformer (ViT) encoder, copes well with strong physical deformation and high visual similarity between classes in imbalanced data. In the experiments, the model achieved an overall accuracy of 87.0% and a macro F1-score of 0.85, clearly outperforming the standalone convolutional baselines. By pairing dense spatial feature extraction with global sequence encoding, the architecture balances local texture recognition with an understanding of global context. In addition, adaptive learning rate scheduling and early

stopping kept in check the overfitting that usually appears when purely convolutional models are trained on small and heterogeneous waste datasets. The model was especially accurate on texture-rich fibrous materials such as paper and cardboard, but the results also revealed the limits of standard optical sensors. Confusion between transparent synthetic polymers and silicate glass, as well as the difficulty of representing shapeless residual trash, indicates that future research should incorporate multi-modal sensing to resolve such ambiguities. Even so, the hybrid model provides an accurate, interpretable, and scalable computational framework. Its main contribution is a theoretical and practical basis for building reliable, efficient, and automated sorting systems for smart-city recycling facilities, thereby supporting sustainable waste management.

Declarations

AI Usage Statement

No AI tools were used in the preparation of this manuscript.

Author Contributions

L.R.: conceptualization, methodology, software, data curation, formal analysis, writing – original draft; C.S.: supervision, validation, writing – review and editing.

Funding Statement

This research received no external funding.

Conflict of Interest Statement

The authors declare no conflict of interest.

Ethical Clearance

This study used only a publicly available image dataset of waste objects and did not involve human participants or animal subjects; therefore, ethical approval was not required. The authors confirm that the research was conducted in accordance with applicable ethical standards for scientific research.

Data and Code Availability

The Garbage Classification dataset analyzed in this study is publicly available on Kaggle at https://doi.org/10.34740/KAGGLE/DS/81794 [18]. The source code used to build and evaluate the proposed model is available from the corresponding author upon reasonable request.

References

[1] N. A. Alsadhan, "DeepWasteSort-SI-SSO: A Vision Transformer-Based Waste Image Classification Framework Optimized with Self Improved Sparrow Search Optimizer," *Sustainability*, vol. 18, no. 4, p. 2080, 2026, doi: 10.3390/su18042080.

[2] F. Y. Manik *et al.*, "Explainable CNN-Based Multiclass Household Waste Classification Using Grad-CAM for Smart Waste Management," *Int. J. Adv. Comput. Sci. Appl.*, vol. 17, no. 1, 2026, doi: 10.14569/IJACSA.2026.0170155.

[3] C. Dewi, R. A. Susanto, H. P. Chernovita, M. V. V. Santoso, and A. P. S. Chen, "Evaluation of Garbage Detection Efficacy Utilising Deep Learning," *J. Inf. Commun. Technol.*, vol. 24, no. 4, pp. 36–57, Oct. 2025, doi: 10.32890/jict2025.24.4.2.

[4] N. S. Kolla, M. Anumula, S. Sujana, and M. Ratnababu, "Road Garbage Classification Using ResNet50," in 2023 Int. Conf. Signal Process. Comput. Electron. Power Telecommun. (IConSCEPT), May 2023, pp. 1–5, doi: 10.1109/IConSCEPT57958.2023.10170073

[5] N. Nnamoko, J. Barrowclough, and J. Procter, "Solid Waste Image Classification Using Deep Convolutional Neural Network," Infrastructures, vol. 7, no. 4, p. 47, 2022, doi: 10.3390/infrastructures7040047.

[6] K. Liu and X. Liu, "Recycling Material Classification using Convolutional Neural Networks," in 2022 21st IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Nassau, Bahamas, 2022, pp. 83-88, doi: 10.1109/ICMLA55696.2022.00019

[7] U. L. M. Rijah and P. K. W. Abeygunawardhana, "Smart Waste Segregation for Home Environment," in 2023 3rd Int. Conf. Adv. Res. Comput. (ICARC), 2023, pp. 184-189, doi: 10.1109/ICARC57651.2023.10145659

[8] L. Yong, L. Ma, D. Sun, and L. Du, "Application of MobileNetV2 to waste classification," PLoS ONE, vol. 18, no. 3, 2023, doi: 10.1371/journal.pone.0282336

[9] T. Tabata, "Environmental and Social Impacts of Community-Based Household Plastic Waste Collection for High-Value Recycling," *Appl. Sci.*, vol. 16, no. 3, p. 1326, 2026, doi: 10.3390/app16031326.

[10] N. Umar, B. Asrul, and Y. Yuyun, "Type Deep Learning Model for Multi-Label Waste Classification in Canal Environments: A Comparative Study with CNN Architectures," *J. Appl. Data Sci.*, vol. 7, no. 1, pp. 261-276, 2025, doi: 10.47738/jads.v7i1.1066.

[11] A. Alkılınç, F. Y. Okay, İ. Kök, and S. Özdemir, "Deep Ensemble Learning Model for Waste Classification Systems," Sustainability, vol. 18, no. 1, p. 24, 2026, doi: 10.3390/su18010024.

[12] D. Kumar *et al.*, "Optimizing Litter Detection in Images: An Integration of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs)," *IEEE Access*, vol. 14, pp. 3325-3344, 2026, doi: 10.1109/ACCESS.2025.3646075.

[13] Z. Lv, Y. Diao, C. Zeng, W. Wang, and S. An, "Research on Encrypted Transmission and Recognition of Garbage Images in Low-Illumination Environments," *Electronics*, vol. 15, no. 2, p. 302, 2026, doi: 10.3390/electronics15020302.

[14] M. Tarequzzaman, M. Akash, Z. Nayon, M. S. Reza, and S. Haque, "Enhancing waste management through municipal solid waste classification: a convolutional neural network approach," *IAES Int. J. Artif. Intell.*, vol. 14, no. 6, pp. 4775-4786, 2025, doi: 10.11591/ijai.v14.i6.pp4775-4786.

[15] H. C. Yoong, S. L. Jong, and K. G. Tay, "Waste Classification Using NasNet-Mobile: A Multi-Stage Deep Learning Approach for Environmental Sustainability," *Indones. J. Electr. Eng. Informatics*, vol. 13, no. 4, pp. 952-967, Dec. 2025, doi: 10.52549/ijeei.v13i4.7198.

[16] H. T. N. Le and H. Q. T. Ngo, "Application of the vision-based deep learning technique for waste classification using the robotic manipulation system," *Int. J. Cogn. Comput. Eng.*, vol. 6, pp. 391-400, 2025, doi: 10.1016/j.ijcce.2025.02.005.

[17] J. Cui *et al.*, "Research on optimization of waste sorting and transportation network in smart cities based on garbage volume prediction," *Discov. Comput.*, vol. 28, p. 54, 2025, doi: 10.1007/s10791-025-09537-x.

[18] Cchang, "Garbage Classification," Kaggle, 2018, doi: 10.34740/Kaggle/DS/81794.

[19] I. Wiguna, I. Desnanjaya, and I. Sandika, "Performance evaluation of pre-trained deep learning model on garbage classification with data augmentation approach," *IAES Int. J. Artif. Intell.*, vol. 14, no. 6, pp. 4971-4981, 2025, doi: 10.11591/ijai.v14.i6.pp4971-4981.

[20] V. Lingamuthu *et al.*, "ARCNet: A Novel Deep Learning Model for Robust Solid Waste Image Classification in Waste Management Systems," *Int. Arab J. Inf. Technol.*, vol. 22, no. 6, Nov. 2025, doi: 10.34028/iajit/22/6/6.

[21] S. Shetty, S. Kallianpur, R. Fernandes, A. P. Rodrigues, and V. Padmanabha, "ECO-HYBRID: Sustainable Waste

Classification Using Transfer Learning with Hybrid and Enhanced CNN Models," *Sustainability*, vol. 17, no. 19, p. 8761, 2025, doi: 10.3390/su17198761.
[22] G. Celik, "Multi-layer feature fusion for high-accuracy solid waste classification using a hybrid deep learning model," *Vis. Comput.*, vol. 41, pp. 10227–10249, 2025, doi: 10.1007/s00371-025-04031-3.
[23] R. G. Tiwari and V. K. Shukla, "Hybrid Squeeze-and-Excitation Convolutional Neural Network with Elastic Weight Consolidation for Longitudinal Learning in High-Accuracy Waste Classification," *J. Online Inform.*, vol. 10, no. 2, pp. 395-406, Dec. 2025, doi: 10.15575/join.v10i2.1628.
[24] V. S. Malagitti, M. C. Elemmi, S. Jagadeesha, and P. Kiran, "Cognitive Approach and Modified Vision Transformer for Improved E-Waste Management System," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 6, 2025, doi: 10.22266/ijies2025.0731.35.
[25] S. Potharaju *et al.*, "Optimizing Waste Management with Squeeze-and-Excitation and Convolutional Block Attention Integration in ResNet-Based Deep Learning Frameworks," *J. Artif. Intell. Technol.*, vol. 5, pp. 211–220, 2025, doi: 10.37965/jait.2025.0709.
[26] I. Dawar *et al.*, "A systematic literature review on municipal solid waste management using machine learning and deep learning," *Artif. Intell. Rev.*, vol. 58, p. 183, 2025, doi: 10.1007/s10462-025-11196-9.
[27] O. Chacón-Albero et al., "AI for Sustainable Recycling: Efficient Model Optimization for Waste Classification Systems," Sensors, vol. 25, no. 12, p. 3807, 2025, doi: 10.3390/s25123807.
[28] H. Zhou *et al.*, "Optimizing Backbone Networks Through Hybrid–Modal Fusion: A New Strategy for Waste Classification," *Sensors*, vol. 25, no. 10, p. 3241, 2025, doi: 10.3390/s25103241.
[29] M. Castro-Bello *et al.*, "Convolutional Neural Network Models in Municipal Solid Waste Classification: Towards Sustainable Management," *Sustainability*, vol. 17, no. 8, p. 3523, 2025, doi: 10.3390/su17083523.
[30] M. Castro-Bello *et al.*, "R3sNet: Optimized Residual Neural Network Architecture for the Classification of Urban Solid Waste via Images," *Sustainability*, vol. 17, no. 8, p. 3502, 2025, doi: 10.3390/su17083502.
[31] 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, doi: 10.29207/resti.v8i4.5875.
[32] A. Afifah, E. R. Arumi, M. Maimunah, and S. Nugroho, "Comparative Evaluation of Preprocessing Methods for MobileNetV1 and V2 in Waste Classification," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 9, no. 3, pp. 444-452, 2025, doi: 10.29207/resti.v9i3.6211.
[33] M. Nahiduzzaman *et al.*, "An automated waste classification system using deep learning techniques: Toward efficient waste recycling and environmental sustainability," *Knowledge-Based Syst.*, vol. 310, p. 113028, 2025, doi: 10.1016/j.knosys.2025.113028.
[34] F. R. Sayem et al., "Enhancing waste sorting and recycling efficiency: robust deep learning-based approach for classification and detection," Neural Comput. Appl., vol. 37, pp. 4567–4583, 2025, doi: 10.1007/s00521-024-10855-2.
[35] A. Pandey *et al.*, "Deep Learning Enabled Garbage Classification and Detection by Visual Context for Aerial Images," *Appl. Comput. Intell. Soft Comput.*, vol. 2025, p. 9106130, 2025, doi: 10.1155/acis/9106130.
[36] M. Chhabra et al., "Intelligent waste classification approach based on improved multi-layered convolutional neural network," Multimedia Tools Appl., vol. 83, pp. 84095–84120, 2024, doi: 10.1007/s11042-024-18939-w.
[37] V. K. Yogamadhavan and G. Mannayee, "Evaluation of Performance of Different Optimizers of Convolutional Neural Network in the Classification of Images of Urban Domestic Solid Waste," J. Internet Serv. Inf. Secur., vol. 14, no. 3, pp. 42-77, Aug. 2024, doi: 10.58346/JISIS.2024.I3.004.
[38] C. J. Yi and C. F. Kim, "AI-Powered Waste Classification Using Convolutional Neural Networks (CNNs)," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 10, 2024, doi: 10.14569/IJACSA.2024.0151009.
[39] H. Santoso, I. Hanif, H. Magdalena, and A. Afiyati, "A Hybrid Model for Dry Waste Classification using Transfer Learning and Dimensionality Reduction," Int. J. Informatics Vis., vol. 8, no. 2, pp. 623-634, May 2024, doi: 10.62527/joiv.8.2.1943.
[40] R. Singh et al., "Performance Analysis of Garbage Classification Using Balanced and Unbalanced Dataset with EfficientNetV2B1 Architecture," J. Comput. Sci., vol. 20, no. 10, pp. 1310-1321, 2024, doi: 10.3844/jcssp.2024.1310.1321.
[41] Sunardi, A. Yudhana, and M. Fahmi, "Improving waste classification using convolutional neural networks: An application of machine learning for effective environmental management," Rev. d'Intelligence Artificielle, vol. 37, no. 4, pp. 845-855, 2023, doi: 10.18280/ria.370404.