RTDRNet-lite: A Lightweight Real-Time Detection Framework for Robotic Waste Sorting

Md Jawadul Karim [a], Sirajum Munir [b], Amith Khandakar [c,1] , Mominul Ahsan [d], Julfikar Haider [e]

[a] *Department of Computer Science and Engineering, BRAC University. Dhaka-1212, Bangladesh. Email: md.jawadul.karim@g.bracu.ac.bd*

[b] *Department of Electrical and Computer Engineering, Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh. Email: 2010013@student.ruet.ac.bd*

[c] *Department of Electrical Engineering, College of Engineering, Qatar University, Doha, 2713, Doha, Qatar. Email: amitk@qu.edu.qa*

[d] *Department of Computer Science, University of York, Deramore Lane, Heslington, York YO10 5GH, UK. Email: mominul.ahsan@york.ac.uk*

[e] *Department of Engineering, Manchester Metropolitan University, Chester Street, Manchester M1 5GD, UK. Email: j.haider@mmu.ac.uk*

[1] *Corresponding Author: Department of Electrical Engineering, College of Engineering, Qatar University, Doha, 2713, Doha, Qatar. Email: amitk@qu.edu.qa*

Abstract

In the age of global urbanization, waste recycling remains a critical challenge, impacting the environment and societies from small communities to entire nations. This research aims to address these gaps by proposing a comprehensive and fully automated waste management framework that integrates advanced AI-based detection with robotic hardware to enable intelligent, real-time waste sorting. The fundamental framework of this work is the RTDRNet-lite model, a modified lightweight version of the high-performing object detection variant RT-DETR, which achieved an impressive mAP@50 of 97%. Developed with real-time applicability in mind, the model uses lightweight C2F modules within its head architecture, reducing the computational complexity without any dramatic change in accuracy. A unique approach to training the model was employed, leveraging both real-world waste image data and highly detailed synthetic images generated using the Stable Diffusion model, the Realistic Vision v5.1. This hybrid approach enriches visual

diversity and improves the model's generalizability, especially in handling complex object boundaries. The model is trained on four high-frequency waste categories, paper, plastic, glass, and metal, using over 12,929 annotated instances. Additional qualitative evaluations, including IoU-based visual analysis, external validation, and heatmap visualization, confirm the model's robustness, spatial accuracy, and resilience in complex scenes. To demonstrate real-world applicability, a custom 4-degree-of-freedom (DoF) robotic arm was developed and integrated with the model, successfully validating its performance in live sorting tasks. The results confirm both the numerical performance and the practical deployment potential of the proposed system for large industrial-scale waste management facilities and environments.

Keywords: Waste detection, stable diffusion model C2F block, robotic arm 4 DoF, inverse kinematics, graphical user interface, GUI

1. Introduction

As cities expand and consumption increases, the global waste crisis is becoming one of the most pressing yet overlooked challenges of our time. From overflowing landfills to polluted rivers and oceans, the sheer volume and diversity of waste produced daily pose a critical threat to our ecosystems, public health, and future sustainability (Jain & Shah, 2019). Waste management, once seen as a basic municipal service, has now escalated into a global priority, deeply intertwined with environmental preservation, public health, and economic efficiency. With urban centers producing millions of tons of solid waste each day, the urgency to implement intelligent, scalable solutions has never been greater.

The effects of poor waste handling can be seen everywhere on our planet. Plastics choke marine ecosystems and release toxins into waterways, which further degrade into microplastics that are ingested by animals and enter food chains with potentially toxic impacts on wildlife and humans (Emenike et al., 2023). Open dumping and incineration are commonly practiced in third-world countries, and these methods are known to release toxic fumes and greenhouse gases that fuel climate change and respiratory diseases (Sheriff et al. 2025). E-waste, in addition to hospital and industrial hazardous waste, is frequently released from these facilities, is not properly contained, and may result in long-term soil and groundwater pollution (Hasan et al., 2023). These impacts are not isolated, as they ripple through whole ecosystems and populations, often most intensely among the vulnerable and marginalized regions of the world. In addition to the environmental and health consequences, there is also an increasing economic cost. Municipalities around the world spend billions annually on waste collection and disposal, yet recycling rates remain disappointingly low (World Bank, n.d.). Sorting waste correctly, especially in an urban context with mixed waste, is a labor-intensive task and is subject to frequent errors (Sayem et al., 2024).

In traditional systems, human workers are tasked with the dirty, dangerous, and monotonous job of manually separating waste materials. This endangers workers and slows the scale and pace of their activities (Jerie, 2016). Furthermore, human error in classification often results in cross-contamination of recyclables, reducing the effectiveness of recycling facilities and increasing landfill dependency. These facts reveal the challenges facing traditional waste practices. Manual sorting, mechanical shredding, and basic optical/visual separation can be acceptable at the basic level; however, they are not robust enough to cope with the variety and unpredictability of modern waste (Fang et al., 2023). Conventional systems are not adaptable for identifying new waste materials and their ever-changing orientations or maintaining consistent performance across shifts and facilities (Alsabt et al., 2024).

As waste streams become more varied and contaminated, the cracks in these conventional systems become even more apparent. This is where the convergence of artificial intelligence and robotics began to reshape the narrative. In recent years, automation has gained traction in several industries, and its application in waste management is particularly promising because of the nature of the problem, repetitive tasks, hazardous environments, and the need for real-time decision-making (Jaouhari et al., 2024). AI, particularly computer vision and deep learning, offers a powerful toolset for recognizing patterns in waste items (Zhang et al., 2021), whether it distinguishes PET bottles from PVCs or identifies organic matter from synthetic packaging, tasks that often baffle even trained human workers (Torres et al., 2021). By training AI models on large datasets of labeled waste images, machines can learn to detect and classify waste materials with high accuracy. These detection systems can be mounted on conveyor belts in sorting facilities or integrated into smart bins in households and urban infrastructure.

However, a critical gap remains in much of the existing research in this domain. Many previous studies focused solely on developing and evaluating AI models for waste detection, classification, or segmentation, reporting results on the basis of accuracy, precision, or IoU metrics. While these contributions are valuable from a machine learning standpoint, they often fail in addressing the practical deployment of such systems. The discussion frequently ends at the model training phase, leaving the post detection phase, robotic manipulation, sorting strategies, and operational logistics largely unaddressed. This disconnect between algorithm development and system-level implementation limits the translational value of otherwise promising research. To physically act on this classification, robotic arms and automated manipulators are required to pick, sort, and place waste items into appropriate categories. These robots need to be designed to mimic human dexterity but operate with greater speed, consistency, and immunity to fatigue. The fusion of AI detection with robotic actuation represents a turning point for waste management systems (Lubongo et al., 2024). No longer confined to static roles, modern systems can now learn from data, adapt to new

waste patterns, and perform precise physical actions autonomously. A number of companies have already proven such concepts, such as AMP Robotics and ZenRobotics, which have deployed AI-driven sorting rigs capable of separating hundreds of tons of waste every hour more accurately and quickly than a human operator ever could. Such systems not only lower the dependence on human labor but also increase the recovery of valuable materials, such as aluminum, copper, and recyclable plastics, which directly contributes to the circular economy (Lakhouit et al., 2025).

Both of our previous studies on waste recycling automation by Sayem et al. (2024) and Nahiduzzaman et al. (2025), had several critical limitations that motivated the present work. In the study of Sayem et al. 2024), the use of image classification was impractical for real-time, multi-object sorting scenarios. Additionally, the dataset used suffered from severe class imbalance, with many categories containing very few samples, limiting generalization. In Nahiduzzaman et al. (2025) study, although a larger dataset was employed, it was primarily composed of web-scraped images, resulting in mislabeling, noisy backgrounds, and the presence of irrelevant objects. None of the study integrated sophisticated hardware implementation in a real-time physical setting and the robotic interaction was limited to mostly simulation. Building upon these insights and limitations identified in our previous two studies on waste recycling, this work presents a significantly more refined and efficient framework. Multiple scientific contributions are outlined in this paper:

1. **Development of a Lightweight Waste Detection Framework:** A streamlined detection model (RTDRNet-lite) was designed by simplifying an existing architecture to achieve efficient performance with significantly reduced computational requirements, making it suitable for real-time applications.
2. **Enhanced Dataset through Synthetic Image Generation**: A hybrid dataset was created by combining natural waste images with synthetically generated ones, addressing issues of class imbalance and limited data availability for certain categories.

3. **Integration with Physical Robotic Hardware**: The detection model was deployed on a custom-built 4-degree-of-freedom robotic arm, enabling real-time waste item identification and positioning for sorting operations.
4. **Evaluation on External Image Sets**: The proposed framework was tested on independent waste image datasets not used during training, demonstrating reliable performance across varying backgrounds and object types.
5. **Interpretability and User Accessibility**: A visual explanation mechanism was included to highlight the system's focus areas during detection, along with a graphical user interface (GUI) to support real-time monitoring and manual control.
6. **Improved Annotation and Data Processing Pipeline**: A semi-automated labeling approach was implemented for the synthetic data, improving annotation efficiency while maintaining quality through confidence-based filtering.
7. **Hardware-Oriented Optimization**: The overall system was designed with real-world constraints in mind, balancing model accuracy with reduced power consumption, memory usage, and hardware compatibility.
8. **Comparison with Prior Studies** This work addresses the limitations of earlier systems by enabling simultaneous multi-object detection and physical testing, moving beyond single-object classification and simulation-only environments.

The rest of the paper is structured as follows. In Section 2, we review existing works regarding deep learning-based waste detection, classification, and segmentation in a whole-spectrum manner, including model classification and deployment. In Section 3, we describe the dataset, materials, model architecture, and experimental procedures used in the present work. The proposed model is evaluated in Section 4, which discusses the performed metrics, validation methods, and real-time testing inside hardware integrated within a software environment. A comparison with the state-of-

the-art methods is given in Section 5, and concluding remarks as well as future works are presented in Section 6.

2. Literature Review

The rise of research on AI-based waste management systems in recent years has led to the development of a variety of innovative methodologies for waste classification, detection, and technology. While sharing common performance goals such as accuracy, model speed, or scalability, these goals have exploited various architectural approaches or datasets. A consistent thread across several works is the use of convolutional neural networks (CNNs) and deep learning models, with variations tailored to specific deployment contexts and waste types.

For example, papers by Majchrowska et al. (2021), Prasad et al. (2025), and Sayem et al. (2024) introduce dual-stage or dual-stream models for waste detection and classification: the first uses EfficientDet-D2 for localization and EfficientNet-B2 for classification and operates on seven categories of waste. It provides approximately 70% precision and 75% classification accuracy and real-time performance at 30 fps. Similarly, Sayem et al. (2024) introduced a dual-stream model coupled with the GELAN-E detection network on a comprehensive dataset of 10,406 images across 28 categories, achieving 83.11% classification accuracy and 63% mAP50 in detection. Both demonstrate how splitting detection and classification processes into specialized modules enhances performance, especially when backed by diverse datasets. The methods presented by Nahiduzzaman et al. (2025), Hossen et al. (2024), and Ahmed et al. (2023) focus on classification efficiency and model compactness. The first introduces a three-stage waste classification pipeline that efficiently categorizes waste into 2, 9, and 36 categories and delivers 96%, 91%, and 85.25% classification accuracy, respectively.

It employs a lightweight DP-CNN architecture (~1.09 M parameters) and an ensemble extreme learning machine (En-ELM), emphasizing real-time applicability with extremely low inference times. Hossen et al. (2024), followed the RWCNet model trained on TrashNet, achieving 95.01% overall accuracy with individual F1 scores exceeding 93% in five out of six categories. On the other hand, Ahmed et al. (2023) leverages transfer learning using DenseNet169, MobileNetV2, and ResNet50V2, where ResNet50V2 achieved a classification accuracy of 98.95%. The consistent use of pre-trained models in this work highlights the efficacy of transfer learning in boosting performance over custom CNNs. The segmentation of waste, particularly in cluttered and complex scenes, is another significant avenue explored in studies such as Sirimewan et al. (2024), Prasad et al. (2025), Qiu et al. (2022)), and Kiyokawa et al. (2021). In Sirimewan et al. (2023), segmentation of construction and demolition (CRD) waste via DeepLabv3+ and U-Net with backbones such as ResNet-101 yielded IoU values of 0.74 and mAP values of up to 0.85. Despite the use of a small dataset of 430 images, the work achieved reasonable performance, although limitations in class balance and manual labeling were noted. In contrast, Prasad et al. (2025) introduced ShARP-WasteSeg, which incorporates RGB and depth data to enhance boundary detection and instance segmentation. The integration of shape-aware and boundary-sensitive features improved the mask AP by 7.91% and the boundary AP by 11.44%. Qiu et al. (2022) took this further with ETHSeg for X-ray-based waste inspection, allowing penetration of occlusions in waste bags. The method achieved a mAP50 of 63.22%, driven by an “easy-to-hard” segmentation strategy and a ResNet-101-FPN backbone, revealing how novel data modalities can overcome visibility challenges in traditional imaging.

Comparative analysis of smart bin integration and real-time deployment features is evident in Wang et al. (2021), Gunaseelan et al. (2023), Rahman et al. (2020), and Sallang et al. (2021). Wang et al. (2021) presented a cloud-integrated system using seven CNNs, where MobileNetV3 reached 94.26% accuracy with a model size of 49.5 MB and inference time of 261.7 ms. Similarly,

Gunaseelan et al. (2023) utilized a modified ResNeXt architecture coupled with ResNet-50 for dual-network classification across biodegradable and hazardous categories, achieving an impressive 98.9% overall accuracy. The study also features a smart bin with real-time control and solar-powered hardware. Rahman et al. (2020) reported a simpler two-class classification system for digestible vs. indigestible waste at 95.31% accuracy, integrated with a sensor-driven Android interface. Sallang et al. (2021) rounds out this group with an SSD-MobileNetV2-based solution, achieving an mAP of 92.16% on Raspberry Pi, enabling automated sorting and real-time monitoring via LoRa and GPS.

By optimizing the IRD hyperparameters via an Arithmetic Optimization Algorithm, the system achieves 98.61% accuracy. Moreover, Kiyokawa et al. (2021) employs DeepLabv3+ on a 5,366-image dataset of construction waste, achieving 0.56 mIoU with robustness to real-world variations such as lighting and moving vehicles. These works emphasize the importance of tailored model designs and data strategies to accommodate object scale and environmental complexity. A unique take-on problem comes from Iqbal et al. (2022), which uses video analytics on edge devices for plastic bag contamination detection. YOLOv4 and CSPDarkNet_tiny achieved an mAP of 63% at 24.8 fps on a Jetson TX2, proving that real-time deployment of high-speed models in constrained environments is feasible. Continuous training and deployment loops enhanced long-term system performance and minimized false detections. The system's alignment with industrial settings represents a trend toward sustainable, data-driven operations.

Finally, Mookkaiah et al. (2022) added another dimension by incorporating hybrid pooling and batch normalization into a ResNet V2-based architecture for MSW classification, yielding a 19.08% improvement in accuracy over traditional methods. This demonstrates how nuanced architectural choices can provide substantial performance improvements even in basic binary classification tasks. Together, these studies illustrate a rapidly evolving landscape where real-time capability, accuracy, and deployment efficiency are equally valued. Papers such as Majchrowska

et al. (2021), Nahiduzzaman et al. (2025), Ahmed et al. (2023), and Gunaseelan et al. (2023) consistently push for high accuracy through network innovation, whereas Prasad et al. (2025), Qiu et al. ((2022), and Kiyokawa et al. (2021) emphasize segmentation robustness in realistic scenarios. Works such as Wang et al. (2021), Rahman et al. (2020), and Sallang et al. (2021) stress end-to-end smart system integration, indicating a holistic approach to waste management through AI. Collectively, the field is moving toward scalable, explainable, and context-aware AI systems capable of functioning across diverse real-world waste management environments.

Finally, invaluable insights into dataset diversity and representation, particularly for small or complex objects, are addressed in Alsubaei et al. (2022) and Kiyokawa et al. (2021). Alsubaei et al. (2022) outline DLSODC-GWM, a method focused on small object detection, which consists of using an improved RefineDet (IRD) with a Functional Link Neural Network (FLNN). While recent studies have made notable advancements in waste classification and segmentation through dual-stage models, lightweight networks, and smart bin integration, still challenges remain towards developing a system ready for real-life deployment. Most approaches either rely on image classification with limited real-world applicability or require high computational resources unsuitable for embedded deployment. Additionally, segmentation models often struggle with cluttered, overlapping waste in unstructured environments, and synthetic data generation is rarely explored to improve dataset diversity. These gaps highlight the need for a unified, efficient, and deployable system that combines robust detection, real-time performance, and adaptability to complex waste scenarios.

3. Dataset and Methodology

The overall research framework is systematically structured into several distinct phases, encompassing fundamental data acquisition, the design and development of the model architecture, comprehensive numerical and visual analyses, and ultimately, the deployment of the trained model

integrated with robotic hardware for real-time application and evaluation. Figure 1 provides a detailed representation of this study's complete technical workflow and analytical components, illustrated through structured block diagrams for enhanced interpretability.

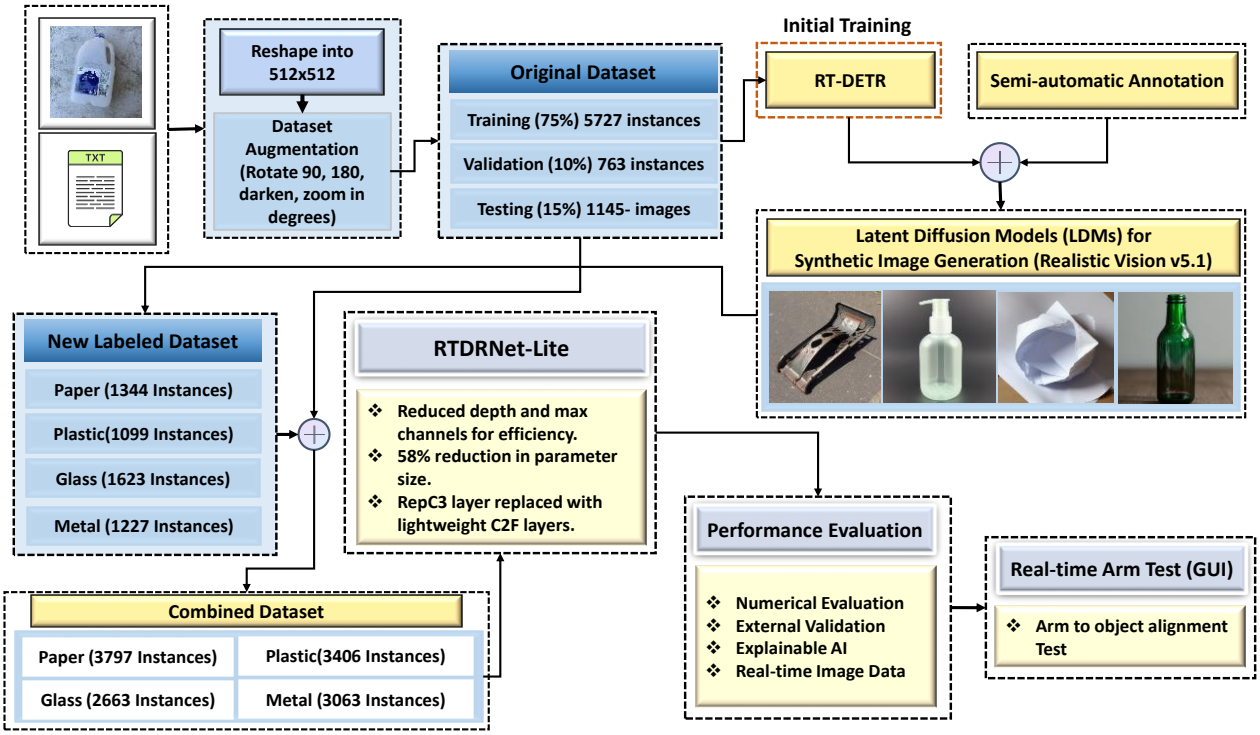


Fig. 1. Block diagram of the proposed waste detection framework integrating real-time robotic evaluation.

3.1.1 Dataset Preprocessing

The foundational dataset utilized in this research was from Kaggle, titled "RealWaste Image Classification" by Joakim Arvidsson (2024). This dataset originally comprised nine distinct classes of waste, encompassing both organic and inorganic materials. Given the objective of this study, to detect and classify waste materials based on their visual characteristics and material composition, certain class consolidations were performed. "Vegetation" and "Food organics" classes were excluded from the dataset due to their handling inefficiency in robotic arm sorting systems and

their often moist, deformable, or irregular physical consistency (Kharola et al., 2022), which poses challenges for conventional robotic grippers.

The ‘Cardboard’ and ‘Paper’ classes were combined into a single type due to their interchangeable physical features and visual resemblance. The ‘Miscellaneous Trash’ and ‘Textile Trash’ classes were removed from the dataset because of their complexity for model detection in a cluttered environment, weak economic importance, and very low number of appearances in the data. The final dataset was restructured into 4 primary dominant classes, one representing organic waste as paper waste, and three representing inorganic waste: metal, plastic, and glass. The preprocessing phase commenced with data augmentation to increase the dataset's diversity and improve model generalizability. Each image underwent three types of augmentations: rotation at 90°, 180°, mirror, and zoom-in transformations. This process aims to simulate various real-world orientations and scales of waste objects. Following augmentation, an average hash algorithm was applied to identify and eliminate redundant or near-duplicate images, ensuring appropriate data for the model by removing unnecessary and repetitive data that cause model overfitting (Ying, 2021). The LabelImg image annotation tool was subsequently employed for manual labeling. Bounding boxes were created around each object, and labels were assigned to their respective categories to facilitate object detection training. Figure 2 desmostrates the overall dataset preprocessing pipleline.

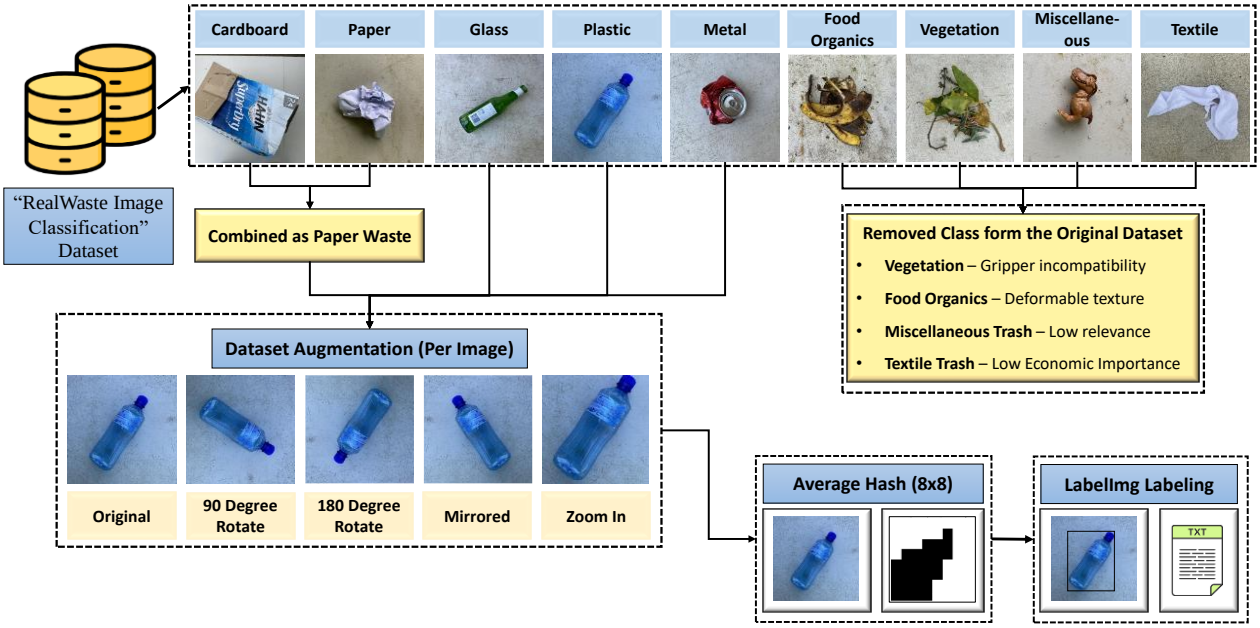


Fig. 2. Dataset preprocessing pipeline

3.1.2 Semi-Automatic Annotation

One of the major distinctions of the proposed method is that, given the class imbalance in the original dataset, instead of solving this problem conventionally, artificial synthetic images were generated. To generate synthetic waste images for underrepresented categories, we utilized the Realistic Vision v5.1 (SG161222, n.d.) model, a high-fidelity text-to-image diffusion model based on Stable Diffusion, which was trained on millions of image-caption pairs from the LAION-5B dataset using the latent diffusion framework. This model is the result of extensive fine-tuning and checkpoint merging of photorealistic diffusion models, specifically designed to enhance texture, object clarity, and scene realism. It builds upon the Stable Diffusion 1.5 backbone, using classifier-free guidance and large-scale datasets to generate high-quality, prompt-aligned images. Using the Diffusers library, a generation pipeline capable of producing realistic variations of target objects, such as regular plastic, metal, paper, and glass items, was constructed by prompting the model with descriptive phrases. The guidance scale was set to 8.5 to ensure strong alignment between the

prompt and the generated image, while maintaining visual diversity. Additionally, 60 inference steps were applied to achieve a balance between generation quality and computational efficiency.

This particular model was chosen over other diffusion-based models like Stable Diffusion v1.5 or 2.1 due to its strength in creating photo-realistic outputs with natural textures, consistent lighting, and realistic object shapes. These qualities made it especially suitable for producing synthetic images that closely resemble the real waste images used in training. By carefully writing prompts that described common waste materials like plastic bottles, metal cans, and glass fragments, we were able to generate images that matched the visual style, background simplicity, and lighting conditions of our real-world dataset. This helped reduce any noticeable difference between real and synthetic images, ensuring that the model wouldn't overfit to either domain. All synthetic images were also resized to 512×512 pixels and underwent the same augmentation steps (rotation, zooming) as the original images to further align their appearance. This approach allowed us to increase the number of examples for underrepresented classes like glass and metal, while maintaining visual consistency across the dataset. As a result, the detection model could learn to recognize a wider range of appearances, including different angles, partial views, or occlusions of the same object type, conditions that often occur in real-world waste environments. Object detection models do not depend on whether an image is real or synthetic, they learn from repeated patterns, object shapes, and spatial features, all of which the diffusion model captures effectively. Once the synthetic images were generated, we used a semi-automatic labeling process. A pre-trained detection model was run on the new images to predict bounding boxes, and only those with high confidence, above 90% were retained. We then reviewed these predictions manually to ensure accuracy. This process saved significant time while still maintaining high annotation quality. An overview of this image generation and labeling workflow is provided in Figure 3.

Initially, a detection model was trained using the RT-DETR architecture on the primary dataset mentioned in Table 1. Following the development of the initial detection model, the synthetic

images were compiled into a single directory. To complete the annotation of the new dataset, a Python-based script running the trained detection model was used to automatically detect and localize objects present in these unannotated images. For each image, the model was used to predict the object classes as well as their respective bounding boxes. These predictions were stored in a record, and the annotation files for each image were automatically generated using the predicted class indices and the bounding box coordinates. In order to ensure the high quality of the generated annotations with maintained dataset integrity, a confidence threshold of 90% was implemented, where any prediction below the threshold was ignored. The threshold was used to filter out low-confidence detections to ensure that only the truly reliable object annotations were preserved.

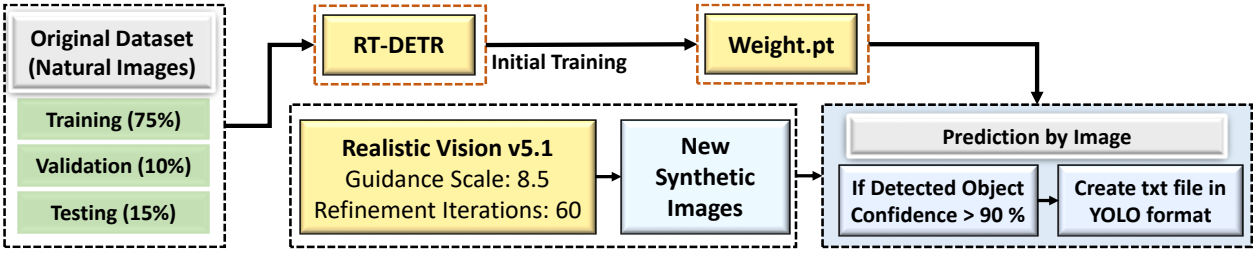


Fig. 3. Semi-automatic annotation for synthetic image labeling.

Afterwards, manual verification was performed for all the predicted datasets to disregard any mislabeled predictions. As a result, this semi-automated annotation process returned a total of 5293 additional object instances, substantially increasing the size of the dataset. The expansion was proven beneficial for improving the model's robustness, generalization ability, and overall performance in different realistic multimodal and multi-class waste classification environments. Table 1 shows a comparison between the previous and current class image numbers, and Figure 4 shows some samples of the natural images against the synthetic images present in the final dataset. In comparison to our previous papers on AI-based waste recycling, Sayem et al. (2024) utilized both image classification and object detection approaches using the WaRP-C and WaRP-D datasets

(Parohod, 2023), respectively. While both datasets initially appeared reliable due to the presence of multiple waste classes, further inspection of WaRP-C revealed significant class imbalance, with most categories containing very few images. Similar issues were observed in WaRP-D, where some classes had over 200 instances while others had fewer than 30, resulting in highly skewed data that was unsuitable for robust training. On the other hand, the dataset used in Nahiduzzaman et al. (2025) included a higher number of images; however, closer examination revealed multiple issues, such as misclassified waste categories, random irrelevant objects due to web scraping, and generally small and noisy samples. In contrast, the dataset used in this research combines naturally captured images with high-quality, AI-generated synthetic data to provide a more balanced and representative training set.

Table 1.

Dataset comparison consisting of both natural and synthetic images.

Combined Dataset (Natural + Synthetic Data)		**Natural Dataset**	
Class	**Instances**	**Class Name**	**Instances**
Paper	3797	Paper	2453
Plastic	3406	Plastic	2307
Glass	2663	Glass	1040
Metal	3063	Metal	1836

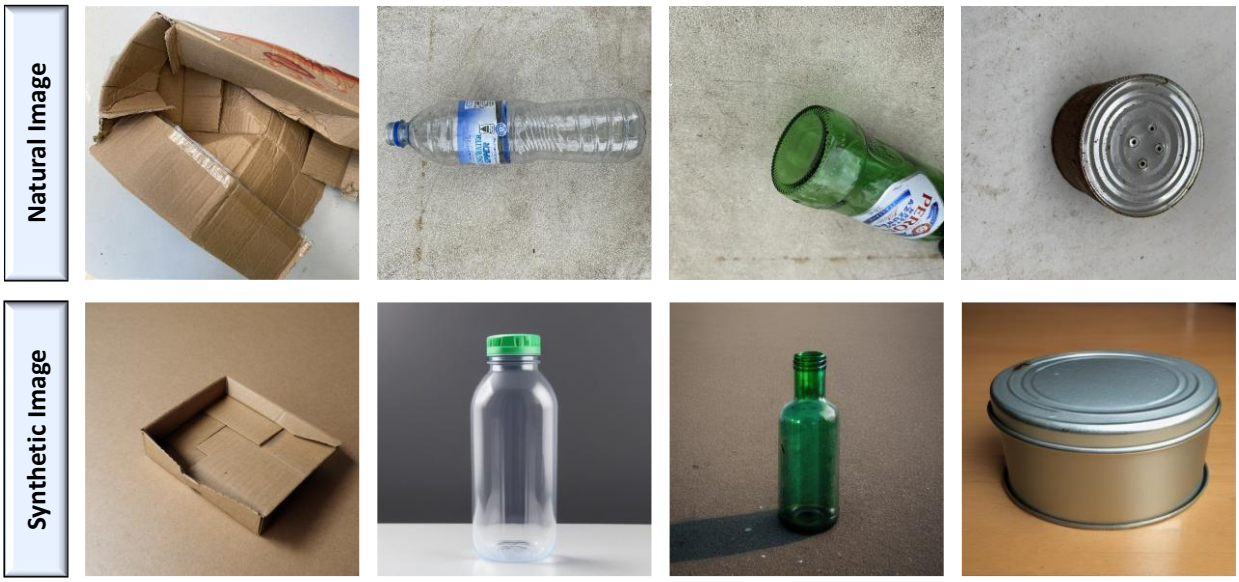


Fig. 4. Samples and visual comparison of natural and synthetic images in the dataset.

3.2 Model Architecture

3.2.1 Proposed Model

This modified version of the RT-DETR-large architecture, RTDRNet-lite, features a carefully re-engineered design to improve computational efficiency without significantly sacrificing detection performance. The main goal of this revision is to address the challenges encountered in resource-constrained environments, where computational power, memory bandwidth, and power consumption are heavily limited. Additionally, these AI and robotic hardware systems require substantial power, as they are designed to operate continuously; therefore, integrating a lightweight AI model is crucial to ensure energy efficiency and align with the primary goal of building a sustainable, low-resource waste sorting solution. To achieve this, the model implements several strategic architectural changes that greatly reduce the number of parameters and floating-point operations per second (FLOPs), enabling it to operate effectively in real-time or near-real-time deployment scenarios. Figure 5 shows the block diagram of the proposed RTDRNet-lite architecture. A key aspect of change occurs in the backbone of the architecture, which has been intentionally downscaled to provide a more compact and efficient representation of input features.

The initial layers of the network, including the HGStem and HGBlock modules, have been optimized. Originally configured with dimensions of (32, 48) and (48, 128), these have been reduced to (24, 32) and (32, 96), respectively.

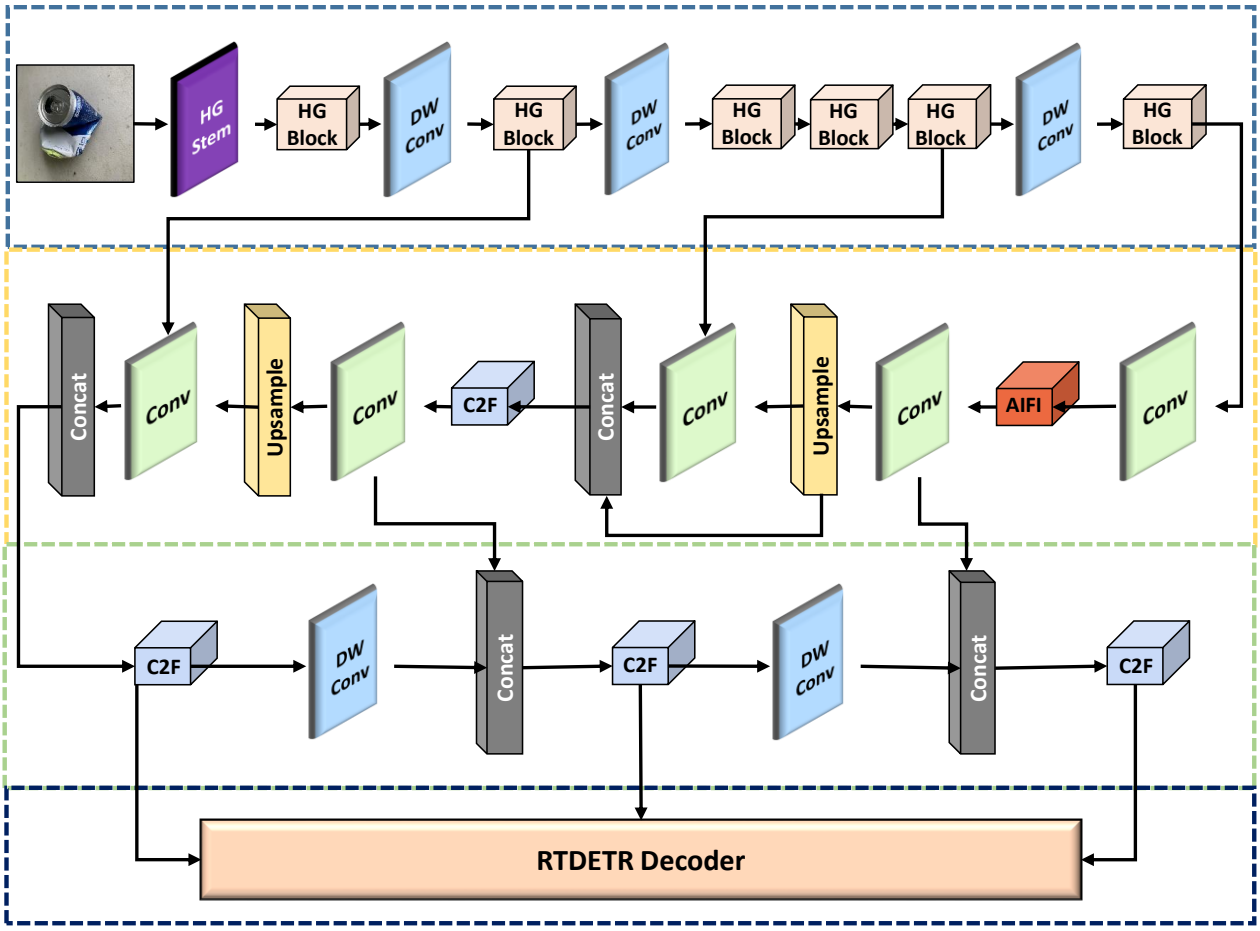


Fig. 5. Block diagram of RTDRNet-lite model architecture.

This downscaling guarantees an initial reduction in the computational requirements of feature extraction, especially in the early stages, which is important for low-latency computing. Additionally, the models have been minimized in both depth and channel width as they advance through deeper stages. This reduction leads to a substantially reduced computational footprint, while the model still has the capacity for hierarchical characteristic processing, which results in robust object detection. In other words, efforts have been made to increase the representational capability of a network while simultaneously optimizing its computational expenditure.

Furthermore, the number of HGBlocks replicated in each stage is reduced, with a limit of three. This avoids the risk of over-parameterization and a gigantic depth that is not supported by the achievable precision. Furthermore, the channel width in deeper layers, particularly those that execute DWConv and HGblocks, has been kept minimal. This not only produces smaller intermediate feature maps but also ensures that there is less memory access and time to be invested. Figure 6 shows the block diagram for DWConv, where each input channel of shape $\boldsymbol{H} \times \boldsymbol{W} \times \boldsymbol{C}$ is filtered independently using a $\boldsymbol{K} \times \boldsymbol{K}$ kernel, resulting in an output of shape $\boldsymbol{H'} \times \boldsymbol{W'} \times \boldsymbol{C}$. Unlike standard convolution, there is no cross-channel mixing. This operation, illustrated above, reduces computational complexity and is well suited for real-time applications. Equation (1) provides a mathematical formulation of depthwise convolution, demonstrating how filtering is applied independently to each channel, thereby reducing the computational cost.

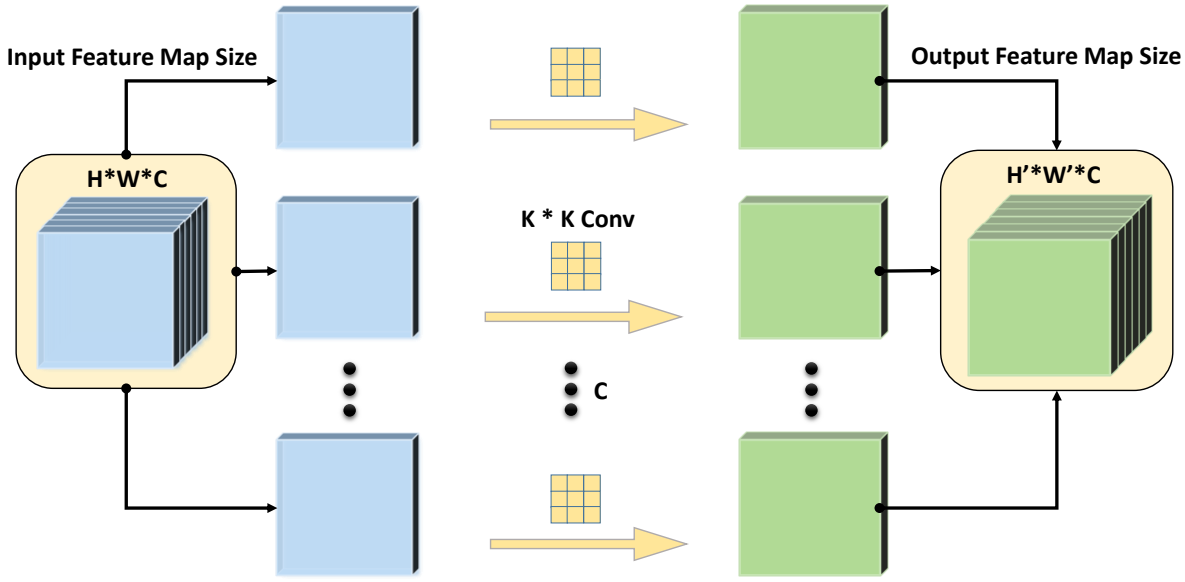


Fig. 6. Block diagram for depth-wise convolution (DWConv) module.

$$Y_{i,j,c} = \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} X_{i+m,j+n,c} . W_{m,n,c} \quad (1)$$

$Y_{i,j,c}$ is the output at location (i,j) in channel C, X is the input feature map, W is the depthwise filter with shape $K \times K \times C$, where K is the kernel size and C is the number of channels. Unlike standard convolution, filtering is applied *per channel*, with no cross-channel mixing. DWConv reduces complexity from $O(K^2 \cdot C_{in} \cdot C_{out})$ to $O(K^2 \cdot C_{in})$, which is ideal for real-time models. In the model's detection head, key components have been retained and adapted rather than eliminated. The AIFI (Attention-Integrated Feature Interaction) module remains a central part of the architecture in capturing global contextual information and improving the robustness of detection outcomes. However, this module and its associated components have been simplified to reduce their parameter burden. The initial projection layer that feeds into the AIFI module now reduces the channel dimensionality to 192. Furthermore, the AIFI module itself operates with six attention heads instead of the original eight, decreasing its parameter load while retaining much of its functional efficacy. Equation (2) shows the scaled dot-product attention mechanism used in the AIFI module to compute attention scores based on the relationship between query and key representations.

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

The input feature map X is linearly projected into three distinct representations: queries $Q = XW_Q$, keys, $K = XW_K$, and values, $V = XW_V$. d_k are dimensionality keys, and W_Q, W_K, and W_V are learnable weights. In designing RTDRNet-lite, we replaced the original RepC3 modules with C2F (Cross-Stage Partial Fusion) modules to reduce computational load and model size without sacrificing detection performance. While RepC3 blocks are effective at capturing features, they involve deep stacking of convolutions and introduce considerable parameter overhead, which can be excessive for real-time applications on limited hardware. In contrast, the C2F module takes a more efficient approach by splitting input channels, transforming only part of the data, and then merging it back. This allows the model to retain important feature information while using fewer

computations and less memory. From a learning standpoint, the C2F module also improves gradient flow and feature reuse, which helps the network learn more effectively even with fewer layers. This structure encourages the model to focus on the most relevant spatial features without introducing unnecessary complexity. The impact of this change is evident in the results: switching to C2F helped reduce the model's parameter size by around 58%, yet the performance remained strong, achieving 97% mAP@50, only slightly below the original RT-DETR. As shown in Section 4.1, the model continued to perform reliably across all waste categories, indicating that the C2F modules provided a good balance between efficiency and feature extraction quality. Figure 7 presents the block diagram for the HGBlock and C2F modules.

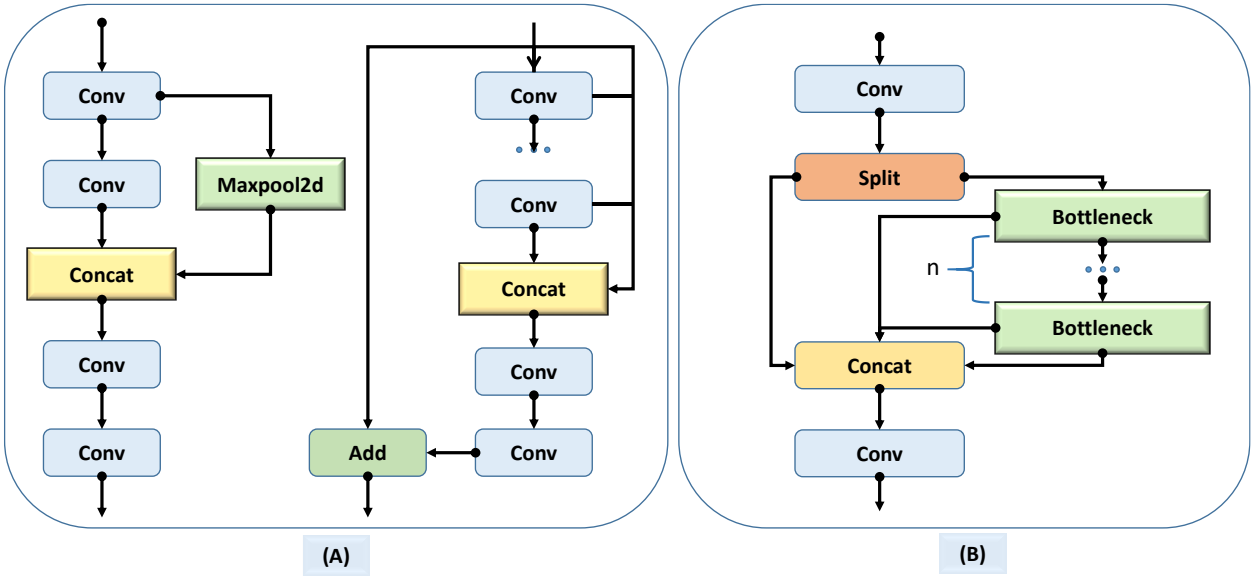


Fig. 7. Block Diagram for HGBlock (A) & C2F (B) module

In summary, this updated version of the RT-DETR-l model represents a comprehensive optimization of the original architecture. By aggressively compressing the backbone, simplifying the detection head, and integrating lightweight alternatives to standard components, the model achieves a highly efficient design. While this streamlined architecture may introduce trade-offs, particularly in detecting small or highly complex objects, the resulting gains in speed,

deployability, and energy efficiency make it an excellent candidate for real-time applications in edge AI and autonomous robotics. The model stands as a clear example of how intelligent architectural simplification can meet the demanding performance constraints of modern low-power, real-time systems without entirely sacrificing the sophistication of deep learning-based object detection.

3.3 Model Performance Evaluation

To evaluate the detection accuracy of the RTDRNet-lite model, mAP@50 is used as a vital performance indicator. This metric reflects the model's ability to balance both precision, that is, how many detected items are truly relevant, and recall, how many relevant items are correctly detected. The "50" in mAP@50 refers to the IoU (Intersection over Union) threshold of 50%, implying that a predicted bounding box is correct if it overlaps the ground truth box by at least half. At this threshold, the average precision for each class (paper, plastic, glass, and metal) is calculated, and the mean is taken across all classes, as shown in equation (3):

$$mAP@50 = \frac{\sum \left(AP_{Class\_Paper} + AP_{Class\_Glass} + AP_{Class\_Metal} + AP_{Class\_Plastic} \right)}{4} \quad (3)$$

In addition to mAP@50, the model's performance is also evaluated using mAP@50:0.95, which is a more rigorous metric commonly used in COCO-style evaluations. It averages the precision across IoU thresholds ranging from 0.50 to 0.95, with a step size of 0.05. This variation provides a better understanding of how accurately the model can localize objects with different bounding box overlap tolerances and is more challenging than a fixed threshold metric.

Precision is the proportion of all positive detections that are relevant to the sum of positive detections that are actually correct. The larger the value is, the more reliable the model is at

determining relevant findings. On the other hand, recall evaluates the model's capacity to identify all actual instances of a given waste class in the dataset, defined by the proportion of correctly detected instances among all actual occurrences of that class, including those missed or incorrectly labeled. Equation 4 and 5 gives a detail on Precision and Recall calculation.

$$\mathrm{Pr}\,ecision = \frac{\text{Number of Correctly Identified Waste Class}}{\text{Number of Correctly Identified Waste Class} + \text{Number of Incorrectly Identified Waste Class}} \quad (4)$$

$$\mathrm{Re}\,call = \frac{\text{Number of Correctly Identified Waste Class}}{\text{Number of Correctly Identified Waste Class} + \text{Number of Incorrectly Mislabeled Waste Class}} \quad (5)$$

To capture the balance between these two, the F1 score is computed. It is the harmonic mean of precision and recall and helps validate the model's reliability, especially in scenarios where both false positives and false negatives are critical. The F1 score is computed using Equation (6).

$$F1\ Score = \frac{2 * \mathrm{Pr}\,ecision * \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \quad (6)$$

These combined metrics, mAP, precision, recall, and F1-score offer a detailed assessment of the RTDRNet-lite model's performance in waste detection, confirming its ability to perform accurately in both numerical evaluations and real-world sorting conditions.

3.4. Software and Hardware Integration

The objective of this research is not only to develop an AI-based detection model for waste item classification but also to establish a sustainable approach for the post-processing stage, specifically, the development of a robotic arm for automated waste sorting on the basis of the output of the AI

model. The aim is to reduce reliance on conventional waste sorting methods, such as manual human-based sorting, which are time-consuming, labor-intensive, and prone to error. In contrast, a robotic arm can operate continuously, day and night, without fatigue or performance degradation. Figure 8 illustrates the general working pipeline of the proposed system, which integrates both AI software and robotic hardware. Once the detection model is trained, it is deployed onto a prototype platform featuring a conveyor-like system with a lateral robotic arm. A Logitech C270 webcam is positioned above the platform to capture every object placed on it. The AI model detects the object within a designated quadrilateral bounding box and calculates its center pixel position, which is considered the object's central location. These coordinates are then transmitted to a connected slave device, an Arduino Mega 2560, which uses the data to compute the angular positions for each of the four robotic arm joints through inverse trigonometric calculations.

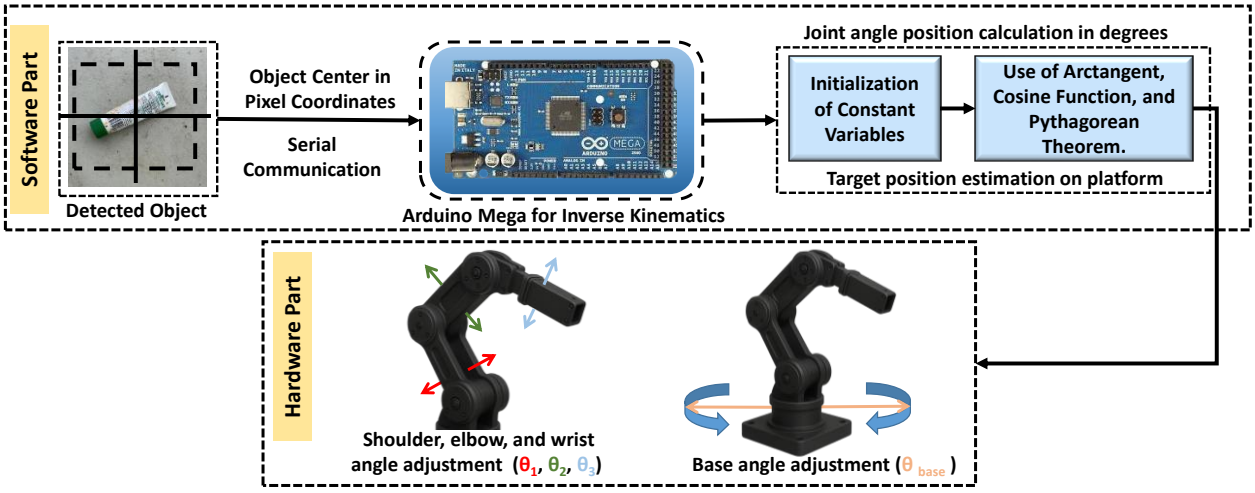


Fig. 8. Software & hardware integration pipeline with RTDRNet-lite model and 4 DoF robotic arm.

Figure 9(a) and Figure 9(b) present graphical representations of the working principle of the 4-DoF robotic arm used in this research. In Figure 9(a), the quadrilateral surface ABCD denotes the field of view captured by the webcam positioned above the detection platform. The coordinates of these points represent the pixel positions from the camera's perspective, which operates at a resolution of 640×480 pixels (width × height). Points F and O correspond to the camera center and the robotic

arm base, respectively, in the camera's coordinate frame. Considering two random points, G and H represent the centers of two detected objects located on the platform within the area ABCD. The respective joint rotations of the robotic arm required to reach these points are determined using a series of Pythagorean and inverse trigonometric calculations. Considering G(Ti, Tj) as the pixel coordinates of the first detected object center, two perpendicular lines GI and GL are drawn from point G to lines EO and DC, respectively. To calculate the rotation angle $\boldsymbol{\theta_{base\_1}}$ required to turn the base servo of the robotic arm toward G, an inverse trigonometric function is applied, as shown in Equation (7). Here, GI equals (Ti – 320), and IO equals (Tj – 480), both in pixel units. Once $\boldsymbol{\theta_{base\_1}}$ is computed, the base servo rotates accordingly, initiating the arm's movement. The same process is followed when the approaching point H is situated on the right side of line EO. However, in this case, the angle $\boldsymbol{\theta_{base\_2}}$ is always negative, whereas $\boldsymbol{\theta_{base\_1}}$ is positive, reflecting the mirrored rotational directions of the base servo depending on the object's location relative to the central axis.

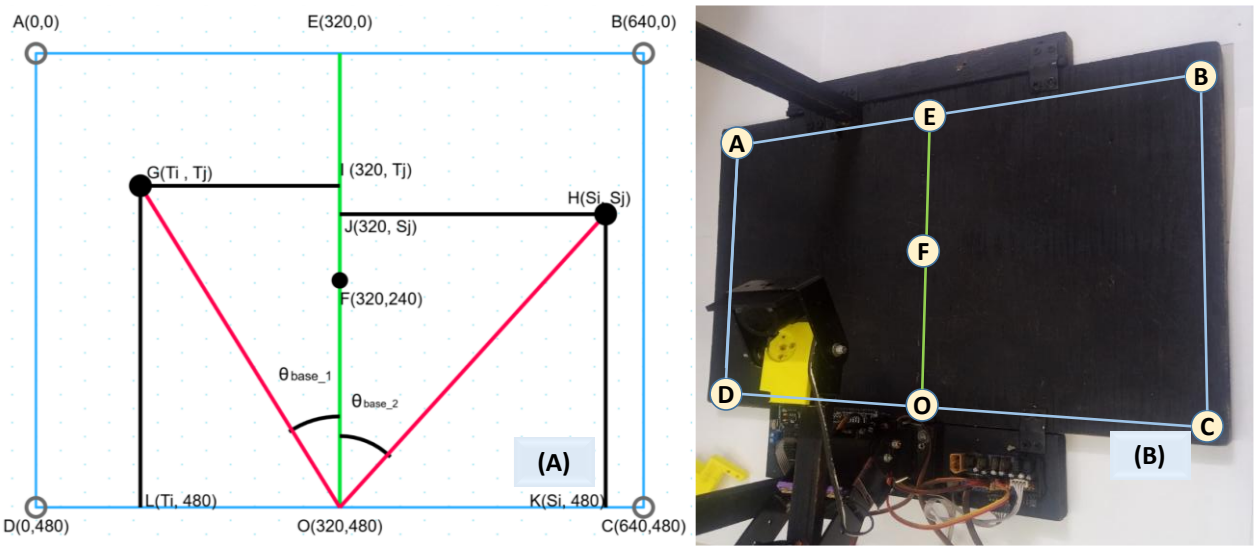


Fig. 9. Working principle for 4 DoF robotic arm base rotation.

$$\theta_{base} = \tan^{-1}\left(\frac{T_i - 320}{T_j - 480}\right) * \left(\frac{180}{\pi}\right) \quad (7)$$

Once the base is aligned, the computation proceeds in the vertical 2D plane formed by the shoulder-elbow-wrist linkage (S-E-W) in Figure 10. The arm joint lengths are defined as follows: L1 is the distance from the platform base ABCD of the arm to the shoulder (S) joint; L2 is the distance from the shoulder (S) joint to the elbow (E) joint; L3 spans from the elbow (E) joint to the wrist (W) joint; and L4 is the distance from the wrist (W) joint to the end effector (P), all in cm. Before proceeding further, camera calibration was performed using pixel-to-distance mapping, through which each pixel was found to correspond to 0.0625 cm. This value was then used to convert all arm joint lengths from their centimeter units to equivalent pixel units for alignment and computation ease. On the basis of the position of the detected object on the platform, the total extension distance D, which represents how far the robotic arm needs to reach, is calculated using the Pythagorean theorem. In this context, D corresponds to the hypothetical GO, as illustrated in Figure 9.

Later, the distance from the shoulder (S) joint to the wrist (W) joint, denoted as **R**, is determined using the Pythagorean theorem again from two known components: the horizontal displacement ***d*** and the vertical offset $\boldsymbol{Y_{offset}}$. Both equations are shown in equation (8). Notably, in this setup, the end-effector lies below the shoulder joint; hence, $\mathbf{Y_{offset} = L1 - Y_{ee}}$, where **L1** is the vertical length of the first link and $\mathbf{Y_{ee}}$ is the minimum vertical height of the end-effector when it reaches for objects.

From equation (9), $\boldsymbol{\alpha_1}$ is defined as the angle between line **R** and the horizontal line (equal to **d**), computed using the arctangent function, $\tan^{-1}$ of $\mathbf{Y_{offset}}$ over **d**. The second angle, $\boldsymbol{\alpha_2}$, represents the angle between links **L2** and **R** and is calculated using the cosine rule, as expressed in equation (9). The effective shoulder joint angle, $\boldsymbol{\theta_1}$, is then obtained by subtracting $\boldsymbol{\alpha_1}$ from $\boldsymbol{\alpha_2}$, thereby aligning

the upper arm with the target point. The elbow angle, θ_2, is calculated via the law of cosines, as expressed in equation (10), on the basis of the known side lengths, and the required bend at Joint 3, the elbow, is determined. Finally, the wrist angle, θ_3, is computed via equation (11) to compensate for the accumulated joint rotations, ensuring that the final link, **L4**, which holds the end-effector, remains horizontally aligned and exactly $\mathbf{Y_{ee}}$ height above the platform ABCD. This guarantees that the gripper or tool at the end maintains the desired orientation, typically parallel to the base reference plane. This stepwise inverse kinematics approach allows precise joint positioning in response to any arbitrary target point within the reachable workspace, enabling the robotic arm to perform accurate pick-and-place objects, in this case, and waste materials through visual guidance.

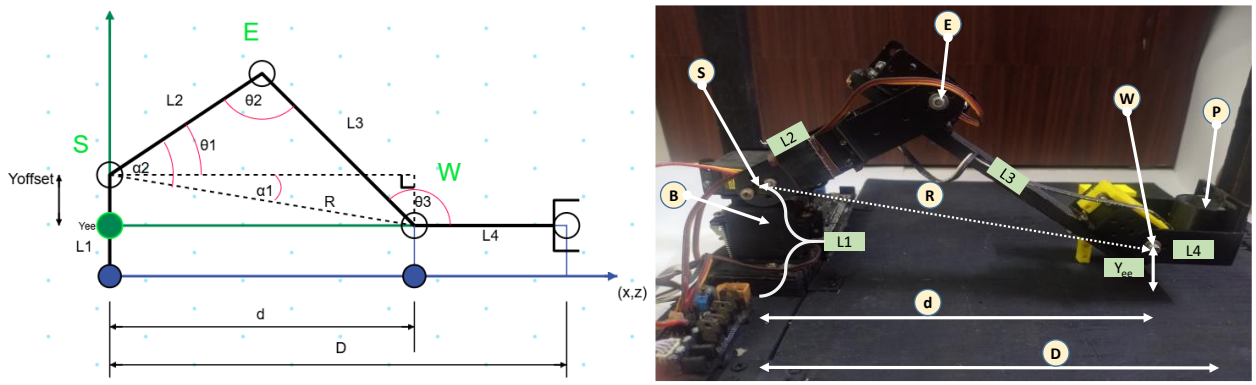


Fig. 10. Working principle for 4 DoF robotic arm shoulder-elbow-wrist linkage

$$D = \sqrt{(T_i - 320)^2 + (T_j - 480)^2}, \quad R = \sqrt{d^2 + Y_{offset}{}^2} \tag{8}$$

$$\alpha_1 \tan^{-1}\left(\frac{Y_{offset}}{d}\right) * \left(\frac{180}{\pi}\right), \quad \alpha_2 = \cos^{-1}\left(\frac{L2^2 + R^2 - L3^2}{2 * L2 * R}\right) * \left(\frac{180}{\pi}\right) \tag{9}$$

$$\theta_1 = \alpha_2 - \alpha_1, \quad \theta_2 = \cos^{-1}\left(\frac{L2^2 + L3^2 - R^2}{2 * L2 * L3}\right) * \left(\frac{180}{\pi}\right) \tag{10}$$

$$\theta_3 = 180 - [(180 - (\alpha_2 + \theta_2)) + \alpha_1] \tag{11}$$

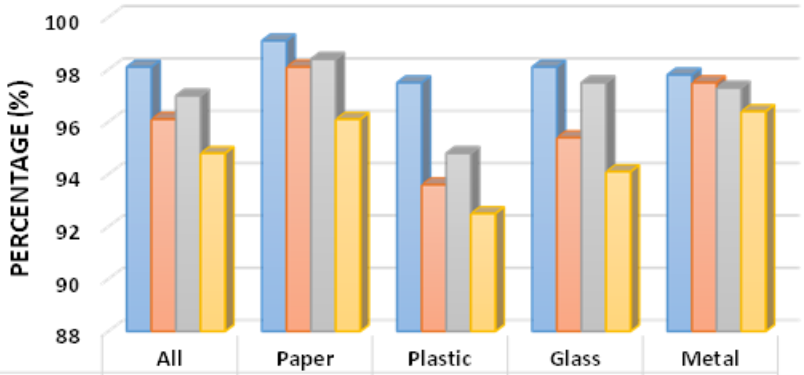
4. Analysis of experimental results

The proposed RTDRNet-lite is thoroughly evaluated in this section with quantitative and qualitative experiments. The evaluation includes general evaluation metrics such as precision, recall, F1-score, mAP, and confusion matrix, as well as an in-depth study of localization performance based on Intersection over Union (IoU) comparisons. The model's generalization capability is assessed through external validation on unseen datasets, while heatmap visualization and GUI-based real-time testing are used to demonstrate interpretability and practical deployment feasibility. Each of the sub-sections has relevance both from a numerical validation point-of-view and also in the real world.

4.1. Performance Metrics Evaluation

The evaluation metrics in Figure 11 highlight a significant performance increase in the proposed RTDRNet-lite model, reflecting its maturity for real-world deployment. The model achieves an overall precision of 98.1 % and a recall of 96.1%, with a mAP of 97% at an IoU of 0.5 and 95.8 at an IoU of 0.5–0.95. These results indicate excellent localization accuracy and class confidence across diverse waste categories. Among the individual classes, Paper consistently outperforms the other categories, achieving 99.1% precision, 98.1% recall, 98.4% mAP@50, and 96.1% mAP@50–95. This suggests that the model can detect and localize paper waste with high consistency and minimal ambiguity. Glass and Metal also exhibit robust results, both exceeding 97% precision and scoring above 94% mAP@50–95, reaffirming their well-separated feature representation in the model's learned space. Plastic, while performing well with 97.5% precision and 94.8% mAP@50, remains relatively weak in terms of recall (93.6%) and mAP@50–95 (92.5%), which is likely due to background interference or intra-class variability. Nevertheless, all classes exhibit mAP scores well above 0.90, demonstrating strong generalizability and reliability. This balanced distribution

of precision, recall, and mAP metrics confirms RTDRNet-lite's effectiveness not only in numerical terms but also in multi-class stability, making it a promising candidate for scalable waste classification systems.



	All	Paper	Plastic	Glass	Metal
Precision	98.1	99.1	97.5	98.1	97.8
Recall	96.1	98.1	93.6	95.4	97.5
mAP@50	97	98.4	94.8	97.5	97.3
mAP@(50-95)	94.8	96.1	92.5	94.1	96.4

Fig. 11. Performance metrics for RTDRNet-lite model

The training performance of the proposed RTDRNet-lite model was assessed using the mAP@50 and mAP@50–95 metrics for 30 epochs, indicating no significant gains with further training as shown in Figure 12 (A). This plateau suggested that the model had already converged well, and further training would have increased the risk of overfitting without notable benefit. Both curves exhibit a steep rise within the initial epochs, indicating rapid convergence and strong early learning dynamics. mAP@50 increased from 56% to over 90% within just six epochs, whereas mAP@50–95 followed closely, rising from 45% to nearly 87% in the same period. These trends suggest that the model quickly learned core spatial and class-level features, although mild oscillations, particularly between epochs 3 and 7, indicate some sensitivity to training data variation or label noise. Both curves show the trend of progressing stabilization past epoch 10. However, the mAP@50 curve consistently outperforms the mAP@50–95 metric, as expected, because of the

latter's stricter averaging across multiple IoU thresholds. The RTDRNet-lite model achieved optimal performance at epoch 27, with an mAP@50 of 97% and an mAP@50–95 of 94.8%. These results demonstrate strong generalizability and fine-grained detection competency. This implies that the chosen architecture is capable of precise object detection even under the more constrained IoU condition. This is in part due to the adaptable attentional power of the C2F modules and the efficient attention mechanism, which is integral to the RTDRNet-lite model design. This steady performance trend confirms that RTDRNet-lite has the necessary performance for high-accuracy waste object detection in real-time settings.

The normalized confusion matrix in Figure 12(B) provides a comprehensive overview of the class-wise prediction accuracy of RTDRNet-lite. The RTDRNet-lite model retains high class-fidelity for all four major waste classes: paper, plastic, glass, and metal. Each class has more than 94% correctly classified instances, with the paper achieving the highest precision at nearly 98%. Glass and metal also maintained nearly equal performance at 97% and 98% recognition, respectively. Although more prone to mixing with the background class at 4%, strong class prediction with over 94% accuracy was retained. The lower performance of Plastic could mean that overlapping with the background class was a great challenge. A significant observation is the high false-positive rates of the background class for plastic and metal. This 48% and 36% misclassification, respectively, means that the model confuses some plastic and metallic objects with background clutter or misapplies the class due to occlusion. Conversely, the background class had high purity but low recall with paper and glass, likely due to faint edges or reflective surfaces. With respect to intra-class confusion, RTDRNet-lite maintains very high fidelity for solid and well-defined classes, whereas a few adjustments in spatial context, awareness, or hard negative mining can reduce background noise misclassification.

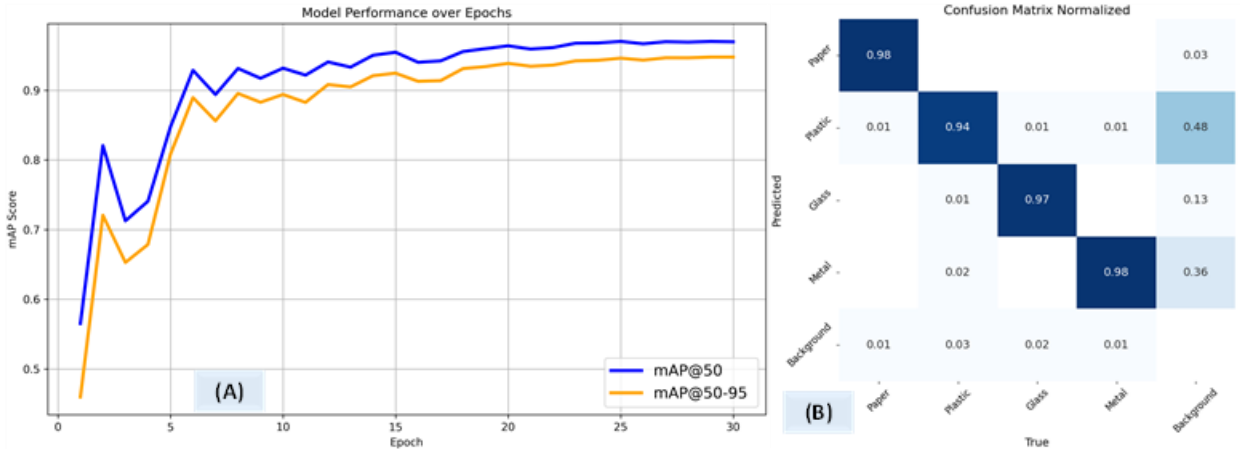


Fig. 12. Model performance curve with mAP@50 and mAP@50-95 (A) & confusion matrix (B)

The F1-confidence curve from Figure 13 (A) offers valuable insight into the prediction confidence thresholds at which the RTDRNet-lite model delivers optimal classification performance. Each class, Paper, Plastic, Glass, and Metal, is represented separately, where their individual F1 scores monotonically increase from low confidence values up to a point of near-optimal performance before they start to drop again due to false negatives arising from overconfidence. The composite F1 score across all classes is shown in bold blue and peaks at 97% at a confidence threshold of 0.766. This point is chosen to achieve the best trade-off between precision and recall to ensure reliable detection performance without risking too many missed detections and false positives. Among the classes, the paper has the most stable confidence and keeps the F1 score at approximately 1.0 over a wide confidence range, which means that the class is easy to isolate and has stable features. In comparison, Plastic has the flattest confidence curve, which means that more uncertain predictions result from the various textures, translucency, and background similarity. Glass and Metal behave similarly but achieve high F1 scores across most of the threshold space. In general, the smoothness and fast reach of the F1 curves of all classes indicate that RTDRNet-lite makes high-quality predictions with high confidence across different waste materials. This

confidence threshold is an essential parameter for real-time waste sorting implementations where fast classification is needed to achieve acceptable processing speeds.

Figure 13 (B), Precision–Recall curves provide a holistic view of the class-by-class performance of the RTDRNet-lite model at varying confidence thresholds, notably highlighting its ability to maintain high precision while preserving a high standard recall value. The detection strength is clearly demonstrated by the high-PR area scores across all classes, which are 98.4% for papers, 94.8 for plastic, 97.5% for glass, and 97.3% for metals. The overall mAP@0.5 over all classes is 97%, as illustrated by the bold blue curve. This paper presents the most stable PR relationship, maintaining near-perfect precision throughout almost the entire recall range. Glass and Metal also achieve strong, consistent performance with minimal decreases in precision even at high recall levels. Plastic, while still achieving a high area under the curve, displays a steeper decline in precision as the recall approaches 1.0. This phenomenon may be a result of instances where the model overpredicts plastic or misclassifies materials such as background noise and semi-transparent objects. The sharpness and affinity of the curves suggest that the model is competent in distinguishing false positives from true positives with high confidence and minimal generalizability. This performance ensures that the feature representation is strong and resilient against overfitting, making the model efficient in real-world waste detection, where recall is essential and should protect against false negatives.

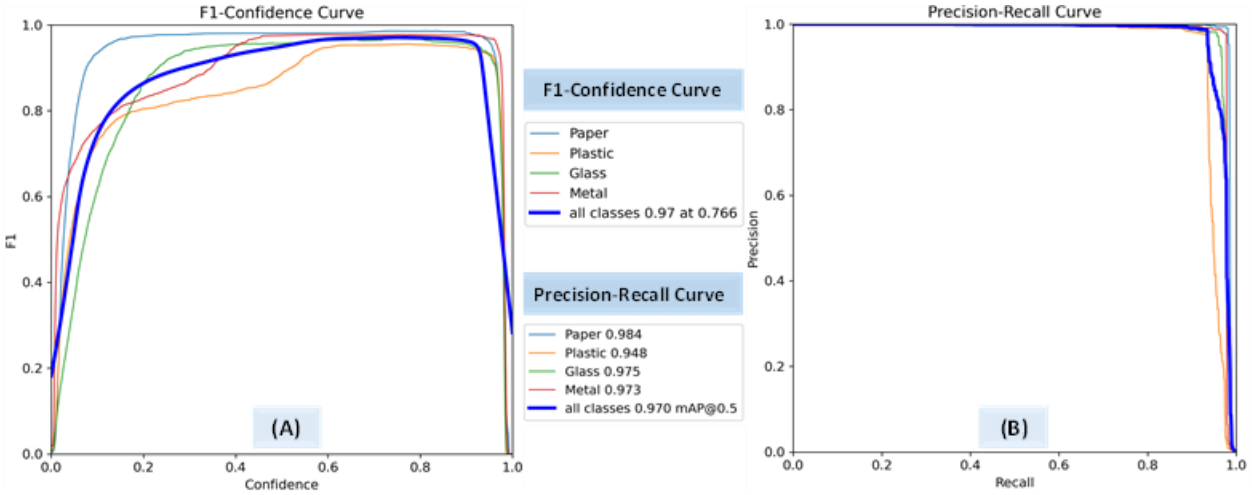


Fig. 13. F1-Confidence Curve (A) & Precision-Recall Curve (B)

4.2.Visual Intersection over union (IoU) test

The qualitative evaluation shown in the 4×5 grid in Figure 14 illustrates the Intersection over Union (IoU) performance of the proposed RTDRNet-lite model across a range of real-world and synthetic waste objects. In each image, the predicted bounding boxes (in red) are compared against the ground truth annotations (in black), offering visual insight into the spatial localization accuracy of the model. Across most samples, the predicted boxes align closely with their ground truth counterparts, demonstrating strong spatial reasoning and high IoU values. Objects with irregular textures, deformities, or varying lighting conditions, such as crumpled plastic, transparent bottles, and metallic wrappers, are accurately enclosed, suggesting the model's resilience to noise and deformation. The consistent overlap across diverse object scales and aspect ratios further reinforces the robustness of the model's localization capability. In a few instances, the red boxes marginally exceed the black ones, indicating slight over-coverage. This behavior may stem from the model's tendency to conservatively estimate object boundaries, potentially as a strategy to avoid under-detection. Interestingly, the objects in the fourth column of the second and last rows exhibited misaligned ground truth boxes. However, the model was able to correctly bind these objects despite

the annotation errors, demonstrating the model's ability to infer object boundaries accurately, even in the presence of imperfect human labeling. Importantly, no major misalignments or omissions are observed, confirming the model's generalization strength. Overall, the visual IoU assessment highlights RTDRNet-lite's high-fidelity bounding box predictions and its effectiveness in complex, cluttered waste scenarios.

Fig. 14. Intersection over Union test, detection by RTDRNet model (Red box) vs ground truth label (Black box)

4.3. External Validation & heatmap analysis.

Figures 15 (A) and (B) show the external validation results of the proposed detection model in two benchmark datasets, the *Trash Detection* dataset and on the *TriCascade Waste Image* dataset. Both datasets consist of various types of paper plastic metal and glass waste, which were not part of the

689 training data. The model exhibits strong generalization capability and can accurately detect and
690 classify various waste categories with different object appearances, backgrounds and lighting
691 conditions. This external validation provides evidence on the generalization performance and the
692 applicability of the model in unseen situations.

693

(A) Trash Detection Dataset

694

Fig. 15. External validation of the RTDRNet-lite model on (A) Trash Detection Dataset & (B) TriCascade Waste Image Dataset

In this study, we adopted EigenCAM to visualize the internal attention of the RTDRNet-lite model, as it is particularly well-suited for transformer-based architectures where traditional gradient-based methods often fall short. Unlike conventional CAM techniques that rely heavily on convolutional spatial gradients, EigenCAM leverages the dominant eigenvectors of activation maps, allowing it to produce robust and visually coherent heatmaps without requiring gradients. This method offered clear and consistent localization of attention, highlighting the most semantically relevant regions of each object, even under occlusion or clutter. The results confirm that EigenCAM provides meaningful visual explanations of the model's reasoning process, further supporting the transparency and trustworthiness of our detection system.

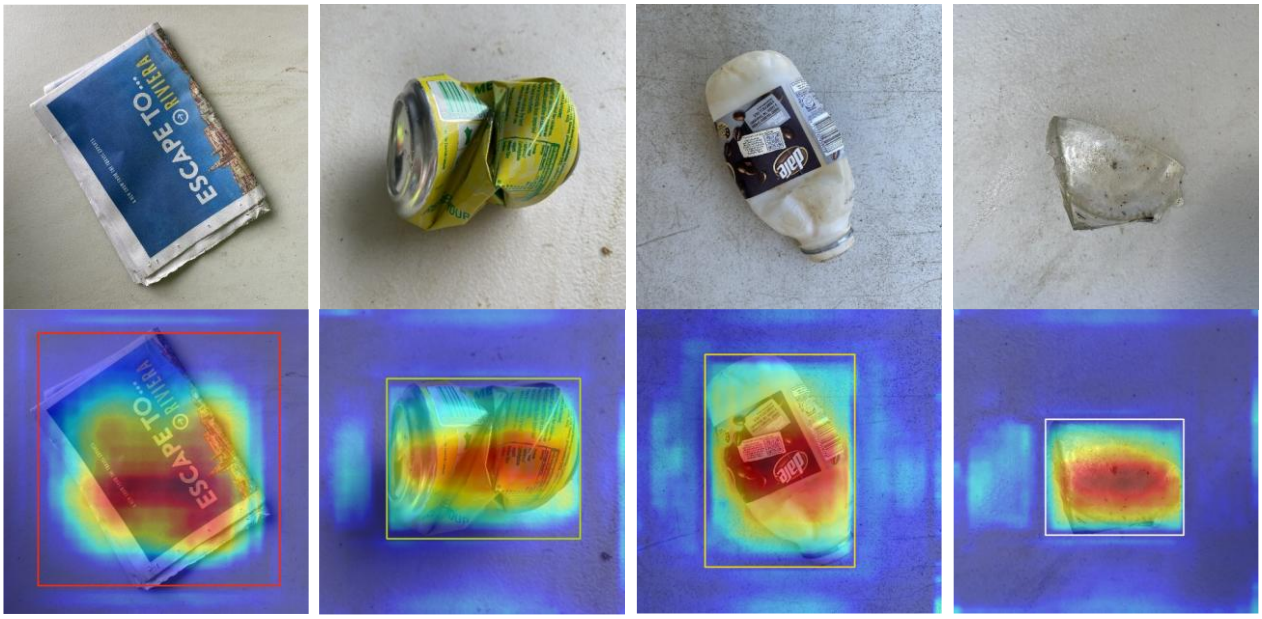


Fig. 16. EigenCAM heatmap analysis for the RTDRNet-lite model

4.4. Real-time analysis with GUI

To facilitate real-time testing of the detection model integrated with the robotic arm hardware, a graphical user interface (GUI) application was developed using PyQt5, enabling live video feed display, detection result visualization, and automated transmission of pixel coordinates to the robotic controller. This interface streamlined the hardware testing process and enhanced user understanding by providing a visual representation of the robot and software operation.

Unlike our previous approaches in Sayem et al. (2024) and Nahiduzzaman et al. (2025) which lacked proper hardware validation and accurate real-time testing, this implementation bridges software and robotics through a functioning GUI and live object detection pipeline, demonstrating practical deployability. As the primary focus was on research and evaluating the detection model's accuracy and the robotic arm's precision in reaching target objects, the hardware was programmed only to position itself over detected objects. The robotic arm did not perform grasping or removal

actions, as these actions were beyond the scope of this phase of the project. The GUI operates in a straightforward sequence, as shown in Figure 17. Upon initialization, it prompts the user to press the “Start Video” button, which loads the trained model and activates the webcam mounted above the hardware platform (ABCD). Once initialized, the model processes a single frame to identify the first object among multiple items. The program was initially executed on a laptop equipped with an AMD Ryzen 5 5500U CPU (overclocked to 4.0 GHz), where each frame required approximately 300 ms for processing, yielding an estimated 3.3 frames per second (FPS). In contrast, when tested on a more powerful server environment such as Kaggle, the model achieved significantly faster performance: around 33 ms per frame (~30 FPS) on an NVIDIA T4 GPU, and approximately 400 ms per frame (~2.5 FPS) on an Intel Xeon 2.2 GHz CPU. In contrast to our previous works (Sayem et al., 2024 and Nahiduzzaman et al. 2025) , which were limited to single-object classification per frame, the proposed system enables simultaneous detection of multiple waste items within a single frame, significantly improving operational speed and practical viability. The results, including the detected object’s class, confidence score, and size, are displayed on the bottom left side of the screen within a designated text area. On the basis of the detection output, the corresponding pixel coordinates are extracted and used to compute the required joint positions for the robotic arm to align itself above the identified object. During this period, the AI model remains idle, conserving computational resources, while the robotic arm completes its movement. This process is repeated iteratively for each object on the platform. This frame-by-frame inference strategy significantly reduces power consumption, as the model processes only one frame per object. For a platform containing **n** objects, the model processes exactly **n+1** frames, **n** frames to detect each object individually, and one additional frame to confirm that no further objects remain on the platform.

A video demonstration showing the system's working principle can be accessed via the following link: Link

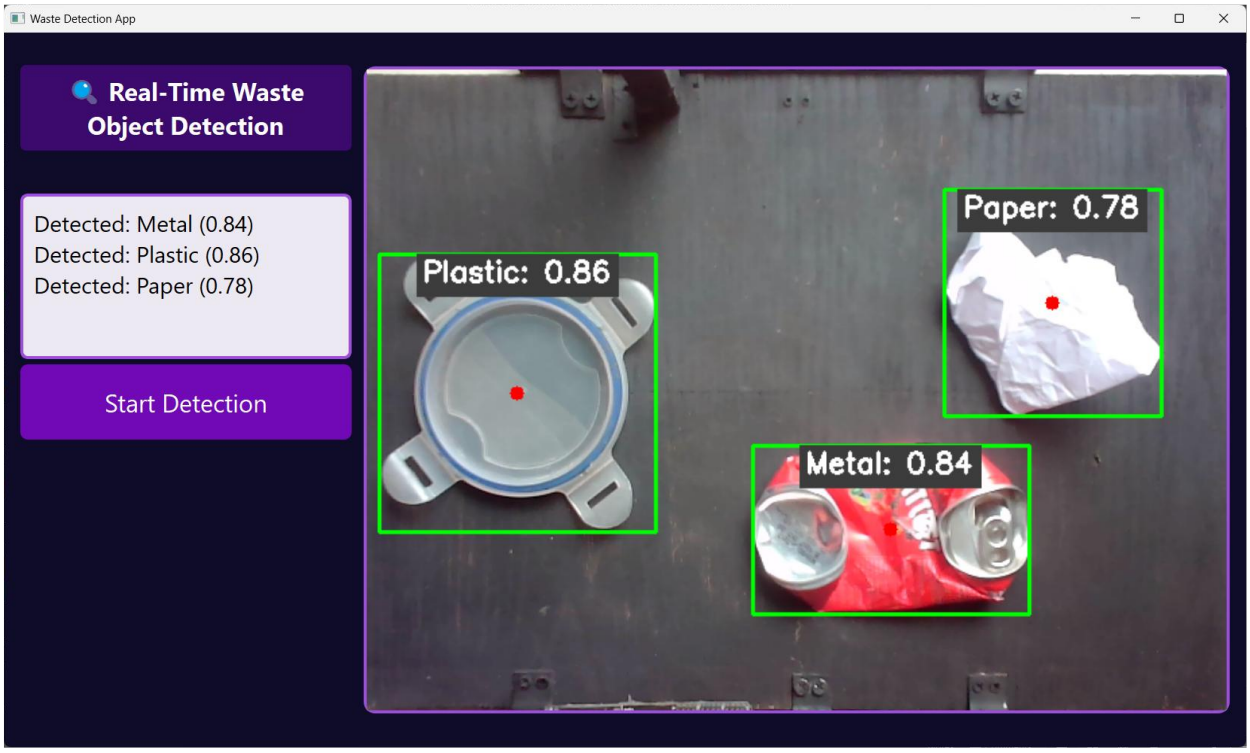


Fig. 17. GUI app for real-time detection and localization using robotic arm.

5. Discussions

A comprehensive review of existing AI-based waste detection and classification systems reveals several notable contributions, each addressing distinct aspects of the problem, ranging from classification accuracy to real-time processing and explainability, as shown in Table 2. However, holistic system balancing performance, explainability, external validation, and hardware deployment remain rare. The proposed RTDRNet-lite approach effectively fills this gap.

While several existing works have achieved high classification accuracy, none of these works focused on deployment feasibility or explainability of the classification. For example, the models presented in Ahmed et al., (2023), Gunaseelan et al., (2023), and Alsubaei et al., (2022) achieved a classification accuracy as high as 98.95% using an optimized ResNet50V2, RefineDet, and a modified ResNeXt. While all of these studies record impressive accuracy, these studies are limited

to classification tasks, often tested only in isolated digital environments, without external validation or hardware testing. In contrast, the RTDRNet-lite model achieves a superior mAP@50 of 97%. In addition, the RTDRNet-lite model was also externally validated on different platforms. Moreover, it was deployed in a real-time robotic system, thereby proving its deployability for future applications. Other works, such as Majchrowska et al., (2021) and Sayem et al. (2024), adopt object detection strategies alongside classification. For example, Majchrowska et al., (2021) used EfficientDet-D2 for detection and classification, and the two models reached 70% AP and 75% accuracy, respectively. Sayem et al. (2024) obtained a 63% mAP50 with the GELAN-E detector. However, in all of these models, there has been some performance loss, as those systems were not operational in real time. Additionally, neither integrates explainability nor hardware validation, reducing their credibility for practical deployment. RTDRNet-lite outperforms both in detection performance and operational readiness through real-world deployment and interpretability features.

Given the selected context of semantic segmentation and cluttered waste environments, most works in the comparison table, for example, Sirimewan et al., (2024), Prasad et al., (2025), Kiyokawa et al., (2021), are related to segmentation models, such as DeepLab-v3+, U-Net, and ShARP-WasteSeg, which focus on boundary accuracy. While beneficial for more detailed and fine-grained analysis, these methods are not designed to prioritize speed and hardware efficiency. In contrast, RTDRNet-lite simplifies RT-DETR with a C2F block and reduces the number of parameters to find the necessary practical balance between high accuracy and low-latency inference, which is critical for robotics. The role of explainable AI becomes more beneficial and essential when the sensitivity and critical nature of applications, such as waste management, increase. Only a few works, Nahiduzzaman et al. (2025), Sayem et al. (2024), include some form of interpretability in the chosen field. RTDRNet-lite is superior because of its adoption of XAI support, increasing the model's transparency and enabling trust-based real-world decisions, which is an important competitive advantage for smart waste distribution systems.

Finally, RTDRNet-lite uses a hybrid training method that combines real-world waste image data with high-resolution synthetic data produced by stable diffusion (Realistic vision v5. 1). To the best of our knowledge, this is the first work in waste detection to utilize synthetic data generation for object detection, offering significant gains in visual diversity and generalizability, particularly in handling edge cases and complex object boundaries. Furthermore, based on object detection, RTDRNet-lite addresses a basic flaw of traditional classification methods presented in Nahiduzzaman et al. (2025), Hossen et al. (2024) , Ahmed et al. (2023), Sayem et al. (2024), Wang et al. (2021), Gunaseelan et al. (2023), Alsubaeiet al. (2022), Rahman et al. (2020), and Mookkaiah et al. (2022), which lack exact localization and real-time robotic incorporation. Such models are, therefore, not suitable for hardware implementation to process cluttered multi-object waste. On the other hand, RTDRNet-lite not only realizes accurate detection of multiple objects but can also be directly embedded into the robotic arms for precise waste localization in real time. Such combined contributions, synthetic dataset generation, object detection-based architecture, external validation, real-time performance, hardware deployment, and XAI support, are rarely considered together in related works.

Table 2.
A comparative analysis on the performance of the proposed model with the available models in literature

Ref.	Approach Used	Dataset	Accuracy/mAP	External Validation	Explainable AI	Real-time test
Majchrowska et al., (2021)	EfficientDet-D2 + EfficientNet-B2	Detect-Waste	70% AP (detection), 75% (classification)	✓	✗	✗
Sirimewan et al., (2024)	DeepLabv3+ and U-Net (ResNet-101, etc.)	CRD Skip Bin Dataset (430 images)	84% & 85% mAP	✗	✗	✗
Nahiduzzaman et al. (2025)	DP-CNN + En-ELM	TriCascade WasteImage	96% accuracy	✗	✓	✓
Hossen et al., (2024)	RWCNet	TrashNet (2,527 images)	95.01% accuracy	✗	✓	✗

Prasad et al., (2025)	ShARP-WasteSeg (CSP backbone)	Real-world CDW dataset	55.5% mAP	✗	✗	✗
Ahmed et al. (2023)	Custom CNN, ResNet50V2	Custom Garbage Dataset	88.52% & 98.95% accuracy	✗	✗	✗
Qiu et al (2022)	ETHSeg (ResNet-101-FPN)	X-ray Waste Dataset (5,038 images)	63.22% mAP50	✗	✗	✗
Sayem et al., (2024)	Dual-stream classifier & GELAN-E	Custom 10,406 image dataset & WaRP dataset	83.11% (classification), 63% mAP50 (detection)	✗	✓	✗
Wang et al. (2021)	MobileNetV3	17,073 (9 classes)	94.6% accuracy	✗	✗	✓
Gunaseelan et al. (2023)	Modified ResNeXt + ResNet-50	Custom hardware dataset	98.9% accuracy	✗	✗	✓
Alsubaei et al. (2022)	DLSODC-GWM(IRD (RefineDet) + FLNN)	Garbage Classification Dataset (Kaggle)	98.61% accuracy	✗	✗	✗
Kiyokawa et al., (2021)	DeepLabv3+	Construction waste dataset (5,366 images)	56% mIoU	✗	✗	✗
Rahman et al. (2020)	CNN + IoT smart bin	Binary Waste Dataset	95.31% accuracy	✗	✗	✓
Mookkaiah et al. (2022)	ResNet V2-based CNN (transfer learning)	MSW Dataset	87.99% accuracy	✗	✗	✓
Iqbal et al. (2022)	YOLOv4 (CSPDarkNet_tiny)	Remondis Contamination Dataset (RCD)	63% mAP	✗	✗	✓
Sallang et al. (2021)	SSD MobileNetV2 Quantized	Urban waste dataset	92.16% mAP	✗	✗	✓
Proposed	RTDRNet-lite (Modified RT_DETR with C2F module and reduced dimension	Custom Dataset (Real+Synthetic) 4 Class	97% mAP	✓	✓	✓

6. Conclusions

This work presents a fully integrated AI solution that can be readily deployed to address the problems of intelligent waste detection and robotic sorting at the system level, which represents long-standing challenges for any existing waste management system. The framework in question, underpinned by the lightweight real-time detection model RTDRNet-lite, provides an object detection sub-framework that is both lightweight and highly accurate and is designed to be deployed in real-time in resource-constrained environments. The revision of the architecture, specifically utilizing C2F blocks instead of RepC3 modules and diminishing the depth and channel sizes of the RT-DETR backbone, led to a substantial 58% decrease in parameter quantity, which does not impede performance as the mAP@50 parameter reaches 97%. This performance/efficiency combination demonstrated by the architecture is paramount for further industry and urban deployment in the future. Furthermore, this work presents a novel method of data generation and ground truth using a combination of real-world images and high-quality synthetic waste images produced by the Stable Diffusion-based Realistic Vision v5.1 model. The hybrid dataset not only addresses the class imbalance issue, but also facilitates generalization across occlusions, deformations, and illumination variations, the aspects not accounted for in the current leading RTDRNet-lite dataset. The semi-automatic annotation with a pre-trained detector significantly accelerates the process of enhancing the dataset while preserving high-quality annotations. In addition to the AI component, a 4-DOF robotic arm was developed and integrated with the trained model to perform automated object localization and alignment via inverse kinematics. Pixel-to-distance calibration and trigonometric computation enabled accurate joint angle estimations, translating visual detections into physical movements. The custom GUI application ensured thorough testing of the system under real-time conditions and accelerated frame-by-frame inference and robot coordination. To summarize, this paper bridged the gap

between AI algorithm development and its physical deployment by introducing an interpretable and hardware-compatible solution to waste sorting.

Some drawbacks still exist in the existing framework despite the positive outcomes it brings. Although this is improved by synthetic data generation, the training data is only represented by four key waste types, which could prove to have minimal performance in the real-world application of highly mixed or less frequently sampled waste. Also, the robotic arm application is limited to orienting objects without physically gripping them or interacting with it at the moment and performance is limited to a simple servo motor due to its prototype nature. Future work will explore expanding the model's capabilities to include additional waste types such as e-waste, organics, and hazardous materials, thereby increasing dataset diversity and enhancing applicability in more complex domains. Addressing object overlap and occlusion through advanced techniques such as instance segmentation or refined attention mechanisms represents another key direction for improving detection robustness in cluttered scenes. On the hardware side, the integration of high-precision servo motors will be pursued to enable smoother, more reliable arm control during field deployments, allowing the system to be evaluated under realistic operational conditions.

Reference

Ahmed, M.I.B. *et al.* (2023) 'Deep Learning approach to recyclable Products Classification: towards sustainable waste management,' *Sustainability*, 15(14), p. 11138. https://doi.org/10.3390/su151411138.

Alsabt, R. *et al.* (2024) 'Optimizing waste management strategies through artificial intelligence and machine learning - An economic and environmental impact study,' *Cleaner Waste Systems*, 8, p. 100158. https://doi.org/10.1016/j.clwas.2024.100158.

Alsubaei, F.S., Al-Wesabi, F.N. and Hilal, A.M. (2022) 'Deep Learning-Based Small Object Detection and Classification model for garbage waste management in smart cities and IoT environment,' *Applied Sciences*, 12(5), p. 2281. https://doi.org/10.3390/app12052281.

Emenike, E.C. *et al.* (2023) 'From oceans to dinner plates: The impact of microplastics on human health,' *Heliyon*, 9(10), p. e20440. https://doi.org/10.1016/j.heliyon.2023.e20440.

Fang, B. *et al.* (2023) 'Artificial intelligence for waste management in smart cities: a review,' *Environmental Chemistry Letters*, 21(4), pp. 1959–1989. https://doi.org/10.1007/s10311-023-01604-3.

Gunaseelan, J., Sundaram, S. and Mariyappan, B. (2023) 'A design and implementation using an innovative Deep-Learning algorithm for garbage segregation,' *Sensors*, 23(18), p. 7963. https://doi.org/10.3390/s23187963.

Hasan, M.A., Ahmad, S. and Mohammed, T. (2021) 'Groundwater contamination by hazardous wastes,' *Arabian Journal for Science and Engineering*, 46(5), pp. 4191–4212. https://doi.org/10.1007/s13369-021-05452-7.

Hossen, Md.M. *et al.* (2024) 'A reliable and robust deep learning model for effective recyclable waste classification,' *IEEE Access*, 12, pp. 13809–13821. https://doi.org/10.1109/access.2024.3354774.

Iqbal, U. *et al.* (2022) 'Edge-Computing Video Analytics Solution for Automated Plastic-Bag Contamination Detection: A Case from Remondis,' *Sensors*, 22(20), p. 7821. https://doi.org/10.3390/s22207821.

Jain, K. and Shah, C. (2019) 'A Review : Sustainability from Waste,' *International Journal of Scientific Research in Science and Technology*, pp. 134–155. https://doi.org/10.32628/ijsrst196630.

Jaouhari, A.E. *et al.* (2024) 'Turning trash into treasure: Exploring the potential of AI in municipal waste management - An in-depth review and future prospects,' *Journal of*

Environmental Management, 373, p. 123658. https://doi.org/10.1016/j.jenvman.2024.123658.

Jerie, S. (2016) 'Occupational Risks Associated with Solid Waste Management in the Informal Sector of Gweru, Zimbabwe,' *Journal of Environmental and Public Health*, 2016, pp. 1–14. https://doi.org/10.1155/2016/9024160.

Kharola, S. *et al.* (2022) 'Barriers to organic waste management in a circular economy,' *Journal of Cleaner Production*, 362, p. 132282. https://doi.org/10.1016/j.jclepro.2022.132282.

Kiyokawa, T. *et al.* (2021) 'Robotic waste sorter with agile manipulation and quickly trainable detector,' *IEEE Access*, 9, pp. 124616–124631. https://doi.org/10.1109/access.2021.3110795.

Kshirsagar, P.R. *et al.* (2022) 'Artificial Intelligence-Based Robotic technique for reusable waste materials,' *Computational Intelligence and Neuroscience*, 2022, pp. 1–9. https://doi.org/10.1155/2022/2073482.

Lakhouit, A. (2025) 'Revolutionizing Urban Solid Waste Management with AI and IoT: A review of smart solutions for waste collection, sorting, and recycling,' *Results in Engineering*, 25, p. 104018. https://doi.org/10.1016/j.rineng.2025.104018.

Liu, J. *et al.* (2021) 'Garbage Collection and Sorting with a Mobile Manipulator using Deep Learning and Whole-Body Control,' *IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pp. 408–414. https://doi.org/10.1109/humanoids47582.2021.9555800.

Lu, W., Chen, J. and Xue, F. (2021) 'Using computer vision to recognize composition of construction waste mixtures: A semantic segmentation approach,' *Resources Conservation and Recycling*, 178, p. 106022. https://doi.org/10.1016/j.resconrec.2021.106022.

Lubongo, C., Daej, M. a. a. B. and Alexandridis, P. (2024) 'Recent developments in technology for sorting plastic for recycling: the emergence of artificial intelligence and the rise of the robots,' *Recycling*, 9(4), p. 59. https://doi.org/10.3390/recycling9040059.

Majchrowska, S. *et al.* (2021) 'Deep learning-based waste detection in natural and urban environments,' *Waste Management*, 138, pp. 274–284. https://doi.org/10.1016/j.wasman.2021.12.001.

Mookkaiah, S.S. *et al.* (2022) 'Design and development of smart Internet of Things–based solid waste management system using computer vision,' *Environmental Science and Pollution Research*, 29(43), pp. 64871–64885. https://doi.org/10.1007/s11356-022-20428-2.

Nahiduzzaman, Md. *et al.* (2025) 'An automated waste classification system using deep learning techniques: toward efficient waste recycling and environmental sustainability,' *Knowledge-Based Systems*, p. 113028. https://doi.org/10.1016/j.knosys.2025.113028.

Prasad, V. and Arashpour, M. (2025) 'ShARP-WasteSeg: A shape-aware approach to real-time segmentation of recyclables from cluttered construction and demolition waste,' *Waste Management*, 195, pp. 231–239. https://doi.org/10.1016/j.wasman.2025.02.006.

Qiu, L. *et al.* (2022) 'ETHSEG: an amodel instance segmentation network and a real-world dataset for X-Ray waste inspection,' *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2273–2282. https://doi.org/10.1109/cvpr52688.2022.00232.

Rahman, Md.W. *et al.* (2020) 'Intelligent waste management system using deep learning with IoT,' *Journal of King Saud University - Computer and Information Sciences*, 34(5), pp. 2072–2087. https://doi.org/10.1016/j.jksuci.2020.08.016.

RealWaste Image Classification (2024). https://www.kaggle.com/datasets/joebeachcapital/realwaste.

Sallang, N.C.A. *et al.* (2021) 'A CNN-Based smart waste management system using TensorFlow Lite and LORA-GPS Shield in Internet of Things environment,' *IEEE Access*, 9, pp. 153560–153574. https://doi.org/10.1109/access.2021.3128314.

Sayem, F.R. *et al.* (2024) 'Enhancing waste sorting and recycling efficiency: robust deep learning-based approach for classification and detection,' *Neural Computing and Applications* [Preprint]. https://doi.org/10.1007/s00521-024-10855-2.

SG161222/Realistic_Vision_V5.1_NoVAE · Hugging face (2001). https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE.

Sheriff, S.S. *et al.* (2025) 'A comprehensive review on exposure to toxins and health risks from plastic waste: Challenges, mitigation measures, and policy interventions,' *Waste Management Bulletin*, p. 100204. https://doi.org/10.1016/j.wmb.2025.100204.

Sirimewan, D. *et al.* (2024) 'Deep learning-based models for environmental management: Recognizing construction, renovation, and demolition waste in-the-wild,' *Journal of Environmental Management*, 351, p. 119908. https://doi.org/10.1016/j.jenvman.2023.119908.

Torres, Y., Nadeau, S. and Landau, K. (2021) 'Classification and Quantification of Human error in Manufacturing: A case study in complex manual assembly,' *Applied Sciences*, 11(2), p. 749. https://doi.org/10.3390/app11020749.

Trends in solid waste management (no date). https://datatopics.worldbank.org/what-a-waste/trends_in_solid_waste_management.html.

Wang, C. *et al.* (2021) 'A smart municipal waste management system based on deep-learning and Internet of Things,' *Waste Management*, 135, pp. 20–29. https://doi.org/10.1016/j.wasman.2021.08.028.

WARP - Waste Recycling Plant Dataset (2023). https://www.kaggle.com/datasets/parohod/warp-waste-recycling-plant-dataset.

Ying, X. (2019) 'An Overview of Overfitting and its Solutions,' *Journal of Physics Conference Series*, 1168, p. 022022. https://doi.org/10.1088/1742-6596/1168/2/022022.

953 Zhang, Q. *et al.* (2021) 'Recyclable waste image recognition based on deep learning,' *Resources*
954 *Conservation and Recycling*, 171, p. 105636.
955 https://doi.org/10.1016/j.resconrec.2021.105636.
956