INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

OPEN ACCESS
Received: 21/01/2026
Accepted: 13/02/2026
Published: 23/02/2026

Citation: Carol I (2026) A Context-Preserving Vision–IoT Framework for Automated Solid Waste Segregation Using Hierarchical Token Decomposition. Indian Journal of Science and Technology 19(7): 429-442. https://doi.org/10.17485/IJST/v19i7.91

* **Corresponding author**.
carol_it2@mail.sjctni.edu

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

A Context-Preserving Vision–IoT Framework for Automated Solid Waste Segregation Using Hierarchical Token Decomposition

I Carol[1]*

1 Assistant Professor, Department of Information Technology, St. Joseph's College (Autonomous), Trichy, India

Abstract

Objectives: This study aims at creating a smart vision-IoT system. It is to propose an automated waste classification through preserving material, geometric and spatial context clearly. To resolve ambiguity and misclassification that can be seen in current waste segregation frameworks. **Method:** It proposes a Context-Preserving Vision-IoT (CPV-IoT) framework, which includes a Hierarchy Tokens Decomposition (HTD) mechanism in the encoding of surface, geometric and contextual semantics before inference. An attention-based transformer is used to carry out contextual reasoning and then lightweight classification is done. The model is tested with the help of simulation-based experiments based on a publicly available waste image dataset, which simulates the work of IoT-enabled smart bins. **Findings:** The simulation evidence shows that the proposed framework has accuracy of 96.87% and F1-score of 96.41%. The results are higher than classical machine learning models and more recent deep learning baselines. Ablation analysis establishes that both hierarchical tokenization and contextual reasoning lead to performance improvement. **Novelty:** The originality of this work is that hierarchical semantic tokenization of waste images is presented, which makes it possible to represent the visual representation in a structured way and then reason about it. This method will contribute to the existing study on waste classification by showing that context-preserving abstraction can dramatically increase the reliability of classification in simulated IoT-based waste management systems.

Keywords: Waste segregation; Vision transformer; Hierarchical tokenization; IoT smart bins; Sustainable waste management

1 Introduction

Solid waste management has been a constant problem in urban areas in rapid development, where the initial improper sorting on site has a significant effect on the recycling and subsequent result in higher rates of environmental impact [1]. Segregation is time-consuming and involves manual segregation that is prone to errors and it subjects the workers to health risks [2]. Thus, automated waste sorting with the application of artificial intelligence attracts more interest [3].

Recent vision-based models make use of convolutional or transformer-based models which are trained using raw images [4]. Even though the techniques are highly accurate on a controlled environment classification, they tend to be inaccurate when applied in real waste disposal situations because of visual similarity among materials [5], partial occlusion [6], mixed or deformed waste items [7] and absence of contextual reasoning [8].

Recent research proves that deep learning-based models are by far more effective than other traditional and sensor-only methods, but there are still issues in contextual understanding, scalability and edge deployment.

Nahiduzzaman *et al.* [9] in the field of deep learning-based waste classification put forward a multi-stage structure that integrates depth-wise separable CNNs with an ensemble extreme learning machine classifier and demonstrated encouraging multi-level classification on a big composite data of more than 35,000 images. Their system focuses on lightweight feature extraction and explainability but does not run in the end-to-end contextual reasoning, but in a staged classification hierarchy.

The model of automated waste sorting designed by Ahmad et al. can classify various waste types based on the deep learning method and identify the lack of data balance and reduced practicality in the real world as the key challenges [10]. Their work points out the necessity of strong representations that can generalize between appearances of waste that vary. Complementary to that, Dipo et al. investigated real-time waste detection using deep architectures based on the YOLO, indicating that changes in the environment and a scenario with multiple objects are further issues in autonomous classification [11].

Some of the studies have incorporated AI in combination with IoT to enable intelligent waste management in the urban areas. Alourani et al. introduced an intelligent IoT system with the combination of sensors and AI to automate the process of sorting and operational monitoring, which exhibits the possibility of integrating classification and bin status analytics at the system level [12]. Similarly, studies by Adeoye explore IoT sensors relaying bin sensor data and an AI-based classification to optimize routes and collect timetables but this study is more of sensor fusion to logistics than fine-grained vision semantics [13].

Fotovvatikhah et al. review the literature on AI-based waste classification and note that even though classical machine learning and CNN-based solutions have been improved, the issues of dataset imbalance, model generalization and ability to encode contextual information remain, particularly in the real-world implementations [14].

Recently, the DWaste platform discusses resource-constrained, more sustainable AI sorting in mobile devices and edge computing devices in real-time [15]. The other most advanced work presents hybrid approaches to deep learning-machine learning, which attains almost optimal accuracy in a variety of publicly available data, which also demonstrates the possibility of structured representation schemes in waste classification [16], [17].

Existing AI-based waste segregation systems exhibit three fundamental limitations:

1. **Pixel-centric learning:** Raw pixel representations do not explicitly encode material-level semantics.
2. **Context loss:** Visual classifiers often ignore disposal context such as object boundaries, surface continuity and relative geometry.
3. **Edge instability:** High-capacity models suffer from inconsistent inference when deployed on resource-constrained IoT devices.

There is a lack of frameworks that explicitly model visual structure and context prior to classification, like how signal-level decomposition is used in time-series IoT analytics. The CPV-IoT framework presented in this paper fills these gaps by proposing hierarchical semantic tokenization and structured reasoning of a single visionIoT architecture. It facilitates robust and scalable waste segregation to be deployed in real time and embedded.

2 Methodology

Context-Preserving Vision -IoT (CPV-IoT) framework aims to bring about reliable and scalable automated waste segregation through process by model visual structure, material context and spatial relationships prior to classification. The framework adheres to a layered analytical structure, based on recent interpretable IoT and IIoT architectures, in which each layer is a well-defined transformation, which makes the architecture robust, explainable and stable to deployment.

The main concept of CPV-IoT is that waste images are not to be seen as pixel grids without hierarchy. It is broken down into the hierarchically significant visual tokens that retain the contextual and material semantics in the entire inference pipeline.

2.1 Framework Overview

The CPV-IoT framework consists of five sequential processing layers given below:

1. Visual Acquisition Layer
2. Hierarchical Token Decomposition Layer
3. Contextual Reasoning Layer
4. Waste Category Inference Layer
5. IoT Integration and Monitoring Layer

Each layer performs a deterministic transformation that ensures the consistency across edge deployments. It maintains semantic continuity from raw image capture to final waste classification. Figure 1 depicts the workflow of the proposed framework.

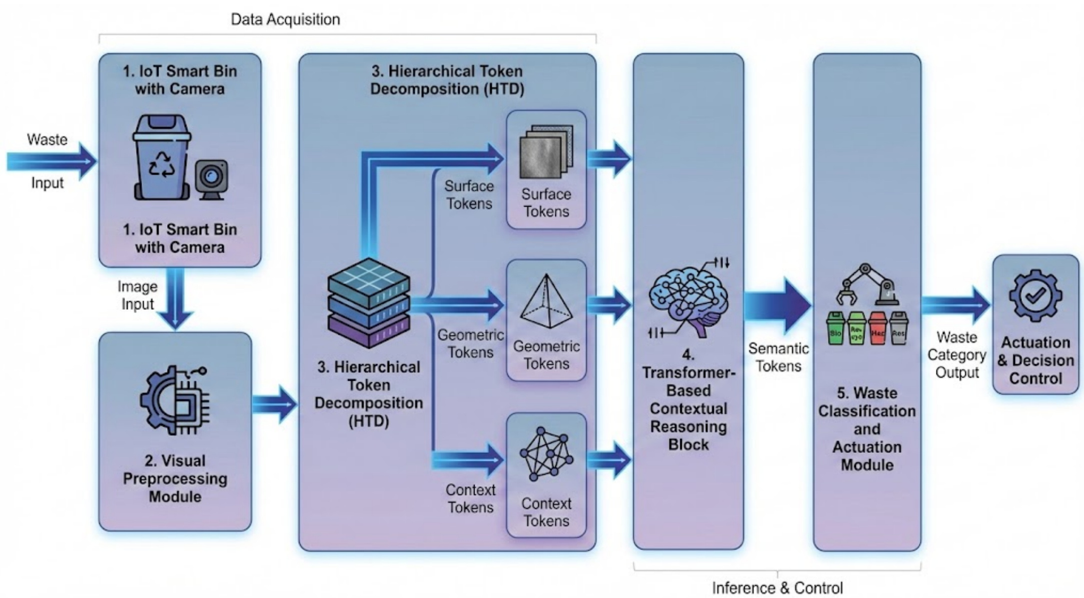

Fig 1. Workflow of the proposed framework

2.2 Visual Acquisition Layer

The Visual Acquisition Layer is the initial operation stage of CPV-IoT framework as it involves the acquisition of waste images in a real-life disposal setting. This layer helps in maintaining uniformity and standardized input of all the subsequent analytical processes as well as deployment-ready visual data to minimize the changes that are introduced by heterogeneous hardware and environmental factors.

The images of waste are obtained via RGB camera modules, which are mounted on smart bins with the IoT. Image capture is a triggered process and is activated with the identification of a disposal action by either a proximity sensor or lid-movement sensor. This is an event-based acquisition strategy that minimizes redundant data capture and also makes sure every waste is recorded when it is disposed.

Let the captured image be represented as:

$$I \in R^{H \times W \times C} \tag{1}$$

where (H) and (W) denote spatial resolution and (C) denotes the number of color channels. To ensure compatibility across different bin installations, all images are resized to a fixed resolution and transformed into a standardized color space.

The following preprocessing operations are applied sequentially:

- Spatial normalization: resizing and aspect-ratio alignment to a predefined resolution.
- Photometric normalization: intensity scaling and contrast stabilization to mitigate lighting variation caused by outdoor and indoor deployment conditions.
- Background intensity smoothing: suppression of abrupt illumination gradients caused by shadows or reflective waste surfaces.

Such operations are not semantically interpreted or extracted or classified and are limited to normalizing low-level signals. The design makes sure that Visual Acquisition Layer is lightweight and computationally efficient and can be used on embedded hardware without making it hard to use.

The output of this layer is a standardized image:

$$\tilde{I} = \mathcal{N}(I) \tag{2}$$

where $\mathcal{N}(\cdot)$ denotes the normalization and standardization pipeline.

The CPV-IoT framework provides a high level of visual capture and normalization separation in a specific layer, leading to the reliable and reproducible inference in a wide variety of deployment settings.

2.3 Hierarchical Token Decomposition Layer

The most important representational stage of the CPV-IoT framework is the Hierarchical Token Decomposition (HTD) Layer. The layer is meant to convert standardized waste images to structured semantic tokens without destroying material characteristics, geometric structure and spatial context before reasoning and classification. The key rationale of this layer is to break the constraints of pixel-based representations with the addition of multi-granular visual abstraction.

Given the normalized image output from the Visual Acquisition Layer:

$$\tilde{I} \in R^{H \times W \times C} \tag{3}$$

the HTD layer performs a sequence of deterministic transformations that decompose $\tilde{I}$ into hierarchically organized tokens. Figure 2 provides the Hierarchical token decomposition workflow.

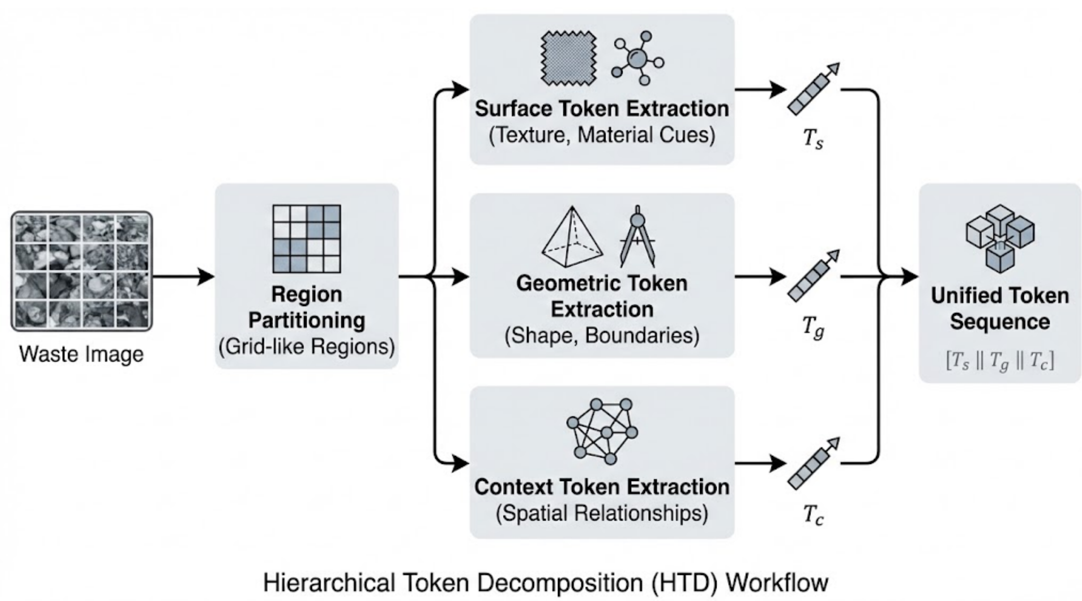

Fig 2. Hierarchical token decomposition workflow

2.3.1. Multi-Scale Region Partitioning

The first step in HTD is the partitioning of the image into a set of fixed, multi-scale regions. This process enables the capture of both local details and global object structures.

$$\mathcal{R} = R_1, R_2, \ldots, R_K \tag{4}$$

Each region R_k corresponds to a spatial sub-area of the image, generated using predefined grid resolutions. Multi-scale partitioning ensures robustness against variations in object size, deformation and partial visibility.

2.3.2. Surface Token Construction

Surface tokens store visual marks at the material level, including texture, reflectance and surface roughness, that are necessary to the differentiation of waste of similar shape with different composition.

For each region R_k, surface descriptors are extracted using convolutional projections and statistical texture measures, forming surface tokens:

$$T_s = \phi_s(R_k) \mid R_k \in \mathcal{R} \tag{5}$$

These tokens capture fine-grained appearance information and provide strong discriminatory power for recyclable and hazardous materials.

2.3.3. Geometric Token Construction

In geometric tokens, structural and shape attributes are reflected. It entails contours, edges and region continuity. This is in the form of a component that facilitates accurate interpretation of a deformed or partially covered waste.

$$T_g = \phi_g(R_k) \mid R_k \in \mathcal{R} \tag{6}$$

The framework removes geometric consistency errors by encoding across regions and eliminating errors due to crushed or irregularly shaped waste items.

2.3.4. Context Token Construction

Context tokens represent spatial relationships and patterns of co-occurrence between the regions. This model is represented in the manner of the relationship between various visual elements in the same instance of disposal.

$$T_c = \phi_c\left(R_i, R_j\right) \mid R_i, R_j \in \mathcal{R}, i \neq j \tag{7}$$

Contextual encoding facilitates the separation between the isolated waste items and mixed or layered waste-arrays as is common in the disposal settings found in the public.

2.3.5. Token Aggregation and Positional Encoding

The complete hierarchical token set is formed by concatenating the three token categories:

$$T = \left[T_s \| T_g \| T_c\right] \tag{8}$$

Positional encoding is applied to retain spatial ordering and regional correspondence across tokens. This ensures that spatial relationships remain interpretable during subsequent reasoning.

2.4 Contextual Reasoning Layer

Contextual Reasoning Layer performs the task of modeling semantic dependencies between the hierarchical tokens produced by the HTD layer. This layer converts the structured representation of tokens into an embedding that is coherent across the entire globe and represents interactions among surface properties, geometrical structure and the spatial context. This layer is mainly aimed at facilitating the comprehension of waste scenes in a holistic manner, opposed to the individual region analysis.

Let the aggregated hierarchical token sequence be denoted as:

$$T = \left[T_s \| T_g \| T_c\right] \tag{9}$$

where $(T_s), (T_g)$ and (T_c) represent surface, geometric and context tokens, respectively.

2.4.1. Attention-Based Token Interaction Modeling

A Contextual Reasoning Layer uses a transformer based reasoning mechanism to model inter-token relationships. The tokens of (T) are mapped to query, key and value space and with self-attention interaction, the modeling is pairwise.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \tag{10}$$

where (d) denotes the embedding dimension.

The mechanism enables the framework to pay selective attention to semantically meaningful interaction of tokens, including correspondence between surface texture and spatial positioning and ignore irrelevant visual noise.

2.4.2. Cross-Token Context Integration

This layer uses heterogeneous semantic tokens as opposed to the conventional vision transformers that use homogeneous patch embeddings. A cross-token attention allows one to combine complementary information across the token categories:

- Surface-geometry interactions refine material shape consistency,
- Geometry-context interactions capture object composition and spatial grouping,
- Surface-context interactions disambiguate visually similar materials based on placement.

These interactions result in enriched token representations that reflect both local properties and global scene structure.

2.4.3. Contextual Embedding Generation

Following attention-based interaction, the refined token representations are aggregated to form a unified contextual embedding: $Z = \text{Transformer}(T)$. This embedding encodes comprehensive information about the waste instance, capturing appearance, structure and spatial relationships in a compact form suitable for classification.

2.5 Waste Category Inference Layer

The Waste Category Inference Layer translates high-level contextual knowledge into practical decisions which has a direct control over waste routing processes. This layer reduces inter-class confusion by working on structured contextual embeddings instead of raw visual features and enables safe and real-time waste collection in smart bin deployments.

It is the last semantic processing step of the CPV-IoT framework and it preconditions the physical actuation and monitoring integration of the output.

Let the contextual embedding obtained from the previous layer be denoted as:

$$Z \in R^d \tag{11}$$

where (d) represents the dimensionality of the contextual feature space encoding material, geometric and spatial semantics.

2.5.1. Classification Mapping

The inference layer employs a fully connected linear projection followed by a softmax normalization to compute class-wise posterior probabilities:

$$\hat{y} = \text{Softmax}(WZ + b) \tag{12}$$

where $W \in R^{|\Omega| \times d}$ and $b \in R^{|\Omega|}$ denote the learnable weight matrix and bias vector, respectively and Ω represents the waste category set.

The output vector $\hat{y}$ corresponds to the probability distribution over the following waste classes: Biodegradable, Recyclable, Hazardous and Residual wastes. This formulation ensures mutual exclusivity among categories, which is essential for physical segregation systems.

2.5.2. Confidence Estimation and Decision Reliability

A confidence score is derived from the maximum posterior probability:

$$p = \max(\hat{y}) \tag{13}$$

This confidence score serves as a reliability indicator for downstream actuation. Low-confidence predictions are treated conservatively to prevent contamination of recyclable or hazardous waste streams.

2.5.3. Lightweight Design for Edge Deployment

The inference layer has a small depth and low number of parameters to enable the efficient deployment of edges. This architecture option decreases the inference latency and minimizes the memory used in-vitro. It can also provide uniform performance when it comes to heterogeneous IoT devices with different computational abilities. The contextual reasoning is addressed on lower levels of the framework and the last layer of classification works as a stable and compact decision boundary. Such a separation makes sure that the representational complexity is not overloaded on top of the edge-level inference process.

2.6 IoT Integration and Monitoring Layer

This layer is the last operational segment of the CPV-IoT architecture and is charged with the responsibility of converting the results of classification to physical actuation, system-level logging and analytical monitoring. This layer allows connecting the gap between visual intelligence and actual waste management process so that automated decision-making can lead to credible segregation and practical insights. Figure 3 depicts the workflow of IoT integration and monitoring layer.

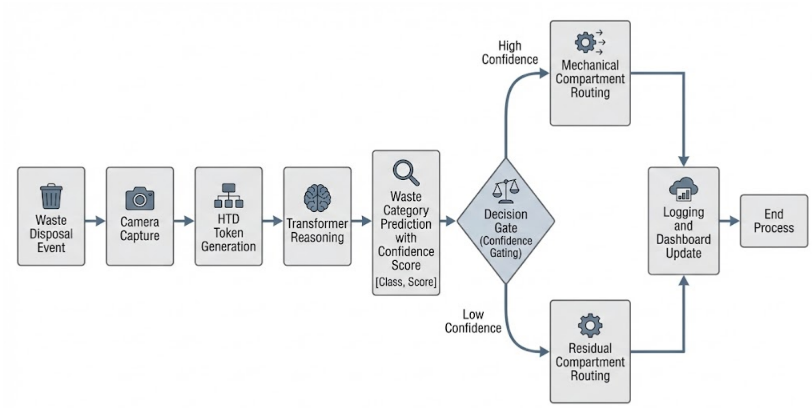

Fig 3. Workflow of IoT integration and monitoring layer

2.6.1. Smart Bin Actuation and Control

The waste category predicted by the Waste Category Inference Layer is transmitted to the embedded control unit of the smart bin. Based on the predicted class label, the control unit triggers the corresponding mechanical actuation mechanism, routing the disposed item to the appropriate compartment. Let the predicted class be denoted as:

$$c = \arg\max(\hat{y}) \tag{14}$$

The actuation logic follows a predefined compartment–class mapping, ensuring deterministic and repeatable segregation behavior. When the associated confidence score (p) falls below a predefined threshold, the system routes the waste item to a residual compartment to avoid cross-contamination of recyclable or hazardous waste streams.

2.6.2. Event Logging and Metadata Capture

A structured log in the system consists of each of the disposal events. The bin identifier (bin_id) and the time (t) of disposal are included in the log. It contains the estimated classification (c) of waste and the confidence measure. Additional sensor data such as proximity measurements (p) is added in case of its availability. This organized logging system allows traceability, performance audit and long term analytics in a process of smart bin deployments.

Formally, a log record is represented as:

$$l = (\mathrm{bin\_id}, t, c, p, s) \tag{15}$$

where (s) denotes auxiliary sensor information.

2.6.3. Communication and Data Transmission

Log records are transmitted from individual smart bins to a centralized monitoring server using lightweight IoT communication protocols. In the absence of network connectivity, records are temporarily cached locally and synchronized once connectivity is restored. This communication strategy ensures resilience and continuity of data collection across distributed deployments.

2.6.4. Monitoring Dashboard and Analytics

The aggregated log data is displayed at the monitoring level in the form of a centralized dashboard interface. The dashboard shows the distribution of waste in or in each category according to bins and location. It also displays bin level fill-level trends and statistics and bin-level usage. The patterns of spatial and temporal segregations are mapped to determine the trend of behavior and inefficiencies in the operations. These analytics can assist in informed operation planning, waste collection optimization scheduling and strategic policy making of municipal waste management authorities.

2.6.5. Separation of Concerns and System Scalability

The IoT Integration and Monitoring Layer is deliberately decoupled from the perception and reasoning components of the framework. This separation ensures that enhancements in monitoring, reporting, or communication infrastructure do not affect the core vision-based inference pipeline.

By maintaining clear boundaries between intelligence, actuation and monitoring, the CPV-IoT framework achieves scalability, maintainability and suitability for large-scale urban deployment.

2.7 Proposed Algorithm

The overview of the training and simulated deployment of the CPV-IoT framework is presented in Algorithm 1. The algorithm starts with the division of dataset into training, validation and testing sets. Computation of class weights is done to deal with imbalance. AdamW optimizer is used to initialize and optimize the proposed model parameters.

While training, every batch is augmented and preprocessed. There is partitioning of images into multi-scale regions. Positional encoding is used to extract surface, geometric and contextual tokens and combine them. The transformer produces contextual embeddings which is inputted into the classification head. Weighted cross entropy loss is calculated and backpropagated. Validation F1-score is used to choose the best model. The test set is evaluated at the final stage.

Once the training is done the inferred learned parameters are applied to simulated runtime inference. Each incoming image is preprocessed and corrected in terms of quality when necessary. Hierarchical tokens are created and sent across the transformer and the classifier. The outcomes of predicting the class and confidence score are calculated. In case the confidence falls below the threshold then the waste is sent to the residual compartment. Otherwise it is mapped into the category. Every event is recorded together with timestamp and sensor records. The algorithm produces the trained model and authenticate the stream of logs.

Algorithm 1: CPV-IoT with HTD (Train + Simulated Runtime)

```
Input: D={(Ii,yi)}i=1..N , E, bs, lr, τ, stream {(It,st,t)}
Output: Θ*, logs L
(Dtr,Dval,Dte) ← Split(D,0.70,0.15,0.15); wc ← ClassWeights(Dtr)
Θ ← Init(); opt ← AdamW(Θ,lr); best ← −∞
for e ← 1..E do
   for B ← Batches(Dtr,bs) do
     T ← []; Y ← []
     for (I,y) ∈ B do
       I ← Preprocess(Augment(I)); R ← MultiScalePartition(I)
         T ← T ∪ {PosEnc(Concat(SurfaceTokenize(R),GeometryTokenize(R),ContextTokenize(R)))}; Y ← Y ∪ {y}
     end for
     P ← Softmax(Head(Transformer(T,Θ)),Θ); loss ← WCE(P,Y,wc)
     opt.zero_grad(); Backprop(loss); opt.step()
   end for
   f1 ← F1(Eval(Θ,Dval)); if f1 > best then best ← f1; Θ* ← Copy(Θ) end if
end for
Report(Eval(Θ*,Dte))
L ← ∅
for each (It,st,t) in stream do
   It ← Preprocess(QualityCorrect(It)); R ← MultiScalePartition(It)
   Tt ← PosEnc(Concat(SurfaceTokenize(R),GeometryTokenize(R),ContextTokenize(R)))
   Pt ← Softmax(Head(Transformer(Tt,Θ*),Θ*),Θ*); ct ← ArgMax(Pt); pt ← Max(Pt)
   if pt < τ then route ← RES else route ← Map(ct) end if
   Actuate(route)
   L ← L ∪ {(bin_id,t,ct,pt,route,st)}
end for
return Θ*, L
```

2.8 IoT-Enabled Smart Bin Integration: A Theoretical and Simulation-Based Model

The smart bin is the deployment platform of the IoT-enabled CPV-IoT framework that allows the visual intelligence translation into automated waste sorting and useful monitoring information. Every smart bin will be built as a self-sufficient cyber-physical system that consists of sensing, computing, actuation and communication devices to enable real-time functionality within a communal waste collection setting.

2.8.1. Smart Bin Hardware Architecture

Each smart bin integrates the following core components:

The waste intake point has an RGB camera, which records the image during the disposal. The picture is taken to ensure that the area covered by the camera is as low as possible and the object of waste is visible as much as possible prior to striking the internal compartments. The fixed focal length and resolution values are employed to make sure that the image quality remains consistent throughout the deployments. The camera is triggered by a proximity (lid movement sensor) to minimize redundant image capture and computation requirements. Figure 4 presents the prototype model of IoT-enabled smart bin.

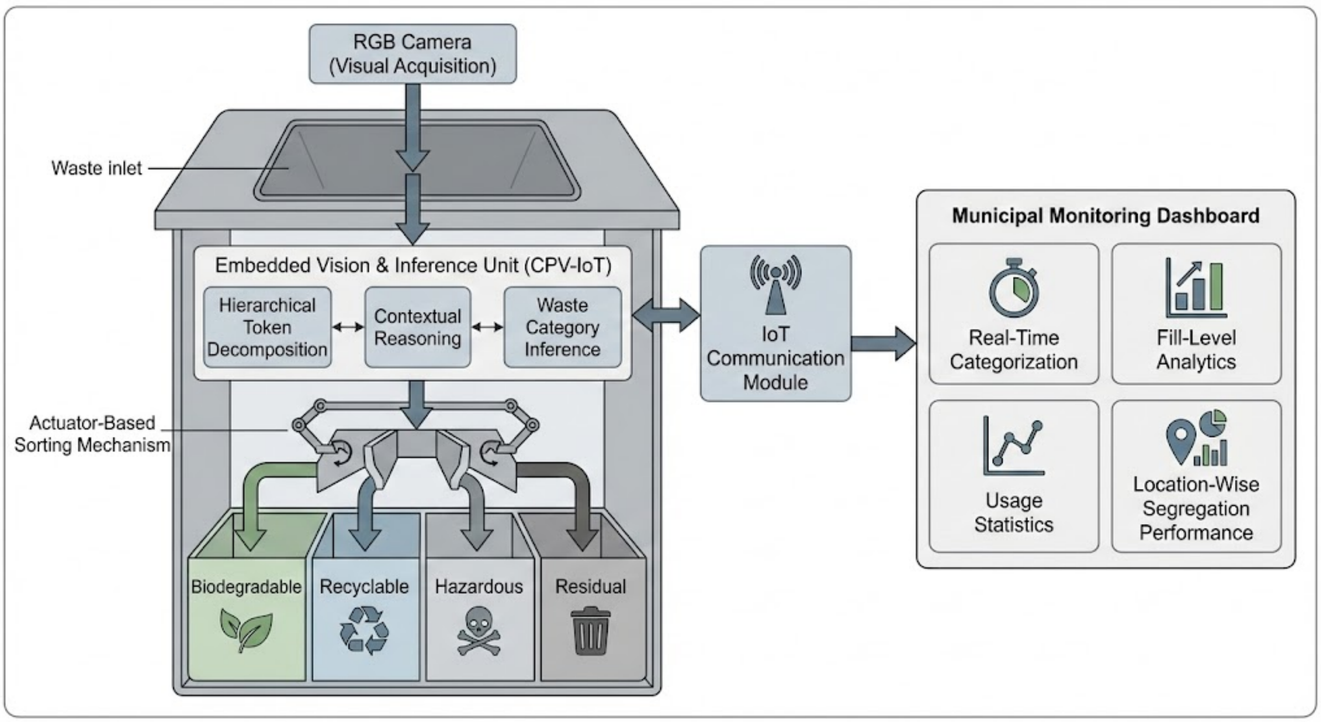

Fig 4. Prototype model of IoT-enabled smart bin

2.8.1.1. Embedded Inference Unit. The inference unit inside the smart bin performs the full CPV-IoT pipeline inside the smart bin. It does visual preprocessing, that is, standardizing the images taken and then analysis is done. The unit then subjects the hierarchical token decomposition to derive surface, geometric and contextual representations. The contextual reasoning makes use of these tokens to produce structured embeddings. Finally, it checks the inference of the waste category and gives a score as a result of actuation confidence. The unified architecture allows making low latency decisions without the need of a continuous connection to the cloud. The unit is energy efficient and it can maintain stable operation in the outdoor conditions of the environment.

2.8.1.2. Actuator-Based Sorting Mechanism. The smart bin has built in a mechanical sorting system that is operated by the inbuilt unit. The actuators dispose of the item into the respective compartment depending on the type of waste that is estimated. The sorting mechanism is predefined category with compartment mapping. It has safety interlocks to avoid mechanical conflicts. Logic based on confidence guarantees that when there is uncertainty in prediction, residual compartment directs waste.

2.8.1.3. Network Interface and Communication Module.
This module allows the smart bins in the city to interact with the municipal monitoring systems. It enables two way communication between the transmission of disposal event logs and the receipt of configuration updates. Another feature of the module is the synchronization of system status information among the platforms that are related. A minimum and robust communication protocol is undertaken to guarantee sure data communication. The design enables its further usage when connectivity is intermittent and enables delayed synchronization when connectivity is minimal.

2.8.2. Functional Capabilities Enabled by CPV-IoT

The integration of the CPV-IoT framework within smart bins enables several key operational capabilities:

2.8.2.1. Real-Time Waste Categorization.
Segregation of wastes is done at the disposal point hence they can be segregated instantly and automatically. The end to end processing pipeline is running on a tight latency budget that provides smooth interaction with users without any obvious delay.

2.8.2.2. Fill-Level and Usage Analytics.
Smart bins check the level of compartments filling continuously and register the frequency of disposal. These measurements give organized data concerning the pattern of waste production in different places. The gathered statistics also show the rate of bin utilization and the periods of the highest usage. The insight of this analytical approach justifies efficient waste collection scheduling. Further, the analysis enhances resource allocation efficiency in municipal waste management.

2.8.2.3. Location-Wise Segregation Performance Monitoring.
The system combines data that is gathered by several smart bins to allow location-specific analysis of the performance of segregation. Such a collection enables local government to assess how recycling is done in various neighborhoods. It can also be used to estimate the contamination of recyclable waste streams. Moreover, the analysis helps to evaluate the awareness campaign and its effectiveness on the segregation behavior of people.

2.8.2.4. System-Level Integration Benefits.
The combination of sensing, intelligent processing, actuation and monitoring that is part of a single smart bin architecture can be fully autonomous and therefore operated with a lack of manual intervention. Its modular construction enables it to be scaled across various locations within cities. It enables its behavior to be consistently segregated across various environmental conditions. Centralized monitoring will allow the same performance standards to be applied in the heterogeneous deployments. This combined design will offer a viable groundwork to large scale implementation of the CPV-IoT model and transform the process of waste segregation into a computer-assisted service of a smart city.

2.9 Experimental Setup

Table 1 provides the dataset used for this study. The experiments use the Waste Classification Dataset (Mendeley Data, Version 2), which comprises 24, 705 RGB waste pictures that are divided into two categories: organic and recyclable [18]. The JPG images included in the dataset were annotated by humans and were taken in home and communal garbage disposal situations and variable resolutions are standardized during preprocessing.

Table 1. Dataset Description

Attribute	Description
Data Type	RGB Images
Total Images	24,705
Classes	Organic Waste, Recyclable Waste
Class Distribution	Organic: 12,500 images; Recyclable: 12,205 images
Image Format	JPG
Image Resolution	Variable (resized during preprocessing)
Acquisition Environment	Household and public waste disposal scenarios
Labeling Method	Manual annotation

Table 2 summarizes the experimental setting of training and evaluation. All the experiments were done on a high-performance workstation featuring 64 GB RAM and an NVIDIA graphics card with 16 GB VRAM with an operating system of

Ubuntu 24.04. The experiment is done with Python and the PyTorch framework with a standardized input resolution and a fixed train-validation-test split. The architecture can be used to facilitating effective training offline and simulation-based inference, which is representative of the computational properties of an edge-compatible deployment environment.

Table 2. Experimental Setup

Component	Specification
Hardware Platform	High-performance workstation
Processor & Memory	64 GB RAM
Graphics Processing Unit	NVIDIA GPU with 16 GB VRAM
Operating System	Ubuntu 24.04
Camera Type	Simulated RGB camera (smart-bin acquisition model)
Programming Framework	Python
Deep Learning Framework	PyTorch
Image Input Size	224×224
Data Split	70% Training, 15% Validation, 15% Testing
Optimization Algorithm	AdamW
Learning Rate	$1 \times 10^{-}$
Batch Size	32
Number of Epochs	50
Loss Function	Weighted Cross-Entropy
Evaluation Mode	Offline training and simulation-based inference

3 Results and Discussion

3.1 Experimental Results

The result of the current baseline model and proposed model in the waste classification dataset is shown in Table 3. The SVM with HOG features had the worst performance with an accuracy of 86.42% and F1-score of 85.80%, which means the limited ability to establish complicated visual features. However, deep learning models showed significant progress and EfficientNet-B0 reached an accuracy of 93.48% with Vision Transformer (ViT-B/16) getting even higher accuracy of 94.21%. The best performance in all measures under consideration was obtained with the proposed CPV-IoT, which has an accuracy of 96.87%, precision of 96.54%, recall of 96.29% and F1-score of 96.41%.

Table 3. Comparative Performance of Baseline Models on Waste Classification Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM (HOG Features)	86.42	85.97	85.63	85.80
Random Forest	88.15	87.94	87.52	87.73
ResNet-50	92.36	92.11	91.84	91.97
EfficientNet-B0	93.48	93.26	93.02	93.14
MobileNetV3-Large	92.89	92.54	92.21	92.37
Vision Transformer (ViT-B/16)	94.21	94.05	93.78	93.91
Smart Waste Segregation [19]	93.30	-	-	-
ViT [20]	96.74	95.00	99.00	96.60
Proposed CPV-IoT	**96.87**	**96.54**	**96.29**	**96.41**

In Table 3, SVM and Random Forest, which are traditional machine learning models, perform the lowest. These models are based on handcrafted features that cannot be relied on to depict the complex visual features of waste images. Consequently, they find it difficult when there are differences in texture, shape and deformation of objects usable in actual disposal situation. ResNet-50, EfficientNet-B0 and MobileNetV3-Large are deep learning baselines that are higher-accuracy models, where the full database is learned by hierarchical visual features directly provided as images. Nevertheless, such models are primarily concerned with the local appearance patterns. They fail to explicitly encode the spatial relationship or the contextual relationship between various portions of the waste object. The result of this weakness is the mixed use of closely related waste categories which is captured by the relatively low precision and recall scores.

Vision Transformer (ViT-B/16) enhances the performance with an ability to capture the global relationships with the help of self-attention. In spite of this advantage, ViT works with homogeneous image patches. It does not differentiate material, geometrical and contextual cues on structural level. This limits its capability to manage littered or half blocked garbage scenes.

The offered CPV-IoT framework also demonstrates the greatest performance in all metrics. This is mainly through the HTD. The HTD distinguishes between surface, geometric and contextual data and then reasons. The structured representation minimizes the ambiguity among visual similar waste classes.

The contextual reasoning layer is another layer that allows better performance by simulating interactions among various types of tokens. This enables the model to view waste objects as coherent structures and not as isolated regions. The attention mechanism enhances the semantic associations and repressed the noisy visuals.

The organic sample is accurately classified with a confidence of 96% and those that are recyclable with a confidence of 97%. These findings imply that the suggested CPV-IoT architecture can be successfully used to draw the line between visually different material categories with a high degree of confidence. The delimiting areas also indicate that the model emphasizes on the pertinent areas of the objects in the inference. The large confidence values in Figure 5 indicate the good quantitative performance in the experimental evaluation.

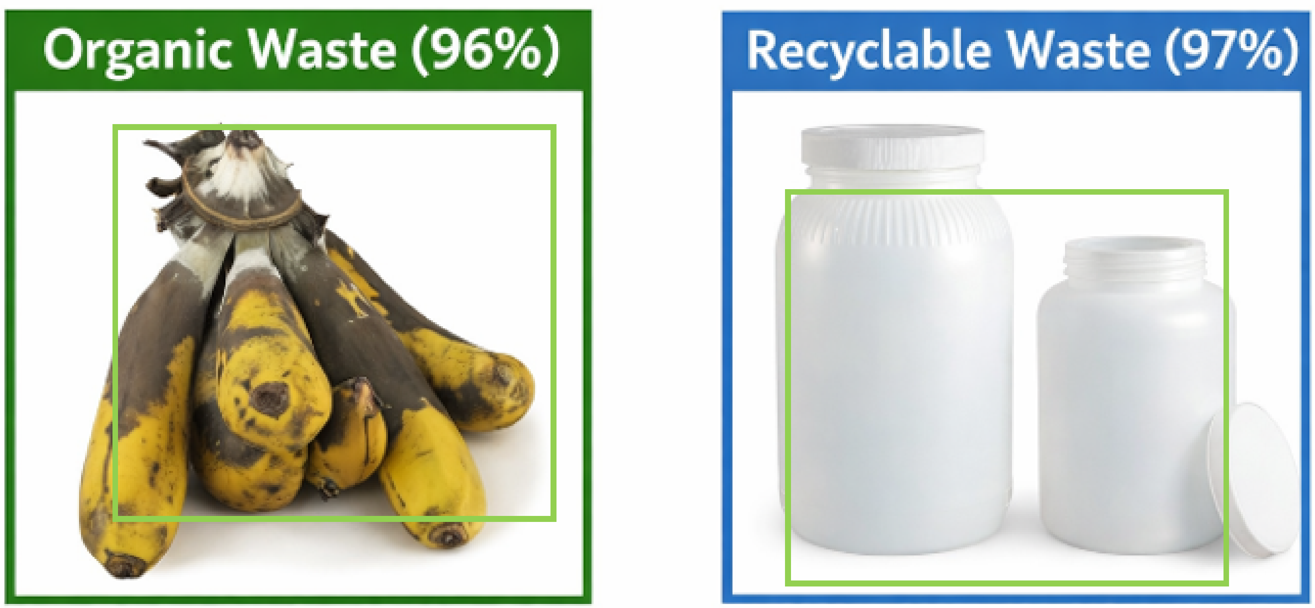

Fig 5. Sample of classified organic and recyclable labels

3.2 Ablation Study

The findings of the ablation study presented in Table 4 show the incremental value of each component of the CPV-IoT framework. CNN backbone only setup has the lowest performance with the average accuracy of 92.36% with F1-score of 91.97%, which means that it cannot detect intricacies in waste appearance variation. The addition of token decomposition with hierarchical structure and tokens without context enhances the precision to 94.02% and F1-score to 93.71, which proves the value of the structured visual representation. Additional incorporation of surface and geometric tokens enhance performance to accuracy of 95.11% and F1-score of 94.86%, which implies the usefulness of material and shape semantics. The architecture that includes contextual reasoning and does not have a transformer reaches 95.74% precision and 95.38% F1-score, which proves the importance of contextual integration. The complete CPV-IoT model attains the best performance, being 96.87% of accuracy and 96.41% of F1-score.

Table 4, which is the result of the ablation, clearly shows the contribution of each component. The accuracy and F1-score decrease significantly when context tokens are removed. Omission of transformer based reasoning is also a limiting

Table 4. Ablation Study: Impact of CPV-IoT Components

Configuration	Accuracy (%)	F1-Score (%)
CNN Backbone Only	92.36	91.97
HTD without Context Tokens	94.02	93.71
HTD + Surface & Geometry Tokens	95.11	94.86
HTD + Contextual Reasoning (No Transformer)	95.74	95.38
Full CPV-IoT (HTD + Contextual Reasoning)	96.87	96.41

performance even in the presence of contextual information. The entire CPV-IoT system is the most effective though it proves that hierarchical tokenization and contextual reasoning are both necessary.

4 Conclusion

This study introduced a Context-Preserving Vision -IoT (CPV-IoT) conceptualized vision-based waste classification, a system that operates on the most critical limitations of the previous vision-based waste management approaches. The primary novelty of the work is the introduction of HTD with contextual reasoning. It explicitly distinguishes between surface, geometric and spatial semantics and then proceeds to classification. This categorical representation is in direct connection to the research objective that is to enhance the reliability of waste separation in visually challenging disposal conditions. Analysis of a publicly available waste dataset on the example of simulation showed that the proposed CPV-IoT framework showed consistent performance compared to the old machine learning and new deep learning baselines. The suggested model had a classification accuracy of 96.87% and an F1-score of 96.41%, which is much higher than the baseline of Vision Transformer and EfficientNet. These quantitative benefits validate that besides context-preserving representation, structured reasoning highly minimizes inter-class confusion. The analysis of ablation confirms the role of each single component further. More than 2.8% accuracy was lost when contextual tokens were not included and loss of 1.1% accuracy was recorded when reasoning was not performed using transformers. These findings prove that hierarchical tokenization and contextual reasoning are critical to the success of the suggested framework.

Despite such contributions, several limitations still exist. The present research will assess the framework to the simulation based experiments and available datasets, without physically implementing it on real smart-bin equipment. The dataset present is restricted to binary wastes and the real-world municipal waste streams have a higher level of complexity. Besides, there were no network-level communication behavior and long-term system reliability tests. Future research will involve the expansion of the framework to multi-class and hazardous waste data, incorporation of other sensor modalities and testing the system on actual IoT-enabled smart bin prototypes. Also, large-scale deployment behavior, energy efficiency and long-term operational stability need to be researched.

5 Acknowledgements

The author gratefully acknowledges the management of St. Joseph's College (Autonomous), Tiruchirappalli, for providing Research Grant-2025 to support this research work.

References

1) Kwakye SO, Amuah EE, Ankoma KA, Agyemang EB, Owusu BG. Understanding the performance and challenges of solid waste management in an emerging megacity: Insights from the developing world. *Environmental Challenges*. 2024 Jan;14:100805. https://doi.org/10.1016/j.envc.2023.100805.
2) Yadav SS, Gupta SK, Yadav MS, Shinde R. .
3) Lakhouit A. Revolutionizing urban solid waste management with AI and IoT: a review of smart solutions for waste collection, sorting and recycling. *Results in Engineering*. 2025 Jan;p. 104018. https://doi.org/10.1016/j.rineng.2025.104018.
4) Hassija V, Palanisamy B, Chatterjee A, Mandal A, Chakraborty D, Pandey A, et al. Transformers for Vision: A Survey on Innovative Methods for Computer Vision. *IEEE Access*. 2025 May. https://ieeexplore.ieee.org/abstract/document/11007557.
5) Single S, Iranmanesh S, Raad R. RealWaste: a novel real-life data set for landfill waste classification using deep learning. *Information*. 2023 Nov;14(12):633. https://doi.org/10.3390/info14120633.
6) Prasad V, Arashpour M. ShARP-WasteSeg: A shape-aware approach to real-time segmentation of recyclables from cluttered construction and demolition waste. *Waste Management*. 2025 Mar;195:231–239. https://doi.org/10.1016/j.wasman.2025.02.006.
7) Hirbod F, Karimi T, Mohammadnazari Z, Rabbani M, Aghsami A. Municipal solid waste management using multiple disposal location-arc routing and waste segregation approach: a real-life case study in England. *Journal of Industrial and Production Engineering*. 2024 Jan;41(1):81–100. https://doi.org/10.1080/21681015.2023.2265925.

8) Resolute P. Applying cultural perspective in Indonesia municipal solid waste management process towards a grounded climate concept and action: A mini-review of opportunities and challenges. *Waste Management & Research*. 2024 Oct;42(10):873–881. https://doi.org/10.1177/0734242X241262711.
9) Nahiduzzaman M, Ahamed MF, Naznine M, Karim MJ, Kibria HB, Ayari MA, et al. An automated waste classification system using deep learning techniques: Toward efficient waste recycling and environmental sustainability. *Knowledge-Based Systems*. 2025 Feb;310:113028. https://doi.org/10.1016/j.knosys.2025.113028.
10) Ahmad G, Aleem FM, Alyas T, Abbas Q, Nawaz W, Ghazal TM, et al. Intelligent waste sorting for urban sustainability using deep learning. *Scientific Reports*. 2025 Jul;15(1):27078. https://doi.org/10.1038/s41598-025-08461-w.
11) Dipo MH, Farid FA, Mahmud MS, Momtaz M, Rahman S, Uddin J, et al. Real-Time Waste Detection and Classification Using YOLOv12-Based Deep Learning Model. *Digital*. 2025 Jun;5(2):19. https://doi.org/10.3390/digital5020019.
12) Alourani A, Ashraf MU, Aloraini M. Smart waste management and classification system using advanced IoT and AI technologies. *PeerJ Computer Science*. 2025 Apr;11:e2777. https://doi.org/10.7717/peerj-cs.2777.
13) Adeoye S. Smart Waste Management Systems: An IoT and AI-Driven Approach for Urban Sanitation. *Computer and Information Science*. 2025;18(2). https://ccsenet.org/journal/index.php/cis/article/view/0/52299.
14) Fotovvatikhah F, Ahmedy I, Noor RM, Munir MU. A Systematic Review of AI-Based Techniques for Automated Waste Classification. *Sensors (Basel)*. 2025 May;25(10):3181. https://doi.org/10.3390/s25103181.
15) Kunwar S. DWaste: Greener AI for Waste Sorting using Mobile and Edge Devices. *arXiv preprint arXiv:251018513*. 2025 Oct;p. 21. https://doi.org/10.48550/arXiv.2510.18513.
16) Nguyen NB, Do TM, Phan CT, Phan TT. Towards Accurate and Efficient Waste Image Classification: A Hybrid Deep Learning and Machine Learning Approach. *arXiv preprint arXiv:251021833*. 2025 Oct;p. 22. https://doi.org/10.48550/arXiv.2510.21833.
17) Wirani Y, Eitiveni I, Sucahyo YG. Framework of Smart and Integrated Household Waste Management System: A Systematic Literature Review Using PRISMA. *Sustainability*. 2024;16(12):4898. https://doi.org/10.3390/su16124898.
18) Nnamoko N, Barrowclough J, Procter J. Waste Classification Dataset. *Mendeley Data*. 2022.
19) Gawai JS, Bhavsar K, Shahare S, Merkhed A, Pahade K. Smart Waste Segregation System Using MobileNetV2. *International Journal of Electrical, Electronics and Computer Systems*. 2025 May;14(1):180–183. https://doi.org/10.17148/IJARCCE.2025.14374.
20) Tanwer PS, Maheshwari S, Behera S, Chauhan A, Kumar ST. Evaluation of vision transformers for the detection of fullness of garbage bins for efficient waste management. *Frontiers in Artificial Intelligence*. 2025 Sep;8:1612080. https://doi.org/10.3389/frai.2025.1612080.