Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.7774

The Garbage Dataset (GD): A Multi-Class Image Benchmark for Automated Waste Segregation

Suman Kunwar [**DWaste** | ***sumn2u@gmail.com***]

DWaste, Baltimore, 21218, USA

Received: 11 March 2026 • **Accepted:** 05 June 2026 • **Published:** 23 June 2026

Abstract This study introduces the Garbage Dataset (GD), a publicly available image dataset designed to advance automated waste segregation through machine learning and computer vision. It is a diverse dataset that covers 10 categories of common household waste: Metal, Glass, Biological, Paper, Battery, Trash, Cardboard, Shoes, Clothes, and Plastic. The dataset comprises 12,259 labeled images collected through multiple methods, including the DWaste mobile app and curated web sources. The methods included rigorous validation through checksums and outlier detection, analysis of class imbalance and visual separability through PCA/t-SNE, and assessment of background complexity using entropy and saliency measures. The dataset was benchmarked using state-of-the-art deep learning models (EfficientNetV2M, EfficientNetV2S, MobileNet, ResNet50, ResNet101) evaluated on performance metrics and computational efficiency measured through energy consumption (kWh). The results of the experiment indicate that EfficientNetV2S achieved the highest performance with an accuracy of 95.13% and an F1-score of 0.95 while exhibiting a trade-off in energy consumption relative to other architectures. Analysis revealed inherent dataset characteristics including class imbalance, a skew toward high-outlier classes (Plastic, Cardboard, Paper), and brightness variations that require consideration. The main conclusion is that GD provides a valuable real-world benchmark for waste classification research while highlighting important challenges such as class imbalance, background complexity, and environmental trade-offs in model selection that must be addressed for practical deployment. The dataset is publicly released to support further research in environmental sustainability applications.

Keywords: Waste Dataset, Waste Classification, Dataset Analysis, Waste Management, Computer Vision

1 Introduction

Effective waste segregation is a critical bottleneck in global recycling systems. With solid waste generation projected to increase by 73% to 3.88 billion tons annually by 2050 [Kaza *et al*., 2021], and the US dumping over two-thirds of its waste despite high per-capita generation [U.S. Environmental Protection Agency, 2020], automated sorting technologies are urgently needed. Computer vision offers a promising solution, but its development is constrained by the availability of robust, large-scale, and well-characterized image datasets. A review of the relevant literature reveals several existing waste image datasets, each with specific limitations. TrashNet [Thung, 2016] is widely used but lacks class diversity. TACO [Proença and Simões, 2020] focuses on waste in natural environments, while UAVVaste [Kraft *et al*., 2021] provides an aerial perspective. SpotGarbage-GINI [Mittal *et al*., 2016] and Trashbox [Kumsetty *et al*., 2022] are sourced from refined web searches, and there are specialized datasets for items such as cigarette butts [Kelly, 2018], plastics [Bobulski and Piatkowski, 2018] or marine debris [Hong *et al*., 2020; Fulton *et al*., 2019]. A significant gap remains for a large-scale, multi-class dataset focused on common household recyclables and trash, curated with detailed characterization to inform model development and highlight inherent data challenges.

Various deep learning approaches have been used for waste classification. MobileNet and ResNet50 are common benchmarks [Poudel and Poudyal, 2022], and ResNet variants are frequently used [Al-Mashhadani, 2023]. To provide a comprehensive evaluation, we benchmark our dataset with multiple model families selected for their architectural innovations and practical relevance. We include ResNet50 and ResNet101, which utilize skip connections to address vanishing gradients in deeper networks, enhancing optimization without a proportional computational increase [He *et al*., 2016]. MobileNet is incorporated for its efficiency on mobile and edge devices through depthwise separable convolutions [Howard *et al*., 2017]. Finally, we evaluated the EfficientNetV2 models (specifically the S and M variants) for their superior training speed and parameter efficiency compared to earlier architectures [Tan and Le, 2021].

This work is based on the idea that a comprehensively curated and analyzed dataset will not only provide a superior benchmark for waste classification models, but will also illuminate critical data-centric factors, such as class imbalance, background complexity, and visual separability that fundamentally impact real-world performance. To test this, we present the GD and adopt a multi-faceted approach: (1) multi-source data collection and rigorous validation, (2) in-depth statistical and visual analysis to quantify dataset properties, and (3) extensive benchmarking using the aforementioned deep learning architectures using the transfer learning approach. The models are evaluated in both original images and standardized versions (resized to 256×256 and 384×384 pixels) using performance metrics (accuracy, recall, F1-score),

training time, and computational efficiency measured in terms of energy consumption (kWh) using Code Carbon [Courty *et al*., 2024].

This approach is justified by the need to move beyond simple model accuracy comparisons and toward an understanding of how dataset attributes influence practical outcomes, including energy footprint measured in kWh. The principal results suggest that GD is a challenging real-world benchmark where the choice of model architecture significantly outweighs the benefit of simple image resizing, and where the highest accuracy (95.13% with EfficientNetV2S) comes with a measurable energy footprint. The main conclusions are that GD fills an important resource gap and that successful waste classification models must be co-designed with consideration of the data inherent biases and the environmental footprint of the training process.

2 Waste Dataset Comparison

The quality and diversity of datasets play a critical role in the performance and generalization ability of deep learning models for waste management. Table 1 provides an expanded overview of widely used waste-related datasets in the literature, covering classification, detection, and segmentation tasks.

In addition to the datasets summarized in Table 1, publicly available benchmarks such as the Kaggle Garbage Classification (12 classes) dataset [Mohamed, 2021] and the HGI-30 dataset [Li, 2021] were also reviewed. The Garbage Classification dataset is widely used for waste image classification benchmarks due to its curated multi-class structure, while HGI-30 provides greater visual diversity across household garbage categories. However, both were excluded because they rely on curated or web-sourced images (e.g., Kaggle was largely web-scraped), and HGI-30 provides less annotation diversity and environmental variability than real-world datasets considered here.

In contrast, datasets such as TACO Proença and Simões [2020], TrashCan 1.0 Hong *et al*. [2020], and UAVVaste Kraft *et al*. [2021] capture more realistic conditions, including occlusions, cluttered backgrounds, multi-object scenes, and sensor noise, making them more suitable for evaluating real-world robustness.

Overall, existing datasets either focus on narrow waste categories, controlled environments, or specific annotation tasks. These limitations motivate the development of GD, which aims to provide a more diverse and realistic benchmark for household waste classification.

3 The Garbage Dataset (GD)

3.1 Data Collection and Curation

The GD was compiled using a multi-method approach to ensure diversity and real-world applicability, capturing natural waste conditions typically observed in practical sorting environments, including variations in lighting, cluttered indoor and outdoor scenes, different viewpoints, and physically deformed items such as crushed or crumpled waste. The collection involved: (i) capture images of waste items against various backgrounds using mobile cameras and the dedicated DWaste mobile application; (ii) obtain and curate images from publicly available repositories and web scraping; and (iii) accept community submissions. This methodology ensures that the dataset reflects the heterogeneity of real-world waste scenarios. The initial dataset contains 20,212 images: approximately 9,800 (48.5%) from public repositories, 6,400 (31.7%) from the mobile application, and 4,012 (19.8%) from community submissions. Following data cleaning and quality standardization, the images were reduced to 12,259 JPEG images, comprising 4,904 images from public repositories (40.0%), 4,290 from the mobile application (35.0%), and 3,065 from community submissions (25.0%).

The images span several classes shown in Figure 1 and passed through cleaning phases that checked for duplication, integrity, and copyright. The collected images went through two image hashing approaches: MD5 hashing and Perceptual hashing.

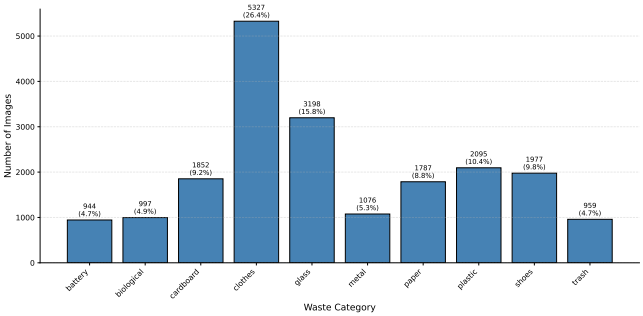

Figure 1. Class distribution of the collected images

MD5 hashing was used to verify the integrity of the file [Shaikh *et al*., 2024] along with the exact duplication, while perceptual hashing was used for images close to duplicate [Zhou *et al*., 2025], as shown in Algorithm 1.

Here, an MD5 hash is computed for every image to identify exact duplicates; if a hash matches a previously seen image, the pair is recorded. If it is unique, a perceptual hash (pHash) is computed after resizing the image to 32×32, providing a compact and standardized representation for perceptual similarity analysis. The pHash is then compared with all previously stored pHashes using the Hamming distance; if the distance is below the empirically selected threshold of 5, the images are considered near duplicates and the pair is recorded. If no near duplicate is found, the pHash is stored for future comparisons. The algorithm outputs exact duplicate pairs and close-to-duplicate pairs. Using image hashing, one exact duplicate and 1,360 near-duplicate images were identified and removed.

Examples of transparent-background images removed during preprocessing are shown in Figure 2. A total of 720 such images were excluded due to the presence of alpha channels, as models are known to struggle with such formats. For physically transparent objects (e.g., glass), similar difficulties have been reported [Wen *et al*., 2025].

In addition, 20 non-RGB images (with color modes P, CMYK, and L) were removed to maintain a uniform input

Table 1. Waste datasets overview

Reference	Dataset Name	Categories	Images	Annotation
Kelly [2018]	Cigarette Butt Dataset	1	2,200	Detection
Haefliger [2020]	DeepSeaWaste	5	3,055	Classification
Serezhkin [2020]	Drinking Waste Classification	4	9,640	Detection
Lynch [2018]	Open Litter Map	11	>100k	Multi-label Classification
Córdova *et al.* [2022]	PlastOPol	1	2,418	Classification/Detection
Mittal *et al.* [2016]	SpotGarbage - GINI dataset	1	2,561	Detection
Proença and Simões [2020]	TACO	28	1,500	Segmentation
Hong *et al.* [2020]	TrashCan 1.0	4	7,212	Instance Segmentation
Fulton *et al.* [2019]	Trash-ICRA19	3	5,700	Detection
Yang and Thung [2016]	Trashnet	6	2,527	Classification
Kumsetty *et al.* [2022]	TrashBox	7	17,785	Classification/Detection
Kraft *et al.* [2021]	UAVVaste	1	772	Segmentation
Bobulski and Piatkowski [2018]	WaDaBa	8	4,000	Classification
Sekar [2019]	Waste Classification data	2	22,500	Classification
Pal [2019]	Waste Classification Data v2	3	∼27,500	Classification
Nnamoko [2023]	Waste Classification Dataset	2	22,500	Classification
Cen [2020]	Waste Images from Sushi Restaurant	16	500	Classification
Sugiyama *et al.* [2022]	BeachLitter Dataset	8	3,500	Classification/Segmentation
Wang *et al.* [2020]	MJU-Waste v1.0	1	2,475	Segmentation

Figure 2. Examples of transparent-background images removed during preprocessing.

Algorithm 1 Duplicate and Near-Duplicate Detection using Hashing

Require: Set of image paths $\mathcal{I}$, Hamming threshold $\tau = 5$
Ensure: Lists $\mathcal{D}_{exact}, \mathcal{D}_{near}$

Initialize dictionaries $\mathcal{H}_{MD5}$, $\mathcal{H}_{pHash}$ and lists $\mathcal{D}_{exact}$, $\mathcal{D}_{near}$
for each $I_i \in \mathcal{I}$ **do**
  $h_{md5} \leftarrow \text{MD5}(I_i)$
  if $h_{md5} \in \mathcal{H}_{MD5}$ **then**
    Append $(I_i, \mathcal{H}_{MD5}[h_{md5}])$ to $\mathcal{D}_{exact}$
  else
    $\mathcal{H}_{MD5}[h_{md5}] \leftarrow I_i$
    Resize I_i to 32×32; $h_p \leftarrow \text{pHash}(I_i)$
    for each $(h_p^j, I_j) \in \mathcal{H}_{pHash}$ **do**
      if $\text{Hamming}(h_p, h_p^j) < \tau$ **then**
        Append (I_i, I_j) to $\mathcal{D}_{near}$
        break
      end if
    end for
    if no match found **then**
      $\mathcal{H}_{pHash}[h_p] \leftarrow I_i$
    end if
  end if
end for
return $\mathcal{D}_{exact}, \mathcal{D}_{near}$

representation and avoid inconsistencies arising from color space differences during conversion. Different color spaces provide different numerical representations, which can affect model performance [Gowda and Yuan, 2019].

Some images contained watermarks, were copyrighted, or had text on the edges. These watermark images are likely to reduce model performance [Yu *et al.*, 2025]. TrustMark was used to detect watermarks [Bui *et al.*, 2023], identifying 493 watermark images. In addition, manual verification was performed to remove copyrighted images, those with text on the edges, and watermarked images. This process reduced the dataset from an initial 18,114 images to 12,259. To reduce severe class imbalance, majority classes such as Glass and Clothes were randomly downsampled by removing excess samples while preserving class diversity. The final dataset remained moderately imbalanced to better reflect real-world waste distributions.

Within a given image, multiple items may appear, but all such items belong to the same annotated class. The dataset captures a wide range of real-world variability, including indoor and outdoor environments, diverse lighting conditions, various background contexts, and physical deformations such as crushed, crumpled, or bent items. This ensures that the images reflect the heterogeneity and challenging conditions encountered in practical waste classification scenarios. However, the single-class-per-image assumption may limit its applicability to highly cluttered real-world scenes where multiple waste categories co-occur. Additionally, while object positioning varies across images, some samples exhibit partial

central bias, which may influence spatial generalization.

The dataset is organized in a flat directory structure, with a main folder containing subdirectories for each class, each housing the respective JPEG files. This structure facilitates straightforward integration with standard machine learning pipelines. Figure 3 illustrates sample images from the dataset.

Figure 3. Example images from the dataset, illustrating variation in object type, scene context, and image quality

3.2 Dataset Statistics and Structure

The dataset comprises 12,259 JPEG images totaling approximately 1.12 GB. A significant class imbalance exists: for instance, the Glass category contains 1,736 images, which is nearly four times the 453 images in the Trash category as shown in Figure 4. This imbalance may bias models toward majority classes without corrective strategies such as oversampling, undersampling, or class-weighted loss functions.

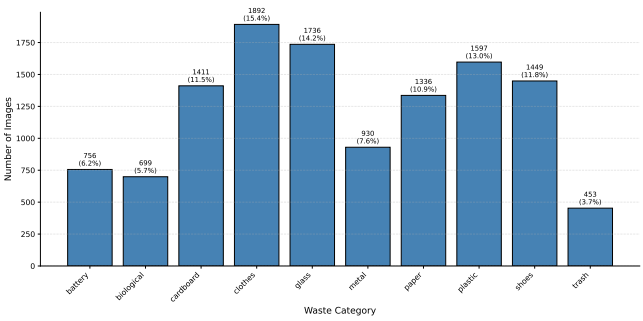

Figure 4. Class distribution of the cleaned dataset

As images are collected from multiple sources, there is notable cross-source variation. Images captured via the DWaste mobile application tend to exhibit consistent resolution, as users are guided to photograph items against neutral backgrounds and are saved with fixed resolution size. In contrast, web-scraped images vary widely in resolution, compression artifacts, and background complexity, ranging from clean studio setups to cluttered real-world scenes.

Community submissions fall between these extremes, often featuring diverse indoor and outdoor environments with uncontrolled lighting and occlusion, as images are collected from recycling centers, parks, and bins. These differences in acquisition settings also result in variations in image resolution across sources, as illustrated in Figure 5. These source differences create potential domain shifts in brightness, color, sharpness, and background context. Although this heterogeneity poses challenges for models trained on homogeneous data [Guo *et al*., 2023], it is precisely this variability that makes the dataset a robust benchmark for real-world generalization [Xiao *et al*., 2024]. Researchers can leverage these cross-source distinctions to study the domain adaptation and robustness of the model, a key to deploying waste classification systems in uncontrolled environments. They are advised to employ standard preprocessing and augmentation, as research confirms that image size [Hieu and Son, 2024] and resolution [Du *et al*., 2025] significantly affect model performance.

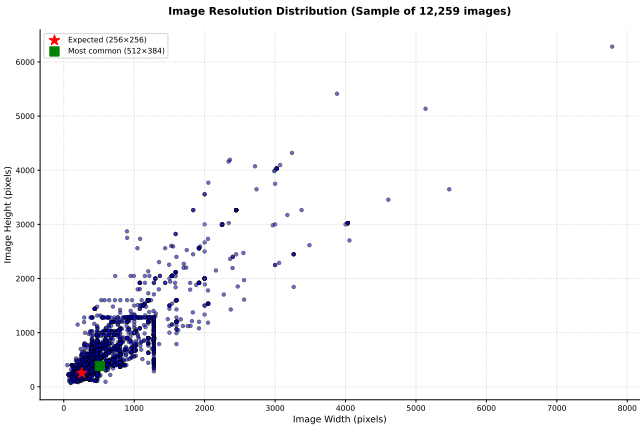

Figure 5. Distribution of image resolutions

To accommodate the input dimensions of the standard model, we provide three versions of the dataset: Original, Standardized_256 [Ghosh *et al*., 2019], and Standardized_384. Normalized versions are created by resizing the images to 256×256 and 384×384 pixels, respectively, using padding to preserve the aspect ratio. Figure 6 shows the data sources, the image counts for each class, and its standardization.

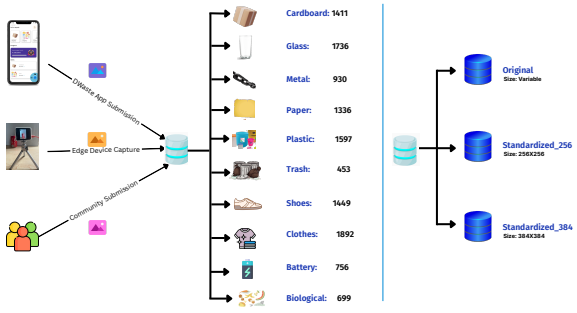

Figure 6. Dataset summary

3.3 Dataset Quality and Validation

Data integrity was verified using checksums and Cyclic Redundancy Check (CRC) [Wada, 2009], confirming that there were no transmission or storage errors. Each image was manually annotated into one of ten predefined categories and verified by at least three volunteers. We ensured that all labels

are accurate and that no personally identifiable or copyrighted images are included. The images collected through the community and the app were obtained with user permission, and a mechanism is provided for users to request the removal of their images if desired.

We define an outlier as an image whose feature representation significantly deviates from the dataset distribution, flagged by an ensemble of unsupervised methods (Isolation Forest, LOF, DBSCAN, K-means distance, Elliptic Envelope) or by being in the top 5% of ensemble scores. Using this definition, 4.3% (527 images) were detected as outliers. Their distribution across classes was very uneven: Plastic (10.8%), Cardboard (6.8%) and Paper (5.8%) had the highest rates, while Clothes (0.2%) and Shoes (0.4%) showed exceptional consistency.

Figure 7, panel (a) shows the PCA projection where darker colors indicate higher outlier scores, highlighting anomalous samples away from the main cluster. Panel (b) compares how many outliers each method detects, revealing varying sensitivities. Panel (c) displays the consensus among methods – most images receive few votes, while true outliers receive three or more. Panel (d) marks the top 15 outliers in PCA space for visual inspection. Panel (e) confirms that Plastic, Cardboard, and Paper have the highest outlier percentages. Panel (f) plots the ensemble score distribution, with a red dashed line at the 95th percentile threshold.

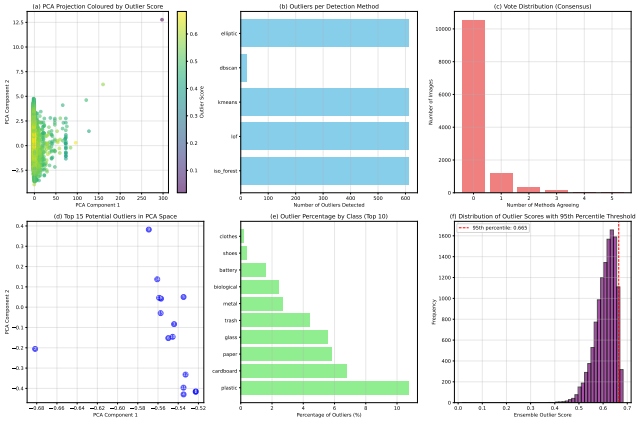

Figure 7. Outlier detection analysis: (a) PCA projection colored by ensemble outlier score; (b) outliers per detection method; (c) vote distribution (consensus among methods); (d) top outliers in PCA space; (e) class-wise outlier percentages; (f) ensemble score distribution with 95th percentile threshold.

4 Dataset Analysis

4.1 Background and Foreground Analysis

Background-object correlation reduces classification accuracy only when training and test domains are mismatched; matched domains maintain high accuracy [Sielemann *et al.*, 2025]. To characterize our dataset, we compute the foreground saliency ratio (pixel ratio) and the Shannon entropy of the background using the grayscale histogram method from [Ma *et al.*, 2025]. The resulting distribution is shown in Figure 8.

Analysis of all 12,259 images revealed high background

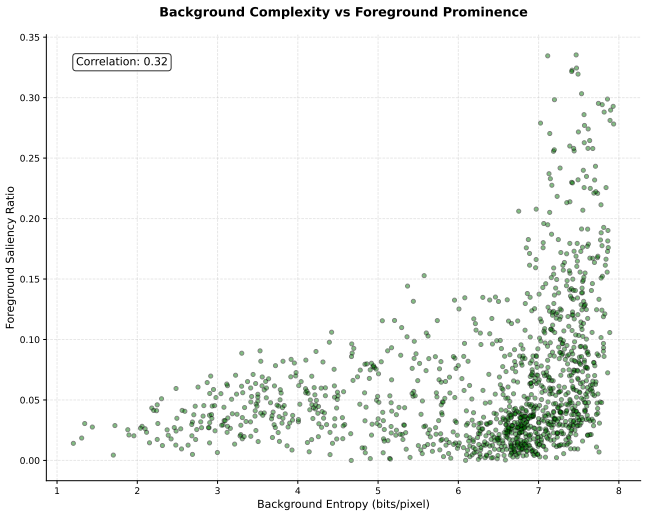

Figure 8. Background entropy vs foreground saliency ratio of images

complexity, with mean Shannon entropy of 6.3 ± 1.5 bits/pixel. This is compounded by a low mean Foreground Saliency Ratio of 0.06 ± 0.06, indicating visually dominant backgrounds. Manual categorization of a 500-image sample found that the backgrounds consisted primarily of indoor floors (35%), outdoor ground (28%), and tables/surfaces (22%).

Saliency detection can improve computational efficiency by focusing on informative regions, reducing search space, and false positives in complex scenes [Qiu *et al.*, 2024]. Furthermore, lighting conditions pose challenges; our analysis indicates a brightness factor skewed toward over-exposure (mean: 1.35), which is likely to reduce model performance due to loss of contrast and degraded feature separability [Wang *et al.*, 2023].

4.2 Class Imbalance and Visual Separability

We assessed visual separability using Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) [Arora *et al.*, 2018] in a balanced subset (400 images per class). PCA revealed poor linear separability, capturing only 31.56% of variance. In contrast, t-SNE showed moderate clustering with a separability ratio of 1.24 (avg distance to the same class: 1.0922, avg distance to a different class: 1.3498) as shown in Figure 9.

The most challenging classes, indicated by high inter-class overlap, were Shoes (0.788), Glass (0.777), Metal (0.759), and Plastic (0.695). Paper and Plastic emerged as the most frequently confused pair, with a recorded centroid distance of 4.12. Contrary to initial assumptions, Biological (0.567) and Trash (0.555) were among the more distinct categories. This complexity confirms the suitability of the dataset not only for image classification but also for research in object detection, data augmentation, transfer learning, and few-shot learning.

5 Benchmark Experiments

5.1 Experimental Setup

The dataset was divided into training sets (80%), validation sets (10%), and test sets (10%) for experimental pur-

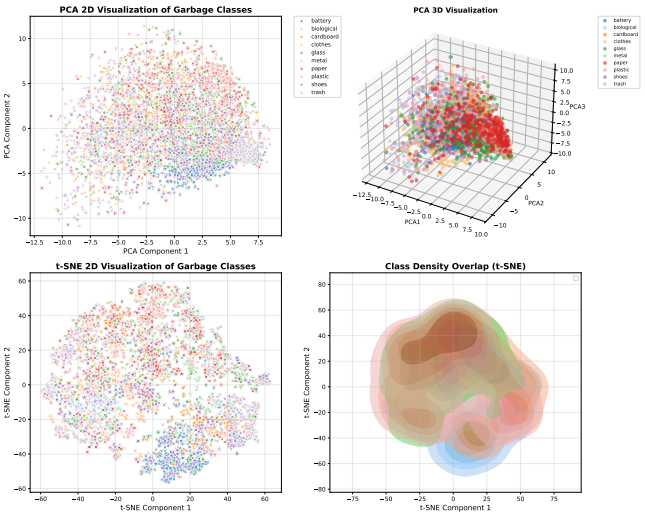

Figure 9. Visualization of PCA/t-SNE of the images

poses. To mitigate class imbalance, random undersampling was applied to the majority of classes during training. All benchmarked models (EfficientNetV2M, EfficientNetV2S, MobileNet, ResNet50, ResNet101) utilized transfer learning from ImageNet weights. Each model was evaluated in three versions of the dataset (Original, Standardized_256, Standardized_384) using a NVIDIA Tesla T4x2 GPU in Kaggle. Performance was measured over 20 epochs using accuracy, recall, and F1-score. Training time and energy consumption were tracked using Code Carbon. These values represent the total energy required for data preparation, training, and inference, reflecting each model's computational cost.

5.2 Results and Discussion

The benchmark results shown in Table 2 reveal EfficientNetV2S as the best performing model, achieving 95.13% accuracy and a 0.95 F1 score with a training time of 6,058 seconds. In contrast, the fastest model, MobileNet_256, trains in 2,480 seconds but sacrifices 28 percentage points in accuracy, scoring only 67.88%. Increasing the input resolution to 384x384 generally provided minimal accuracy gains (under 1%) while significantly increasing the computational cost. Notably, EfficientNetV2 architectures consistently outperformed ResNet models, with EfficientNetV2S exceeding ResNet101's accuracy by 2.36 percentage points despite similar training times.

Further analysis reveals that the minority Trash class (3.7% of the data) consistently yielded the lowest F1-scores across all models. Performance dropped to 0.40 on MobileNetV2 and peaked at only 0.90 on the best-performing model. These results confirm that without targeted interventions, such as class weighting, models remain biased against underrepresented yet critically important categories. Visual overlap between the Paper and Plastic classes remained a challenge, with ResNet101-TL-v2-384 producing identical F1-scores (0.88) for each. Similarly, recall for the Metal category fluctuated wildly (0.49–0.96) depending on the model used. These results indicate that scaling model capacity is not enough; addressing visual ambiguity requires data-centric solutions like specialized augmentation or tailored architectures.

For rapid iteration, EfficientNetV2S_384 offers a compelling alternative, offering 94.50% accuracy with relatively low energy consumption and a fast training time of 4,045 seconds. The results underscore that model architecture selection has a much greater impact on performance than simply increasing input size.

Model selection is also an environmental decision. The least accurate models (MobileNet variants at 66%) exhibit the lowest energy consumption across configurations, reflecting their lightweight design, while the most accurate model (EfficientNetV2S) consistently incurs higher energy consumption across all configurations due to increased model complexity. These values enable a direct trade-off between accuracy and computational cost, where lower kWh indicates reduced energy consumption and better suitability for resource-constrained or large-scale deployments.

For balanced deployments, EfficientNetV2S_384 offers a near-optimal trade-off, reducing energy consumption substantially while maintaining only a minimal reduction in accuracy. Figure 10 illustrates EfficientNetV2M predictions, correctly classifying 2 out of 3 samples. By establishing a baseline using undersampling without augmentation, this study highlights the inherent challenges of the dataset, particularly class imbalance and visual variability.

Figure 10. Predictions on sample image with actual size using EfficientNetV2M model

6 Conclusion

This study presented GD, a comprehensive multi-source image collection designed to advance automated waste segregation. Our main findings demonstrate that GD provides a challenging real-world benchmark characterized by: (1) significant class imbalance, (2) high visual complexity driven by background and lighting variations, and (3) distinct yet overlapping clusters as revealed by t-SNE analysis. These conclusions directly address the problem outlined in the introduction, the lack of a large-scale, well-characterized dataset for household waste classification. The evidence from our systematic analysis and benchmark experiments confirms that while state-of-the-art models like EfficientNetV2S can achieve high accuracy (95.13%), performance is fundamentally constrained by the dataset's inherent properties, particularly class imbalance and background noise.

The implications of this work are both practical and methodological. For practical applications, GD serves as a vital resource for the development of robust waste-sorting systems in recycling facilities, public spaces, and educational tools. Methodologically, our work underscores that achieving real-world robustness requires moving beyond optimizing model

Table 2. Comparative performance of waste classification models.

Model	Time (s)	Accuracy (%)	Recall	F1 Score	Energy Consumption (kWh)		
					Prepare	Develop	Deploy
EN-V2M	7137.81	94.15	0.93	0.93	0.002442	0.312758	0.319029
EN-V2M-256	6169.28	93.61	0.93	0.93	0.002837	0.268021	0.274672
EN-V2M-384	8084.41	94.58	0.94	0.94	0.002475	0.349793	0.356482
EN-V2S	6057.87	95.13	0.95	0.95	0.002492	0.260761	0.266055
EN-V2S-256	5908.89	93.77	0.94	0.93	0.003262	0.253505	0.259520
EN-V2S-384	4044.54	94.50	0.94	0.94	0.002284	0.173350	0.178255
MN	2420.15	66.94	0.65	0.65	0.002205	0.105006	0.108821
MN-256	2480.46	67.88	0.66	0.66	0.001439	0.107556	0.110427
MN-384	2536.73	70.39	0.68	0.68	0.001676	0.109357	0.112462
RN50	5228.93	92.61	0.92	0.92	0.003384	0.225157	0.230935
RN50-256	5830.07	91.91	0.91	0.91	0.002264	0.251346	0.256267
RN50-384	5643.32	92.39	0.92	0.92	0.002716	0.246099	0.251764
RN101	7331.92	92.77	0.93	0.93	0.002506	0.317505	0.323518
RN101-256	5147.28	92.31	0.92	0.92	0.003671	0.222811	0.229675
RN101-384	7445.92	92.64	0.93	0.92	0.002136	0.327963	0.333732

Abbreviations: EN-V2M = EfficientNetV2M, EN-V2S = EfficientNetV2S, MN = MobileNet, RN50 = ResNet50, RN101 = ResNet101.

architecture alone; it necessitates explicit handling of data-centric challenges such as imbalance, background complexity, and computational cost, quantified through energy consumption (kWh).

Future work should focus on addressing the identified limitations through advanced augmentation, imbalance correction techniques, and the development of models that are accurate and computationally efficient. While GD provides a realistic and diverse benchmark, it is still limited by its single-class image annotation and partial structural biases, which may affect performance in fully unconstrained real-world settings. Additionally, the exclusion of transparent-background images introduces a further limitation, as such cases may occur in real-world waste scenarios (e.g., reflective or plastic materials), suggesting the need for models capable of handling such inputs. The release of GD aims to catalyze such research, contributing to scalable and sustainable solutions to global waste management challenges.

Declarations

Authors' Contributions

The author performed all aspects of this study, including the conception and design of the work, data collection, analysis, interpretation of results, and preparation of the manuscript.

Competing interests

The author declares that there are no competing interests.

Acknowledgements

The author thanks the volunteers who assisted in validating the dataset labels and Thierry Haddad for providing images of household and recycling center waste. We also acknowledge the DWaste user community for their image submissions.

Availability of data and materials

The datasets used in this study are available on Kaggle (https://www.kaggle.com/datasets/sumn2u/garbage-classification-v2/). The code used for the experiments is available on GitHub (https://github.com/sumn2u/garbage-dataset-experiments).

References

Al-Mashhadani, I. B. (2023). Waste material classification using performance evaluation of deep learning models. *Journal of Intelligent Systems*, 32(1):20230064. DOI: 10.1515/jisys-2023-0064.

Arora, S., Hu, W., and Kothari, P. K. (2018). An analysis of the t-sne algorithm for data visualization. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1455–1462. PMLR. DOI: 10.48550/arxiv.1803.01768.

Bobulski, J. and Piatkowski, J. (2018). Pet waste classification method and plastic waste database - wadaba. In Choraś, M. and Choraś, R. S., editors, *Image Processing and Communications Challenges 9*, volume 681, pages 57–64. Springer International Publishing, Cham. DOI: 10.1007/978-3-319-68720-9_8.

Bui, T., Agarwal, S., and Collomosse, J. (2023). Trustmark: universal watermarking for arbitrary resolution images. DOI: 10.48550/arXiv.2311.18297.

Cen, A. (2020). Waste images from sushi restaurant. Available at:`https://www.kaggle.com/datasets/arthurcen/waste-images-from-sushi-restaurant`.

Courty, B., Schmidt, V., Luccioni, S., Goyal-Kamal, Marion-Coutarel, Feld, B., Lecourt, J., LiamConnell, Saboni, A., Inimaz, supatomic, Léval, M., Blanche, L., Cruveiller, A., ouminasara, Zhao, F., Joshi, A., Bogroff, A., de Lavoreille, H., Laskaris, N., Abati, E., Blank, D., Wang, Z., Catovic, A., Alencon, M., Stęchły, M., Bauer, C., de Araújo, L.

O. N., JPW, and MinervaBooks (2024). mlco2/codecarbon: v2.4.1. DOI: 10.5281/zenodo.11171501.

Córdova, M., Pinto, A., Hellevik, C. C., Alaliyat, S. A.-A., Hameed, I. A., Pedrini, H., and Torres, R. d. S. (2022). Litter detection with deep learning: a comparative study. *Sensors*, 22(2):548. DOI: 10.3390/s22020548.

Du, X., Sun, Y., Song, Y., Chi, W., Dong, L., and Zhao, X. (2025). Impact of input image resolution on deep learning performance for side-scan sonar classification: an accuracy–efficiency analysis. *Remote Sensing*, 17(14):2431. DOI: 10.3390/rs17142431.

Fulton, M., Hong, J., Islam, M. J., and Sattar, J. (2019). Robotic detection of marine litter using deep visual detection models. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5752–5758. DOI: 10.1109/ICRA.2019.8793975.

Ghosh, S., Das, N., and Nasipuri, M. (2019). Reshaping inputs for convolutional neural network: Some common and uncommon methods. *Pattern Recognition*, 93:79–94. DOI: 10.1016/j.patcog.2019.04.009.

Gowda, S. N. and Yuan, C. (2019). Colornet: investigating the importance of color spaces for image classification. In Jawahar, C., Li, H., Mori, G., and Schindler, K., editors, *Computer Vision – ACCV 2018*, volume 11364, pages 581–596. Springer International Publishing, Cham. DOI: 10.1007/978-3-030-20870-7_36.

Guo, J., Ma, J., García-Fernández, A. F., Zhang, Y., and Liang, H. (2023). A survey on image enhancement for Low-light images. *Heliyon*, 9(4):e14558. DOI: 10.1016/j.heliyon.2023.e14558.

Haefliger, H. (2020). Deepseawaste. Available at:`https://www.kaggle.com/datasets/henryhaefliger/deepseawaste`.

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

Hieu, N. T. and Son, N. H. (2024). The impact of input image size on the performance of deep learning models applied in cybersecurity. In *2024 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 162–167, Danang, Vietnam. IEEE. DOI: 10.1109/RIVF64335.2024.11009067.

Hong, J., Fulton, M., and Sattar, J. (2020). Trashcan: a semantically-segmented dataset towards visual detection of marine debris. DOI: 10.48550/arXiv.2007.08097.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: efficient convolutional neural networks for mobile vision applications. DOI: 10.48550/arXiv.1704.04861.

Kaza, S., Shrikanth, S., and Chaudhary, S. (2021). More growth, less garbage. DOI: 10.1596/35998.

Kelly, A. (2018). Cigarette butt dataset. Available at:`https://www.immersivelimit.com/datasets/cigarette-butts`.

Kraft, M., Piechocki, M., Ptak, B., and Walas, K. (2021). Autonomous, onboard vision-based trash and litter detection in low altitude aerial images collected by an unmanned aerial vehicle. *Remote Sensing*, 13(5):965. DOI: 10.3390/rs13050965.

Kumsetty, N. V., Bhat Nekkare, A., S., S. K., and Kumar M., A. (2022). Trashbox: trash detection and classification using quantum transfer learning. In *2022 31st Conference of Open Innovations Association (FRUCT)*, pages 125–130, Helsinki, Finland. IEEE. DOI: 10.23919/FRUCT54823.2022.9770922.

Li, H. (2021). HGI-30 DATA Set. DOI: 10.5281/zenodo.4646699.

Lynch, S. (2018). Openlittermap.Com – open data on plastic pollution with blockchain rewards(Littercoin). *Open Geospatial Data, Software and Standards*, 3(1):6. DOI: 10.1186/s40965-018-0050-y.

Ma, D., An, Q., and Li, A. (2025). The combined effect of entropy and complexity: Human analysis of AI painting recognition ability. *Systems and Soft Computing*, 7:200378. DOI: 10.1016/j.sasc.2025.200378.

Mittal, G., Yagnik, K. B., Garg, M., and Krishnan, N. C. (2016). SpotGarbage: smartphone app to detect garbage using deep learning. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 940–945, Heidelberg Germany. ACM. DOI: 10.1145/2971648.2971731.

Mohamed, M. (2021). Garbage classification (12 classes). Available at:`https://www.kaggle.com/datasets/mostafaabla/garbage-classification`.

Nnamoko, N. (2023). Waste classification dataset. DOI: 10.17632/N3GTGM9JXJ.3.

Pal, S. (2019). Waste classification data v2. Available at:`https://www.kaggle.com/datasets/sapal6/waste-classification-data-v2`.

Poudel, S. and Poudyal, P. (2022). Classification of waste materials using cnn based on transfer learning. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 29–33, Kolkata India. ACM. DOI: 10.1145/3574318.3574345.

Proença, P. F. and Simões, P. (2020). Taco: trash annotations in context for litter detection. DOI: 10.48550/arXiv.2003.06975.

Qiu, L., Zhang, D., and Hu, Y. (2024). Research on image saliency detection based on deep neural network. *IET Image Processing*, 18(12):3393–3402. DOI: 10.1049/ipr2.13181.

Sekar, S. (2019). Waste Classification data. Available at:`https://www.kaggle.com/datasets/techsash/waste-classification-data`.

Serezhkin, A. (2020). Drinking waste classification. Available at:`https://www.kaggle.com/datasets/arkadiyhacks/drinking-waste-classification`.

Shaikh, M. S., Biswal, A., Pandwal, A., Khodifad, N., and Vaghela, B. (2024). Image forgery detection using md5 & open cv. In *2024 International Conference on Emerging Research in Computational Science (ICERCS)*, pages 1–8. IEEE. DOI: 10.1109/ICERCS63125.2024.10895348.

Sielemann, A., Barner, V., Wolf, S., Roschani, M., Ziehn, J., and Beyerer, J. (2025). Measuring the effect of background on classification and feature importance in deep learning for av perception. In *2025 IEEE In-*

ternational Automated Vehicle Validation Conference (IAVVC), pages 1–8, Baden-Baden, Germany. IEEE. DOI: 10.1109/IAVVC61942.2025.11219547.

Sugiyama, D., Hidaka, M., Matsuoka, D., Murakami, K., and Kako, S. (2022). The BeachLitter dataset for image segmentation of beach litter. *Data in Brief*, 42:108072. DOI: 10.1016/j.dib.2022.108072.

Tan, M. and Le, Q. V. (2021). Efficientnetv2: smaller models and faster training. DOI: 10.48550/arXiv.2104.00298.

Thung, G. (2016). Trashnet dataset and classification repository. Available at:`https://github.com/garythung/trashnet`.

U.S. Environmental Protection Agency (2020). Advancing sustainable materials management: 2018 fact sheet. Available at:`https://www.epa.gov/sites/production/files/2020-11/documents/2018_ff_fact_sheet.pdf`.

Wada, K. (2009). Checksum and cyclic redundancy check mechanism. In Liu, L. and Özsu, M. T., editors, *Encyclopedia of Database Systems*, pages 328–329. Springer US, Boston, MA. DOI: 10.1007/978-0-387-39940-9_1474.

Wang, T., Cai, Y., Liang, L., and Ye, D. (2020). A multi-level approach to waste object segmentation. *Sensors*, 20(14):3816. DOI: 10.3390/s20143816.

Wang, T.-s., Kim, G. T., Kim, M., and Jang, J. (2023). Contrast enhancement-based preprocessing process to improve deep learning object task performance and results. *Applied Sciences*, 13(19):10760. DOI: 10.3390/app131910760.

Wen, H., Zuo, Y., Subramanian, V., Chen, P., and Deng, J. (2025). Seeing and seeing through the glass: real and synthetic data for multi-layer depth estimation. DOI: 10.48550/arXiv.2503.11633.

Xiao, J., Guo, W., Liu, J., and Li, M. (2024). Generalization gap in data augmentation: insights from illumination. DOI: 10.48550/arXiv.2404.07514.

Yang, M. and Thung, G. (2016). Classification of trash for recyclability status. Technical Report CS229, Stanford University. Available at:`https://cs229.stanford.edu/proj2016/report/ThungYang-ClassificationOfTrashForRecyclabilityStatus-report.pdf`.

Yu, J., Liu, X., Zan, F., and Peng, Y. (2025). Robust deepfake detector against deep image watermarking. *PLOS One*, 20(12):e0338778. DOI: 10.1371/journal.pone.0338778.

Zhou, Y., Li, X., Xiong, C., Yao, H., and Qin, C. (2025). A survey of perceptual hashing for multimedia. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(7):1–28. DOI: 10.1145/3727880.