Enhancing Waste Sorting Models with Deep Learning and AI-Generated Synthetic Images*

Iulian Alexandru OGREZEANU[1,2]*, Constantin SUCIU[1,2], Lucian Mihai ITU[1,2]

[1] "Transilvania" University of Brașov, 29 Eroilor Blvd., Brasov, 500036, Romania
iulian.ogrezeanu@unitbv.ro (**Corresponding author*), suciuc@unitbv.ro, lucian.itu@unitbv.ro

[2] Siemens SRL, 15 Noiembrie Blvd., No. 78, Brașov, 500097, Romania

Abstract: This study explores the integration of AI-generated synthetic data into deep learning pipelines for an automated waste detection and classification in industrial recycling systems. Based on the YOLOv12 object detection framework and the publicly available WaRP dataset, the proposed model was trained for recognizing six major waste categories, including bottles, cans, cardboard, and glass. In order to address data scarcity and class imbalance, up to 10,000 synthetic images were generated by using ChatGPT's image generation capabilities, compositing realistic waste objects onto clean conveyor-belt backgrounds. The experimental results demonstrated that augmenting the employed dataset with synthetic samples improved the proposed model`s detection and classification performance, which is proven by an increase from 0.593 (for the baseline model) to 0.622 for the mAP@50 metric and from 0.466 to 0.504 for the mAP@50:95 metric. The best results were achieved for the model augmented with 5,000 synthetic images, after which there was no further improvement in the performance of the employed model. These findings highlight the fact that high-quality synthetic data can effectively enhance deep learning models in waste sorting applications, reducing the dependence on extensive manual data collection. However, for further improvements it would be necessary to enhance the asset realism and diversity rather than simply increasing the dataset size. To sum up, the proposed approach underscores the potential of combining generative AI and computer vision for accelerating industrial automation.

Keywords: Deep learning, Synthetic data generation, Object detection, Waste classification, Recycling automation, Industrial computer vision, Generative AI, Data augmentation, Industry 4.0.

1. Introduction

In recent years, global concern regarding climate change and environmental sustainability has intensified (Zandalinas et al., 2021). Pollution, particularly from unmanaged waste, has prompted coordinated responses from international organizations and policymakers, with waste recycling emerging as a crucial component of the solution. To facilitate recycling, waste sorting plants have been established for preprocessing and separating waste into reusable raw materials. However, conventional methods employed in these facilities continue to rely heavily on manual labor, which introduces significant drawbacks such as increased operational costs, a limited scalability, and unsafe working conditions. These challenges underline the growing necessity of integrating technologies like artificial intelligence (AI), to enhance efficiency and ensure safer working environments.

In this context, deep learning (Dong et al., 2021) has emerged as a transformative tool in the automation of industrial processes. Deep learning, a class of machine learning methods based on multi-layer neural network architectures, has proven effective for learning hierarchical feature representations from large-scale data and has been successfully applied to a wide range of computer vision tasks. When applied to waste sorting systems, these algorithms significantly improve the accuracy of waste detection and classification, thereby optimizing the recycling pipeline.

Recognizing the critical role of waste classification in the recycling process, several researchers have explored the use of AI in this domain. For example, Melinte et al. (2020) utilized enhanced robotic systems for detecting and collecting municipal waste using deep learning algorithms. Building upon such efforts, the present study investigates the use of deep learning models for detecting and classifying waste in sorting facilities. Specifically, it proposes a methodology for training neural networks on annotated datasets to perform classification and object detection for waste items moving along a conveyor belt.

To facilitate the development of these models, a publicly available dataset from Kaggle was employed (Bojer & Meldgaard, 2021), consisting of RGB images of waste on a conveyor belt, captured by a camera mounted above the belt. The dataset enables two primary applications: the full-frame detection of waste along the conveyor stream and

* This article represents an extension of the conference paper: "Automated Waste Sorting Using Deep Learning and Synthetic Data", presented at the CONAT 2024 International Congress of Automotive and Transport Engineering (Part Two: Automobile and Environment).

https://doi.org/10.24846/v35i1y202602

the classification of cropped images representing individual waste items (Yudin et al., 2024).

Furthermore, to address data scarcity and class imbalance challenges which are common in real-world waste sorting applications, the chosen training dataset was enriched with synthetic images (Achicanoy et al., 2021). These were generated using ChatGPT's (Singh et al., 2023) image generation capabilities. This platform enables the generation of diverse synthetic imagery designed to capture a range of object appearances and configurations relevant to waste sorting scenarios, which can support an improved model robustness when combined with real training data.

This research builds upon the previous work of Ogrezeanu et al. (2024), in which an earlier version of the YOLO model was trained for optimizing its performance for waste sorting on a conveyor belt. However, that study did not explore the generative capabilities of modern AI technologies to create new visual assets and construct enhanced synthetic datasets. For the current paper, AI-assisted language tools (mainly ChatGPT) were employed in order to refine readability, clarity, and overall aesthetic of the text. It should be noted that the use of such tools was strictly limited to editorial purposes (e.g. to suggest alternative wording, linking expressions, and stylistic improvements in order to enhance clarity and coherence). All the scientific insights, data analyses, interpretations, and conclusions presented in this work were fully conceived and developed by the authors.

This study relies primarily on a descriptive statistical analysis. While the reported results provide insight into model behavior for the evaluated datasets, no inferential statistical testing was performed. Therefore, the findings should be interpreted as dataset-specific observations rather than statistically generalizable conclusions.

The rest of this paper is organized as follows. Section 2 discusses the related work. Section 3 presents the analyzed dataset and the applied preprocessing methods; it outlines the training procedures and exploratory analyses conducted in the previous work and it introduces the proposed approach for generating a synthetic dataset aimed at enhancing model detection performance. Further on, Section 4 presents the experimental results obtained in this niche. and Section 5 is dedicated to the discussion of the proposed approach and the related experimental results. Finally, Section 6 presents the conclusions of this work and outlines possible future research directions.

2. Related Work

Recent advances in deep learning have significantly influenced waste detection and classification research, driven by the need for automated and scalable recycling solutions. A comprehensive overview of this field is provided by Abdu & Mohd Noor (2022), who survey deep learning architectures, application domains, and benchmark datasets commonly used in waste-related computer vision tasks. Object detection models such as YOLO, SSD, Faster R-CNN, and Mask R-CNN have been widely adopted due to their balance between accuracy and real-time performance, with YOLO-based approaches being particularly suited for industrial and robotic environments. However, despite their steady progress, the object detection performance remains constrained by limited training data, domain variability, and complex scenes involving overlapping objects and changing illumination, challenges that are especially vivid in conveyor-belt sorting systems.

For waste classification, earlier approaches relied on handcrafted features and classical machine learning, while recent studies predominantly employ convolutional neural networks and transfer learning using architectures such as VGG, ResNet, DenseNet, and Inception. Although a high classification accuracy has been reported under controlled conditions, many datasets consist of isolated objects captured against clean backgrounds, limiting generalization to real industrial settings. Nasien et al. (2025), for example, achieved a high accuracy using YOLOv11 on a large, labeled dataset, but without addressing data scarcity or enrichment through synthetic data generation.

The availability of benchmark datasets such as TrashNet, TACO, and TrashICRA19 has facilitated rapid development in the context of waste detection tasks, yet these datasets are often affected by class imbalance, a limited background diversity, and annotation inconsistencies. To address these limitations, data augmentation and synthetic data generation have been explored. Prior work has shown that combining real and synthetic

data can improve robustness, particularly when the real data collection is costly or impractical. Bacchin et al. (2025) demonstrated that GAN-generated synthetic waste images with an enforced semantic coherence can improve the segmentation performance in cluttered environments, while Gautam & Arashpour (2025) reported substantial gains for minority classes in construction and demolition waste segmentation through class-balanced and synthetic augmentation strategies.

Despite these advances, the use of modern, high-fidelity generative AI tools for large-scale synthetic data generation in industrial conveyor-belt scenarios remains underexplored. Most of the existing approaches rely on handcrafted or GAN-based pipelines with a limited diversity and scalability. This gap motivated the present study, which investigates the integration of contemporary generative AI–based synthetic assets to enhance object detection performance under realistic waste-sorting conditions.

3. Materials and Methods

3.1. Data Preprocessing

In this study, a dataset that is available on Kaggle was utilized, namely the Waste Recycling Plant (WaRP) dataset, which contains 28 categories of recyclable materials such as plastic and glass bottles, cardboard, detergent containers, canisters, and aluminum cans. These are organized into six main classes, several of which include multiple subclasses. For instance, plastic bottles are divided into 17 categories based on attributes like color and air content, while glass bottles are split into three subtypes. Cardboard and detergent containers are further separated into two and four categories, respectively. The dataset poses significant challenges for an automated analysis due to factors such as overlapping objects, deformations, and varying lighting conditions.

The dataset is divided into three subsets: WaRP-D (Detection), WaRP-C (Classification), and WaRP-S (Segmentation). This study focuses on the WaRP-D and WaRP-C subsets, which are relevant for object detection (Zhao et al., 2019) and classification tasks. The WaRP-S subset, intended for testing weakly supervised segmentation methods, was excluded as it lies beyond the scope of this research. The WaRP-D subset represents the core of this analysis, comprising 2,452 training images and 522 testing images, each with a resolution of 1920 × 1080 pixels. Every image is paired with a corresponding ".txt" annotation file that defines the bounding boxes and class labels for all the detected objects (see Figure 1 as an example).

Figure 1. Images from the WaRP-D dataset (Anon, n.d.)

As in real industrial environments, a notable class imbalance is present (Figure 2). For example, the least represented class is *canister,* while the most represented class is *bottle* with over 6000 instances appearing throughout the dataset.

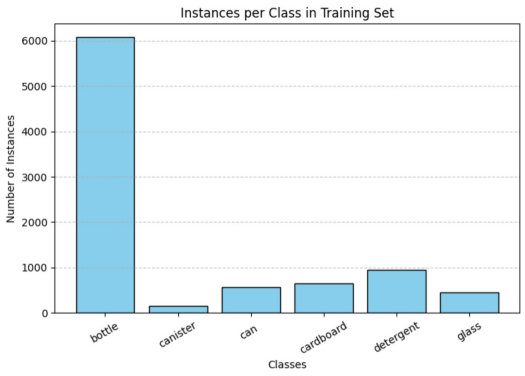

Figure 2. Class histogram

Initially, all subclasses were incorporated, after which the dataset was refined so as to focus on six primary waste categories, herein the WaRP-D dataset was used, which is intended for object detection.

3.2. Waste Sorting Using Deep Learning and Synthetic Data

For detecting diverse waste instances within the WaRP-D dataset, the YOLOv12 (You Only Look Once) model (Hussain, 2023) was employed as the main object detection framework. Recognized for its real-time performance and high accuracy, YOLO represents a state-of-the-art solution in visual recognition. Version 12 of this model (Tian et al., 2025), the most recent release in the YOLO

family, presently offers an enhanced efficiency and deeper feature extraction capabilities. One of its primary advantages is the Ultralytics Python library, which streamlines model setup, training, and evaluation, ensuring a seamless integration into the proposed experimental pipeline.

YOLO-v12 features an architecture of approximately 225 layers and more than 3 million parameters, forming a highly intricate deep network optimized for detecting and classifying objects with precision.

In order to address data scarcity and class imbalance within the WaRP-D subset, the original dataset was augmented with synthetically generated images simulating real industrial environments. Instead of relying only on standard augmentation or random compositing methods, a generative approach using ChatGPT's image creation capabilities was also adopted. Synthetic images were generated using text-based prompts that explicitly constrained the visual characteristics of the target objects. Each prompt specified either a uniform, the blank background, the waste category (e.g. cans), or variations in object color, size, and geometry, and different physical conditions, such as intact or crumpled shapes. Following generation, the images were manually inspected and curated to ensure category correctness, visual clarity, and consistency with real-world waste appearances.

ChatGPT (model GPT-4o, Islam & Moushi, 2025) was selected as the synthetic image generator due to its ability to produce high-resolution, semantically consistent visual assets guided by fine-grained textual prompts. Unlike traditional GAN-based pipelines, which require task-specific training and large curated datasets, the employed generative model allows the direct control over object appearance, deformation, material properties, and the camera perspective through prompt engineering. This flexibility is particularly relevant for waste sorting scenarios, where intra-class variability (e.g. crushed cardboard, partially filled containers, reflective cans) plays a critical role in model generalization. The use of such a prompt-driven generative approach enables a rapid iteration and targeted asset design without retraining a generative model for each category.

Figure 3 illustrates an example of synthetically generated canisters. Each object was individually cropped from the generated image and saved as a separate .png file, thereby expanding the number of available samples for that class. This procedure was repeated across all categories, effectively enriching the dataset and enhancing its diversity. The prompt used for generating this kind of images, with ChatGPT, is the following: "Generate an image with multiple used plastic canisters, oil and detergent containers, isolated individually, realistic wear and dirt, stains, scratches, faded labels, various shapes and sizes, random rotations and orientations, photographed as separate objects, neutral studio lighting, soft shadows, high realism, sharp focus, dataset-style object rendering, white background, no environment, no props, scale consistency, ultra-high resolution arranged as a grid of isolated objects, each object clearly separated, evenly spaced."

Figure 3. Canister assets generated with ChatGPT

Each synthetic asset was carefully designed to visually resemble real-world waste, ensuring a diverse range of appearances in terms of shape, colour, deformation, and the camera perspective. These generated waste items were then composited onto high-resolution images of clean conveyor belts, which had been previously captured without any objects present.

To improve the visual consistency between synthetic assets and real conveyor-belt scenes, a lightweight and controlled compositing pipeline was employed. Synthetic objects were relit using global photometric normalization by matching the mean and standard deviation of the background illumination in the HSV color space, followed by alpha-matting with edge feathering boundaries (Gaussian blur σ = 1–2 pixels). The soft contact

shadows have been omitted. Perspective consistency was maintained by scaling objects according to the vertical image position, reflecting the fixed camera geometry of the dataset, while partial occlusions were introduced through a controlled object overlap limited to 15–20% of the object area.

Background bias was mitigated by compositing assets only onto training-set conveyor backgrounds with randomized placement patterns (randomizing the rotation of the synthetic assets), ensuring a strict separation between the training and test data. To reduce the risk of overfitting synthetic assets, additional stochastic variations were applied during the compositing process. Each synthetic object was randomly rotated, rescaled within a ±10% range, and perturbed with low-amplitude Gaussian noise at the insertion time (reducing the change that the same asset would face in an identical visual configuration multiple times). These controlled variations introduce diversity in appearance while preserving class identity, encouraging the model to learn robust features rather than memorizing specific synthetic instances.

All the generated images were used exclusively for research purposes in accordance with the usage policies of the employed platform.

Figure 4 depicts certain cardboard packages which, as ChatGPT specifically required, were generated as crumbled and not clean so that these instances would closely resemble waste instances that are closer to reality.

Figure 4. Cardboard assets generated with ChatGPT

The placement of each synthetic object within the image was randomized while maintaining plausible spatial arrangements and avoiding excessive overlap to ensure annotation quality. The background bias was mitigated by using multiple conveyor-belt background images and varying the object placement patterns across the synthetic samples, ensuring that the object-background correlations were not fixed. After the assets were inserted, the corresponding label files containing the bounding box coordinates and class identifiers for each object were generated, adhering to the format expected by object detection models such as YOLO. This method enabled the expansion of the training dataset with controlled, annotated synthetic samples, which were integrated into the training process and evaluated through standard performance metrics for a real validation dataset.

For preventing data leakage between the training and evaluation, all the synthetic images were generated and composited exclusively using background images derived from the training part of the WaRP-D dataset. No background images, frames, or visual material originating from the test part were used during synthetic data creation.

The train–test partition for the original WaRP dataset was preserved throughout all experiments, and the test set remained unchanged across all the augmentation conditions. Synthetic images were added only to the training set, ensuring that no synthetic content, object placement pattern, or background texture appeared in the test data.

Since the WaRP-D dataset was captured using a fixed camera and conveyor configuration, the camera perspective and belt appearance are inherently consistent across the two parts; however, a strict separation of the original test images and all synthetic derivatives was maintained to avoid the overlap or indirect contamination.

4. Experiments and Results

To assess the impact of synthetic data on model performance, a series of experiments were conducted using the YOLOv12 model, trained under varying data conditions. The goal was to evaluate whether the inclusion of AI-generated synthetic waste imagery leads to measurable improvements in the object detection accuracy. All models were trained using a strong augmentation baseline provided by the YOLOv12 framework, including mosaic augmentation, RandAugment-based color perturbations, HSV jittering, random scaling, translation, and horizontal flipping.

Synthetic data augmentation was applied in addition to these conventional techniques.

The selected augmentation levels involving 2,500, 5,000, and 10,000 synthetic images were chosen for studying the effect of progressive synthetic data scaling relative to the original WaRP-D training set (2,452 images). These values approximately correspond to a 1×, 2×, and 4× increase relative to the original training set size, enabling a controlled analysis of the early performance gains from moderate augmentation, the refinement effects under a comparable real-to-synthetic data ratio, where synthetic samples do not dominate the training distribution and the potential onset of saturation or diminishing returns at higher synthetic data volumes. For all experiments, the test set (532 images) was kept fixed and unchanged. A validation set corresponding to 10% of the original training data was held out prior to augmentation and was not augmented with synthetic samples.

The synthetic images were added exclusively to the training set, resulting in the following effective training sizes: 2,320 images (the baseline), 4,820 images (the original training set size + 2,500 synthetic images), 7,320 images (the original training set size + 5,000 synthetic images), and 12,320 images (the original training set size + 10,000 synthetic images). The validation set remained constant across all the experiments and it was used for model selection and early stopping, ensuring a fair comparison between the augmentation conditions.

The generated synthetic images differ from the original WaRP-D samples in several controlled aspects. First, the synthetic assets were explicitly designed for increasing intra-class variability, introducing a wider range of deformations, orientations, and surface conditions than in the case of the original dataset. For example, cardboard instances were generated in crumpled and irregular forms, while the plastic containers varied in transparency, fill level, and shape distortion.

Second, unlike the cropped real samples, synthetic objects were composited onto conveyor-belt backgrounds under randomized spatial arrangements, resulting in diverse object scales and relative positions. While the background texture remains consistent with the real acquisition setup, the object appearances exhibit a higher diversity in shape and pose.

From a visual inspection perspective, the synthetic samples preserve the semantic identity of each waste category while intentionally exaggerating appearance variation. This explains the observed performance trend: early improvements reflect an improved generalization, while later saturation suggests that additional synthetic samples increasingly overlap with the already learned visual patterns rather than introducing novel discriminative cues.

The reported results are based on single training runs per experimental condition and are therefore presented as descriptive performance comparisons rather than inferential statistical claims. Given the deterministic nature of the dataset partition and the absence of repeated runs with different random seeds, formal hypothesis testing methods such as ANOVA were not applied. Instead, the experiments aim to identify the relative performance trends and saturation effects as a function of the synthetic data volume.

After training the YOLOv12 framework solely on the original dataset, the baseline results were obtained. The evaluation was performed using the following standard object detection metrics:

- precision: the proportion of true positive detections among all the positive predictions made by the model;
- recall: the proportion of the actual positive instances correctly identified by the model;
- mAP@50: reflects the average detection precision across all classes when the predicted and ground truth bounding boxes overlap by at least 50%;
- mAP@50:95: a more comprehensive metric that averages the precision over multiple IoU thresholds, ranging from 0.5 to 0.95, in steps of 0.05, providing a robust assessment of the detection performance.

The initial experiment aimed to establish a performance baseline by training the YOLOv12 model exclusively on the original WaRP-D dataset, without any synthetic augmentation. This dataset included high-resolution RGB images of waste on a conveyor belt, annotated with bounding boxes and class labels across six major waste categories.

All models were trained for 100 epochs using the Ultralytics YOLOv12 training pipeline with the

AdamW optimizer (learning rate = 0.001, weight decay = 5×10^{-4}, momentum = 0.9), selected automatically by the employed framework. A batch size of 64 was used, with a cosine-free learning rate schedule and a warmup phase of 3 epochs (warmup bias LR = 0.1, warmup momentum = 0.8). The input images were resized to 640×640 pixels. Data augmentation included mosaic augmentation (enabled, closed after epoch 10), horizontal flipping (p = 0.5), HSV color jittering (h = 0.015, s = 0.7, v = 0.4), random translation (±10%), and random scaling (±50%), while rotation, shear, and mixup were disabled. The training was deterministic with a fixed random seed (the seed was equal to zero).

All experiments were conducted on a workstation equipped with a NVIDIA RTX 4090 GPU (16 GB VRAM), 64 GB RAM and CUDA-enabled PyTorch, using 8 dataloader workers. The training time was approximately 30 hours per configuration (≈18 minutes per epoch), with identical settings used across all the augmentation scenarios in order to ensure a fair comparison.

Following training, the model was evaluated using the test set, and its performance was measured using the metrics described previously: Precision, Recall, mAP@50, and mAP@50:95. The results, presented in Table 1, reflect the model's baseline capability to detect waste objects in real-world conditions using the original annotated data and the synthetic data. The experiments started using only the original data in order to create a baseline that serves as a reference point for the subsequent experiments that incorporate synthetic data.

Tables 1 and 2 present the performance of the YOLOv12 model trained under different circumstances (original data and synthetic data). Table 1 is related to the detection performance of the trained YOLOv12 model while Table 2 is related to the classification performance of the trained model (for each training scenario). For the baseline using only the original data, the aggregate performance across all classes yields a precision of 0.665, recall of 0.508, a mAP@50 of 0.593, and a mAP@50:95 of 0.466. These results indicate that while the model demonstrates a solid capability in detecting waste items, there is still room for improvement, especially when it comes to a more precise localization under stricter IoU thresholds.

The class-wise analysis shows a clear variability across the waste categories. The canister class contains only four test instances; therefore, its near-ceiling AP values are not statistically meaningful and are reported for completeness purposes only.

Table 1. YOLOv12 framework waste detection performance across datasets (per class and overall)

Class	No. of instances	Original mAP		+2.5k synthetic images mAP		+5k synthetic images mAP		+10k synthetic images mAP	
		50	50:95	50	50:95	50	50:95	50	50:95
bottle	122	0.812	0.614	0.811	0.622	0.822	0.620	0.824	0.624
canister	4	0.995	0.952	0.995	0.920	0.995	0.921	0.995	0.995
can	28	0.459	0.352	0.499	0.381	0.442	0.345	0.529	0.387
cardboard	35	0.455	0.354	0.427	0.300	0.422	0.333	0.443	0.349
detergent	46	0.376	0.303	0.392	0.320	0.423	0.329	0.453	0.373
glass	21	0.620	0.222	0.520	0.272	0.493	0.278	0.491	0.296
macro	-	0.593	0.466	0.607	0.469	0.599	0.471	0.622	0.504

Table 2. YOLOv12 framework waste classification performance across datasets (per class and overall)

Class	No. of instances	Original		+2.5k synthetic images		+5k synthetic images		+10k synthetic images	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
bottle	122	0.763	0.724	0.755	0.753	0.841	0.690	0.824	0.624
canister	4	0.703	1.000	0.947	1.000	1.000	0.977	0.995	0.995
can	28	0.828	0.311	0.583	0.419	0.642	0.355	0.529	0.387
cardboard	35	0.542	0.295	0.475	0.318	0.546	0.409	0.443	0.349
detergent	46	0.553	0.353	0.521	0.333	0.617	0.333	0.453	0.373
glass	21	0.604	0.366	0.819	0.362	0.755	0.400	0.491	0.296
macro	-	0.665	0.508	0.683	0.531	0.734	0.527	0.597	0.573

Figure 5. Comparison between labels (left) and predictions (right) for the baseline model

The performance trends and conclusions are drawn primarily from classes including a sufficiently high number of samples. The bottle class performed consistently well in comparison with the other classes, reaching 0.812 for mAP@50 and 0.614 for mAP@50:95, benefitting from its larger representation in the dataset.

In contrast, the performance was weaker for cans, cardboard, and detergent, where both precision and recall dropped significantly. For example, cardboard scored 0.455 for mAP@50 and 0.354 for mAP@50:95, while detergent achieved values of 0.376 and 0.303, respectively. These results suggest difficulties in learning robust representations for classes that are more visually ambiguous or that appear with a high variability on the conveyor. Glass achieved moderate detection results (0.620 for mAP@50), but its fine-grained localization remained low (0.222 for mAP@50:95).

These baseline results establish a benchmark for assessing the impact of synthetic data integration in the following experiments. As the model performance is concerned, the image on the left side of Figure 5 displays the ground truth bounding box with only one object labelled as "detergent." The annotation is clear and well-placed. With regard to the right side of Figure 5, the YOLOv12 model detects the same "detergent" object with a high confidence (0.92), showing a precise match. In addition, the model predicts other objects like "bottle" and "canister" with a moderate confidence. These objects are not present in the ground truth bounding box therefore they are considered false positives.

Overall, the model performs well on the known classes and begins to detect other items, though further evaluation is needed for unlabeled objects.

The precision-recall (PR) curves for the baseline model (Figure 6) trained on the original WaRP-D dataset show strong precision–recall trade-offs for the well-represented and visually consistent classes such as *bottle* and *glass*, while the performance degrades rapidly for the underrepresented or visually ambiguous categories such as *can* and *cardboard*. The aggregated curve (a mAP@0.5 of 0.485) indicates a steady decline in precision as the recall increases, denoting sensitivity to false positives at lower confidence thresholds. These curves highlight the impact of class imbalance and appearance variability on the baseline setting, motivating the use of synthetic augmentation to improve robustness for the weaker classes.

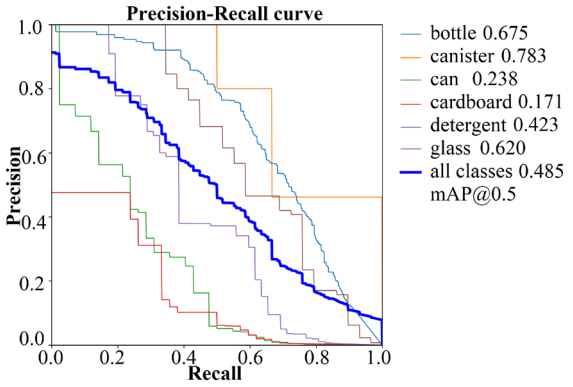

Figure 6. PR curves of the original model

To assess the impact of synthetic data augmentation on model performance, the second experiment involved enriching the original WaRP-D dataset with 2,500 synthetic images (one time the size of the original training set).

As reflected by the performance tables, the inclusion of 2,500 synthetic images led to a slight improvement in the overall model performance in comparison with the baseline. The global mAP@50 increased from 0.593 to 0.607, while mAP@50:95 also showed a marginal increase

Figure 7. Comparison between labels (left) and predictions (right) for the model trained with 2,500 more synthetic images

from 0.466 to 0.469. These results suggest that synthetic data helped enhance the model's generalization, particularly in classifying common items such as *bottle* and *glass*. The overall classification also improved from a Precision/Recall ratio of 0.665/0.508 to one of 0.683/0.531. From a class-wise perspective, the bottle class remained strong (the mAP@50 maintained the value 0.81 and the value of mAP@50:95 increased from 0.614 to 0.622), the can class improved modestly (the value of mAP@50 increased from 0.459 to 0.499), and the detergent class ticked up (the value of mAP@50 increased from 0.376 to 0.392). The glass class dipped somewhat at stricter thresholds but it maintained a reasonable performance. These results suggest that the synthetic set added useful intra-class variation without destabilizing the detector.

In comparison with the baseline, there are fewer false positives, hinting that synthetic data may reduce overfitting while preserving the key detections; there also appears to be an annotation error, as the bottle detection seems plausible.

In summary, this experiment demonstrates that augmenting training data with synthetic images, when properly generated and integrated, can lead to tangible gains in the object detection performance, especially for frequent or visually consistent classes. The left side of Figure 7 depicts the ground truth bounding box with a single "detergent" label, clearly marked in green. With regard to the right side of Figure 7, the YOLOv12 model trained with 2,500 more synthetic images again detects the *detergent* accurately, with a high confidence (0.92). Additionally, it detects a *bottle* with a low confidence (0.33). This indicates that the model starts to recognize new objects, but with uncertainty. In comparison with the baseline model, there are fewer false detections, suggesting that the synthetic data may help reduce overfitting while keeping the key detections stable, and there is also a clear mistake made with regard to the annotations, as the bottle detection performed by the model seems to be correct.

In comparison with the baseline, the PR curves (Figure 8) show a modest but consistent upward shift for the well-represented classes such as *bottle* and *canister*, reflected in the increased overall mAP@0.5 (0.496 vs. 0.485). The minority and underrepresented classes (*can*, *cardboard*) remain limited by a low recall, indicating that a moderate synthetic augmentation improves robustness primarily for the dominant classes. The aggregated curve suggests a slightly more favorable precision–recall trade-off without altering the overall operating regime (supporting the idea that this augmentation level enables incremental gains).

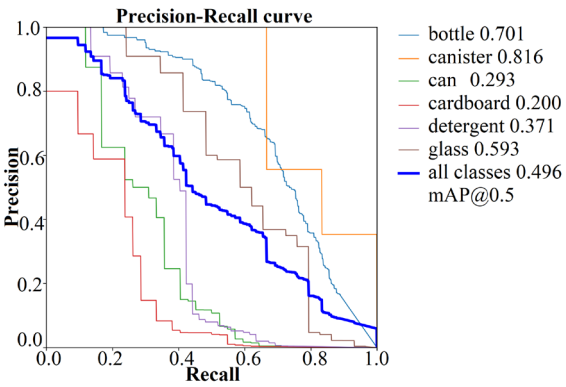

Figure 8. PR curves for the model trained with an augmented training set (+2.5k synthetic images)

The third experiment involved further expanding the training dataset by adding 5,000 AI-generated synthetic images to the original WaRP-D dataset while keeping all the other hyperparameters unchanged.

Figure 9. Comparison between labels (left) and predictions (right) for the model trained with 5000 more synthetic images

With 5,000 more synthetic images, the overall detection performance revealed a mAP@50 of 0.599 and a mAP@50:95 of 0.471. Relative to the 2,500-image setup, this reflects a slight drop for mAP@50 but a small gain for mAP@50:95, indicating a marginally better fine-grained localization. The overall classification increased in precision to 0.734 (from 0.683) with a small decrease in recall to 0.527 (from 0.531), suggesting a more conservative but cleaner classifier. The class-wise trends remained consistent: the bottle class kept its position (0.822/0.620 for mAP@50/mAP@50:95), the detergent class improved again (0.423 for mAP@50), and the glass class stabilized, while the can and cardboard classes remained comparatively underrepresented.

In short, the model trained with 5,000 more synthetic images provided the highest classification precision so far and the best mAP@50:95 value to this point, although the gains were not very high. Figure 9 illustrates a qualitative behaviour similar to that depicted in Figure 7: the detergent class is also detected accurately, with a high confidence" (~0.91) and several "bottle" objects are detected with a slightly higher confidence than in the case of the model trained with the augmented training set of 2,500 images.

For the model augmented with 5,000 synthetic images, the aggregated PR curve (Figure 10) exhibits a clearer upward shift, yielding an overall mAP@0.5 of 0.522 and indicating an improved precision at moderate recall levels in comparison with the baseline model and the model augmented with 2,500 synthetic images. The gains are the highest for the bottle and canister classes, while the can and cardboard classes show a limited recall improvement, suggesting persistent class-specific challenges. The smoother decay of the aggregated curve reflects a more stable confidence ranking, supporting the interpretation that this augmentation level provides the best balance between performance improvement and diminishing returns.

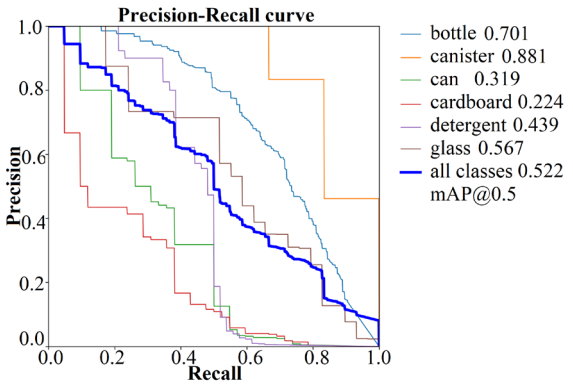

Figure 10. PR curves for the model trained with the augmented training set (+5k synthetic images)

However, the improvements appear to taper off for some classes, suggesting that merely increasing synthetic data volume may not be sufficient without a further refinement in asset diversity and realism, therefore another experiment was carried out in which the training dataset was further expanded by adding 10000 more synthetic images.

With 10,000 more synthetic images, the overall detection performance reached its best values: a mAP@50 of 0.622 and a mAP@50:95 of 0.504. However, the overall classification featured a trade-off: the precision decreased to 0.597 while the recall increased to 0.573, and the classifier captured more positives at the cost of extra false positives. From a class-wise perspective, the *can* class achieved its highest mAP@50 (0.529), the *detergent* class improved further (0.453/0.373 for mAP@50/mAP@50:95), the *bottle* class edged up again (0.824/0.624), and the *canister* class remained saturated near the ceiling (0.995/0.995),

Figure 11. Comparison between labels (left) and predictions (right) for the model trained with 10,000 more synthetic images

while the glass class achieved a lower mAP@50 value in comparison with the baseline model but it achieved a better value for mAP@50:95 in comparison with the intermediate models with synthetic augmentation.

Overall, moving from 5,000 to 10,000 synthetic images yielded the strongest detection metrics but it introduced a precision–recall trade-off in the classification. Figure 11 shows the same comparison for the model expanded with 10,000 synthetic images: the detergent class is also detected with a high confidence (~0.91), and a bottle object is detected with an improved confidence (~0.43) and a tighter bounding box, but the aggregate quantitative gains in comparison with the model expanded with 5,000 synthetic images are still modest, consistent with the diminishing returns. Predictions were also performed by this model in order to visualize its performance, a similar image to that in Figure 9 was used, its left side showing the ground truth bonding box with only the *detergent* class labeled. The right side of Figure 11 corresponds to the model augmented with 10,000 synthetic images, which correctly detects the detergent with a high confidence (0.91). It also identifies a bottle object with an improved confidence (0.43), and the bounding box is more accurate than in the previous model versions. In comparison with the model augmented with 5,000 synthetic images, its performance is slightly better, but the gains are small. This suggests that adding more synthetic data beyond this point may lead to diminishing returns.

For the model augmented with 10,000 synthetic images, the aggregated PR curve (Figure 12) achieves a mAP@0.5 of 0.521. However, the curve also shows a steeper drop in precision at higher recall levels, reflecting an increased false-positive rate and a less conservative classifier. While classes such as *canister* continue to benefit, the gains for *can* and *cardboard* remain limited, reinforcing the observation related to diminishing returns and highlighting residual domain gap effects rather than an insufficient data volume.

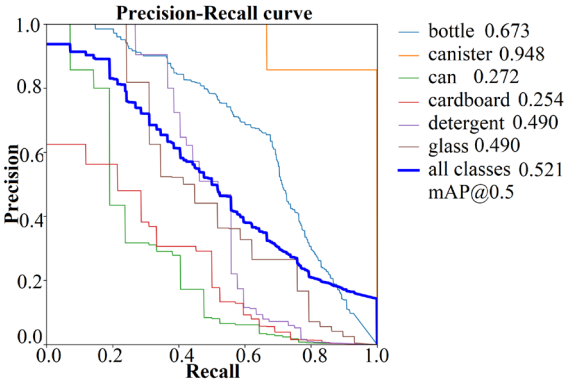

Figure 12. PR curves of the model trained with augmented training set (+10k synthetic images)

In conclusion, this final experiment indicates that beyond a certain point, that is approximately 5,000 more high-quality synthetic images (or almost two times more synthetic images than the original number), the utility of further augmentation diminishes in this case. While synthetic data proves essential for enhancing a model`s performance, especially for underrepresented classes, simply scaling the dataset size does not guarantee continued gains. Future improvements may require enhancing the realism and diversity of the synthetic assets or exploring model-specific optimizations.

5. Discussion

The obtained results reinforce a common finding in waste detection and classification research:

model performance is often constrained more by data availability, class imbalance, and domain variability than by the model architecture. Prior work consistently identified background complexity, occlusion, and illumination changes as key challenges for generalization in realistic industrial environments. In the WaRP conveyor-belt setting, these factors are exacerbated by object overlap, deformation, and camera perspective variability, making robust localization and classification difficult even for modern detectors.

Within this context, the observed improvements based on synthetic augmentation are practically meaningful. The detection performance increases from the baseline (a mAP@50 of 0.593, a mAP@50:95 of 0.466) to the best augmented model (a mAP@50 of 0.622, a mAP@50:95 of 0.504), indicating that the proposed synthetic data pipeline introduces a beneficial intra-class variation. The higher gains under stricter localization criteria (mAP@50:95) suggest that synthetic data contributes not only to the detection frequency but also to an improved bounding-box tightness and localization robustness.

All experiments were conducted using a fixed random seed due to the high computational cost for training large detection models. While this limits formal variance estimation, all configurations were trained and evaluated under identical conditions, enabling a direct comparison of relative trends. The performance saturation observed beyond an augmentation with approximately 5,000 synthetic images is therefore interpreted as an empirical trend rather than a statistically defined threshold. Future work could include multi-seed evaluation and calibration analyses in order to better quantify variability and prediction performance.

The qualitative analysis reveals consistent error patterns across the analysed models. False positives frequently arise from reflective background structures, partial object regions, and confusion between visually similar categories such as bottles and canisters. False negatives are primarily associated with heavy occlusion, strong specular highlights, and severe object deformation, particularly for cardboard. These observations suggest that realism and appearance diversity, rather than dataset size alone, limit further gains from synthetic augmentation.

Residual performance limitations can be attributed to several interacting factors. Despite controlled blending and relighting, the compositing pipeline remains a simplified approximation of real conveyor-belt scenarios and does not fully capture complex lighting or material-dependent effects. Class imbalance also persists, particularly for the can and cardboard classes, whose variability and deformation are difficult to reproduce synthetically.

The class-wise analysis aligns with the known operational challenges. The bottle class consistently achieves good results due to its stable visual features and to being well-represented, while the detergent class benefits from increased appearance coverage through augmentation. In contrast, the can and cardboard classes remain difficult, indicating that augmentation alone is insufficient without a highly realistic deformation, reflectance, and occlusion modeling. The near-ceiling performance of the canister class should be interpreted cautiously, as a very small number of test samples limits the generalization ability.

From a practical perspective, synthetic augmentation offers a low-cost mechanism for improving detector robustness without extensive manual annotation, which is particularly valuable in industrial settings. However, persistent false positives highlight the need for validation in additional real conveyor-belt scenarios and to avoid the confusion between visually similar materials before model deployment.

Based on these findings, several actionable recommendations emerge. First, the results indicate that prioritizing realism over sheer synthetic data volume becomes essential once a moderate synthetic dataset augmentation, approximately 1× to 2× the size of the original training set, is reached, as performance gains were observed to reach saturation beyond this point. In particular, increasing the synthetic dataset size to 4× the original size (~10,000 synthetic images) yielded only marginal improvements for the evaluated dataset. Second, an ablation-driven refinement for the synthetic factors is recommended, as the results show that only a subset of synthetic attributes contributes meaningfully to performance gains. Third, synthetic data is most effective when used for addressing class imbalance and real failure

cases, improving recall for underrepresented categories through targeted sample generation.

6. Conclusions

This study investigated the use of AI-generated synthetic images to augment training data for the automated waste detection and classification in recycling facilities. Using the WaRP dataset and a YOLOv12-based pipeline, four configurations were trained and evaluated: the original dataset and three augmented variants expanded with 2,500, 5,000, and 10,000 synthetic images (which correspond to 1x, 2x or 4x more images in comparison with the size of the original training set) generated via ChatGPT and composited onto clean conveyor-belt backgrounds.

Across the experiments, the synthetic augmentation improved the detection performance in comparison with the baseline model. The best detection results were obtained for the model trained with 10,000 synthetic images, achieving a mAP@50 of 0.622 and a mAP@50:95 of 0.504, in comparison with the baseline with a mAP@50 of 0.593 and a mAP@50:95 of 0.466. As the classification performance is concerned, the model trained with 5,000 synthetic images outperformed the baseline model, improving the precision value from 0.665 to 0.734 while maintaining a higher recall.

Future work could focus on improving synthetic data realism and diversity (e.g. more varied lighting, textures, shadows and object interactions), expanding background variability beyond clean conveyor-belt backgrounds, and validating the proposed approach on additional real-world scenarios and plant configurations in order to better assess its generalization ability across different operating conditions.

Acknowledgements

This work was supported by a grant of the Ministry of Research, Innovation and Digitalization, CNCS/CCCDI-UEFISCDI (National Council for Scientific Research/Romanian National Authority for Scientific Research and Innovation - Executive Agency for Higher Education, Research, Development and Innovation Funding), project code COFUND-EP-PERMED-PERSONALISE-DKD (“Personalised SGLT2i treatment in diabetic kidney disease supported by multiparametric renal magnetic resonance imaging”), within PNCDI IV (National Plan for Research, Development, and Innovation2022-2027).

REFERENCES

Abdu, H. & Noor, M.H.M. (2022) A survey on waste detection and classification using deep learning. *IEEE Access*. 10, 128151-128165, https://doi.org/10.1109/ACCESS.2022.3226682.

Achicanoy, H., Chaves, D. & Trujillo, M. (2021) StyleGANs and transfer learning for generating synthetic images in industrial applications. *Symmetry.* 13(8), Art. ID 1497, https://doi.org/10.3390/sym13081497.

Anon (n.d.) *WaRP - Waste Recycling Plant Dataset.* https://www.kaggle.com/datasets/parohod/warp-waste-recycling-plant-dataset [Accessed 8th September 2025].

Bacchin, A., Marangoni, F., Gottardi, A. et al. (2025) Image Data Augmentation through Generative Adversarial Networks for Waste Sorting. In: *2025 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, *15-18 April 2025, Palermo, Italy*. Barcelona, Spain, PAL Robotics. pp. 1-6. https://doi.org/10.1109/SIMPAR62925.2025.10979009.

Bojer, C. S. & Meldgaard, J. P. (2021) Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting.* 37(2), 587–603, https://doi.org/10.1016/j.ijforecast.2020.07.007.

Dong, S., Wang, P. & Abbas, K. (2021) A survey on deep learning and its applications. *Computer Science Review.* 40, Art. ID 100379. https://doi.org/10.1016/j.cosrev.2021.100379.

Gautam, B. & Arashpour, M. (2025) Advanced data augmentation techniques to enhance instance segmentation dataset for construction and demolition waste management. *Waste Management*. 200, Art. ID 114744, https://doi.org/10.1016/j.wasman.2025.114744.

Hussain, M. (2023) YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines.* 11(7), Art. ID 677. https://doi.org/10.3390/machines11070677.

Islam, R. & Moushi, O. M. (2025) GPT-4o: The cutting-edge advancement in multimodal LLM. In:Arai, K. (ed.) *Lecture Notes in Networks and Systems.* vol. 1426 (*Intelligent Computing -Proceedings of the 2025 Computing Conference*, vol. 4). Cham, Switzerland, Springer Nature, pp. 47-60.

Melinte, D. O., Travediu, A. M. & Dumitriu, D. N. (2020) Deep convolutional neural networks object detector for real-time waste identification. *Applied Sciences.* 10(20), Art. ID 7301. https://doi.org/10.3390/app10207301.

Nasien, D., Adiya, M. H., Farkhan, M. et al. (2025) Automated Waste Classification Using YOLOv11 A Deep Learning Approach for Sustainable Recycling. *Journal of Applied Business and Technology*. 6(1), 68-74, https://doi.org/10.35145/jabt.v6i1.205

Ogrezeanu, I.A., Suciu, C. & Itu, L.M. (2025) Automated Waste Sorting Using Deep Learning and Synthetic Data. In: Chiru, A. & Covaciu, D. (eds.) *Proceedings in Automotive Engineering* (*CONAT 2024 International Congress of Automotive and Transport Engineering, Part Two: Automobile and Environment*), Cham, Switzerland, Springer Nature, pp. 213–223.

Singh, S. K., Kumar, S. & Mehra, P. S. (2023) Chat GPT & Google Bard AI: A Review. In:*2023 International Conference on IoT, Communication and Automation Technology (ICICAT)*, *23-24 June 2023, Gorakhpur, India*. New York, USA, IEEE. 10.1109/icicat57735.2023.10263706.

Tian, Y., Ye, Q. & Doermann, D. (2025) *Yolov12: Attention-centric real-time object detectors*. [Preprint] https://arxiv.org/abs/2502.12524 [Accessed: 3rd October 2025].

Yudin, D., Zakharenko, N., Smetanin, A. et al. (2024) Hierarchical waste detection with weakly supervised segmentation in images from recycling plants. *Engineering Applications of Artificial Intelligence.* 128, Art. ID 107542, https://doi.org/10.1016/j.engappai.2023.107542.

Zandalinas, S. I., Fritschi, F. B. & Mittler, R. (2021) Global warming, climate change, and environmental pollution: recipe for a multifactorial stress combination disaster. *Trends in Plant Science.* 26(6), 588–599, https://doi.org/10.1016/j.tplants.2021.02.011.

Zhao, Z.-Q., Zheng, P., Xu, S.-T. et al. (2019) Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems.* 30(11), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865.