SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Literature Survey on Solid Waste Classification Using Deep Learning Techniques and Explainable AI

K. Gomathi[1], Michael G.[2]

[1]Research Scholar; Institute of Computer Science and Engineering; Saveetha School of Engineering, Saveetha Institute of Medical and Technical Science(SIMATS), Saveetha University, Chennai, Tamil Nadu, PIN code: 602105. E-mail: mathikrishnan55@gmail.com, gomathi.k@kgr.ac.in, Orcid: https://orcid.org/0009-0009-9060-6631
[2]Professor, Department of CSE, E-mail: michaelg.sse@saveetha.com, Orcid Id: 56924800500

ABSTRACT
The rapid growth of municipal solid waste has created significant environmental, economic, and public health challenges, highlighting the need for automated waste classification to support recycling, resource recovery, and smart waste management. This survey systematically reviews deep learning techniques for image-based solid waste classification, with particular emphasis on the role of Explainable Artificial Intelligence (XAI) in enhancing model transparency, interpretability, and trustworthiness. The review examines (i) publicly available waste image datasets and their characteristics and limitations, (ii) deep learning approaches, including convolutional neural networks (CNNs), transfer learning, vision transformers, and object-detection models, (iii) post-hoc XAI techniques, such as Grad-CAM, Grad-CAM++, SHAP, and LIME, and their integration into waste classification pipelines, and (iv) commonly adopted evaluation protocols and performance measures. The reviewed studies indicate that CNN-based and transfer-learning approaches can achieve high classification performance on benchmark datasets, although their effectiveness varies considerably with dataset composition, class distribution, experimental settings, and real-world conditions. XAI methods are increasingly employed to identify the visual features influencing model predictions and to assess whether automated decisions are based on semantically meaningful waste characteristics. The survey further identifies major challenges, including dataset bias and class imbalance, limited cross-dataset and real-world generalization, lack of standardized evaluation protocols, and insufficient interpretability. Finally, future research directions are discussed, including sustainability-aware deep learning, multimodal data integration, robust and generalizable XAI frameworks, and resource-efficient models for real-time waste classification and deployment in smart-city environments.

Keywords: Municipal solid waste, waste classification, deep learning, convolutional neural networks, transfer learning, vision transformers, Explainable Artificial Intelligence, XAI, image classification, smart waste management.

1. INTRODUCTION

Municipal solid waste is one of defining environmental problems of the current century according to recent statistics approximately 2.01 billion tons of municipal solid based was generated every day globally in 2016 and rapid urbanization population growth and economic development have continued to increase waste volumes worldwide(Fang et al., 2023).Inadequate waste processing contributes to pollution inefficient recycling and resource loss motivating the development of automated waste classification systems that can reduce manual labor in based separation and recycling activities(Nahiduzzaman et al., 2025). Traditional manual sorting is slow, labour-intensive, and prone to errors, resulting in reduced recycling efficiency and increased environmental pollution (K. et al., 2025).Waste classification is therefore considered as a critical component of effective recycling because reliable identification and segregation of these materials were essential for efficient resource recovery[4].Deep learning has increasingly emerged as a promising computational approach for addressing various solid waste management challenges.

Systematic reviews have reported that machine learning and deep learning techniques demonstrate strong potential for detecting and classifying different waste categories that supports improved decision making resource recovery and circular economy practices. An AI based identification and sorting systems have reported classification accuracies ranging from 72 to 99% while AI assisted waste logistics have demonstrated reductions in transportation distance,cost and processing time.Despite these advances,deep learning models are frequently regarded as black boxes and their limited interpretability can hinder user trust transparency and practical adoption in real world waste management applications. Consequently an emerging body of research has integrated deep learning classification models with explainable artificial intelligent techniques (XAI), including Gradient Weighted Class Activation Mapping,Shapely Additive Explanations (SHAPE) and local interpretable model agnostic explanations (LIME) to provide interpretable evidence for model predictions and improve accountability.

A. Literature Search and Review Methodology

A structured literature-review approach was adopted to identify, examine, and synthesize research on deep-learning-based solid waste classification and Explainable Artificial Intelligence (XAI). The review focused primarily on peer-reviewed studies addressing image-based waste classification, object detection, transfer learning, vision transformers, and explainability techniques applied to waste-management applications. Relevant studies were identified from the published literature using combinations of keywords such as solid waste classification, waste image classification, deep learning, convolutional neural networks, transfer learning, vision transformers, object detection, Explainable Artificial Intelligence, Grad-CAM, SHAP, and LIME. The identified studies were screened based on their relevance to image-based solid waste classification and their contribution to model development, dataset construction, evaluation, or interpretability. Studies focusing exclusively on unrelated waste-management applications without an image-classification or explainability component were excluded from the thematic synthesis. The selected literature was subsequently organized into four major themes: (i) waste image datasets and dataset-related limitations, (ii) deep learning architectures and classification strategies, (iii) XAI methods and their integration with waste classification models, and (iv) evaluation practices, research challenges, and future directions.Rather than comparing reported accuracy values alone, the review critically examines the relationship among datasets, model architectures, evaluation protocols, computational requirements, real-world generalization, and interpretability. Particular attention is given to whether high reported performance is supported by realistic datasets, balanced class distributions, robust validation strategies, and meaningful explanations. This approach enables the survey to identify methodological gaps that may not be apparent from accuracy-based comparisons alone.

B. Scope of the study and Contribution of this Survey

This survey synthesizes the peer-reviewed literature on deep learning for image-based solid waste classification, with particular emphasis on the integration of Explainable Artificial Intelligence (XAI). The survey extends existing reviews by jointly examining datasets, deep learning architectures, evaluation practices, computational considerations, and interpretability rather than treating classification performance and explainability as separate research themes.

The main contributions of this survey are as follows:

Dataset-centric analysis: The survey provides a structured overview of widely used and emerging waste image datasets, including their class distributions, image characteristics, annotation practices, dataset bias, class imbalance, and limitations for real-world deployment. Comparative taxonomy of deep learning approaches: CNNs, transfer-learning architectures, vision transformers, hybrid models, and object-detection approaches are systematically categorised and critically discussed in terms of reported performance, application context, and deployment requirements. Focused analysis of XAI integration: Unlike reviews that primarily discuss classification performance, this survey specifically examines the integration of post-hoc XAI methods, including Grad-CAM, Grad-CAM++, SHAP, and LIME, into waste-classification pipelines and evaluates their role in supporting model transparency and feature relevance. Evaluation and reproducibility perspective: The survey analyzes the metrics and validation strategies used across studies and highlights the lack of standardised datasets, evaluation protocols, and quantitative XAI assessment that limits meaningful comparison between models.Identification of deployment-oriented research gaps: The review connects model accuracy with practical constraints such as dataset bias, real-world generalisation, computational efficiency, edge-device deployment, and interpretability, thereby identifying challenges that must be addressed before large-scale implementation.Integrated future research roadmap: Based on the identified gaps, the survey proposes future directions involving robust and diverse datasets, cross-domain generalization, resource-efficient architectures, standardized XAI evaluation, multimodal learning, and explainability-by-design for real-time smart waste-management systems.

2. BACKGROUND AND PRELIMINARIES

A. Deep Learning for Image Classification

Convolution neural networks have been widely employed for based classification because of their ability to automatically extract come complex visual features from raw images.Early CNN based waste classification approaches typically consist of limited number of convolution layers hidden layers and a softmax output layer for multi class classification.More advanced architectures have subsequently been developed including multi modal CNN frameworks that integrate network such as DSSD,Yolov4 and R-CNN for domestic waste detection and classification.Hybrid multilayer CNN architectures have also demonstrated promising performance achieving accuracies of up to 92% on the Trashnet data set. Because waste image datasets are often small,transfer learning has become widely adopted strategy.In this approach a CNN is initialized with weight pre trained on Imagenet and subsequently fine tuned using waste image datasets. More recently Vision transformers which employ self attention mechanisms rather than conventional convolutional operations have emerged a competitive alternative to CNN particularly because of their ability to model long range and to overcome limitations associated with the local fields of convolutional architectures.

B. Explainable Artificial Intelligence

Explainable AI techniques comprise a range of concepts taxonomies and techniques designed to improve the transparency, interpretability and accountability of AI systems.In computer vision the choice of XAI technique is often influenced by the underlying model architecture.For Convolutional neural network, commonly used approaches include saliency maps,layer wise relevance propagation(LRP),Integrated gradients and Class Activation Mapping(CAM).For attention based architecture such as Vision Transformers, explanation methods generally focused on attention visualisation and feature attribution techniques.Among convolution-based approaches gradient weight class activation mapping generates localization by using the gradients associated with the target class at the final convolutional layer. The resulting heat map highlights in major regions that contribute to most strongly to the model's prediction and can generally be applied without modifying the model.Grad-CAM also extends the applicability of class activation explanations to a broader range of CNN architectures.An enhanced variant Grad-CAM++, provides improved localization and is particularly effective when multiple instances of the same object appear within an image.In addition to the gradient based techniques model agnostic methods such as shapely additive explanations and local interpretable model agnostic expression can be used to identify the contribution of individual features or image regions to a models prediction.These methods are particularly useful when interpretability is required independently of the underlying image classification model.

The general structure linking deep learning and XAI in solid waste classification is summarized in Figure 1.

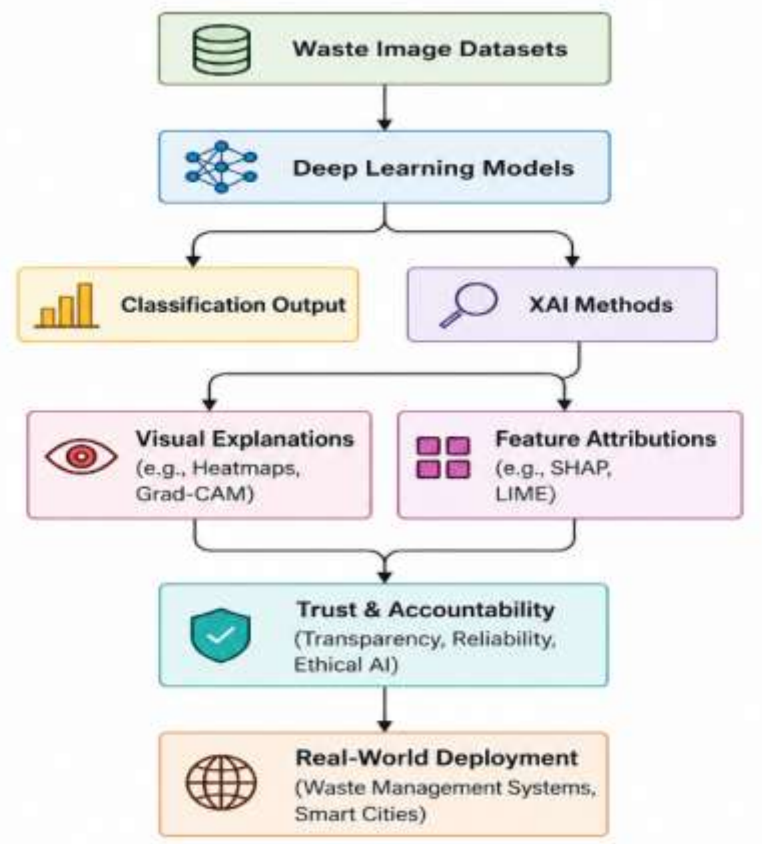

Fig. 1. Conceptual taxonomy of deep-learning-based solid waste classification with explainable AI integration.

3. DATA SETS FOR SOLID WASTE CLASSIFICATION

Data set availability and quality or decisive factors in the performance and generalizability of phased classifiers A systematic literature review of 97 studies identified over 15 publicly available based classification datasets and highlighted key limitations including class imbalance real world variability and standardization issues not discussed Table1 one summarizes the most frequently used datasets in waste classification.

TABLE 1: Most frequently used data set in waste classification

Dataset	No. of images	Classes	Notes
TrashNet	2,527	6	Glass, paper, cardboard, plastic, metal, trash; uniform white-poster-board background
TACO	1,500 (4,784 annotations)	28 supercategories	Trash in the wild, crowdsourced, COCO-format annotations
VN-Trash	5,904	3	Organic, inorganic, and medical waste from Vietnam
TriCascade WasteImage	35,264	36 (hierarchical 2/9/36)	Combines four pre-existing datasets; largest reported waste image set
domestic garbage-classification	2,467	6	TrashNet-equivalent distribution, captured across 500+ cities
RECYCLEYE	3 million	8	Private industrial dataset; labels include material, object, color, etc.

A. TrashNet

TrashNet is the most widely used benchmark for waste image classification. It consists of 2,527 images divided into six classes: 501 glass, 594 paper, 403 cardboard, 482 plastic, 410 metal, and 137 trash images (Bourougaa-Tria et al., 2022; Ma et al., 2022). Images were captured by mounting objects on a white poster board under natural and/or artificial light and scaled to 512 × 384 pixels (Bourougaa-Tria et al., 2022; Ma et al., 2022). The uniform background makes TrashNet easy to learn but unrepresentative of real-world waste encountered after collection (Shaikjee & Haar, 2022). Experiments show that datasets with uniform structure such as TrashNet enable high classification accuracy but may lead to overfitting, with decreased validation performance; on TrashNet, models achieved overall accuracy of 0.66 and best F1-scores for cardboard (0.84), paper (0.75), and glass (0.64) (Chomicki et al., 2025).

B. TACO and Real-World Datasets

TACO is an open, crowdsourced image dataset for litter detection and segmentation, containing 1,500 images and 4,784 annotations of trash in diverse environments such as woods, roads, and beaches (Proença & Simões, 2020). It follows COCO formats, is grouped into 28 supercategories, and provides binary masks for each waste object, enabling instance segmentation (Shaikjee & Haar, 2022). Because each image may contain several waste objects, TACO better represents solid waste in the wild than TrashNet (Shaikjee & Haar, 2022). However, satisfactory trash detection in the wild for deployment requires many more manual annotations (Proença & Simões, 2020). The VN-Trash dataset (5,904 images, three classes of organic, inorganic, and medical waste) was introduced for smart waste-sorting machines (Vo et al., 2019). Majchrowska et al. performed a critical analysis of over ten existing waste datasets and proposed merged benchmark collections — detect-waste and classify-waste — with unified annotations spanning seven categories (bio, glass, metal and plastic, non-recyclable, other, paper, and unknown) to address the lack of widely accepted benchmarks and metrics (Majchrowska et al., 2021).

C. Dataset Limitations

The literature consistently identifies three dataset-related obstacles: (i) the images used to train transfer-learning and CNN models for waste classification are far less complex than real-world waste, reducing accuracy in practical applications (Fang et al., 2023); (ii) class imbalance and high visual variability in realistic datasets create training difficulties (Chomicki et al., 2025); and (iii) the absence of standardization hampers comparability across studies (Fotovvatikhah et al., 2025; Majchrowska et al., 2021). Data availability, quality, and privacy are likewise recurrent concerns in AI-enabled waste management (Olawade et al., 2024).

4. Deep Learning Techniques for Solid Waste Classification

A. CNN-Based Approaches

Custom CNN architectures remain a workhorse for waste classification. Deep convolutional networks classify municipal waste into multiple categories with reported accuracies as high as 97.63% using single-shot detectors and Faster R-CNN features (Fang et al., 2023). Zhang et al. added a self-monitoring module to a residual network model that divides waste into six types by material, achieving 95.87% accuracy, although the small dataset and limited realism of the images restricted practical applicability (Fang et al., 2023). Fahmi and Lubis developed a CNN-based waste recognition system classifying waste into inorganic and organic

substances, each subdivided into five subclasses, with an average accuracy of 90% (Fang et al., 2023).
The ScrapNet model, based on the EfficientNet architecture, achieves 98% accuracy on TrashNet while using comparatively fewer parameters than existing methods, and 92.87% accuracy on a newly created standardized dataset of 8,135 images (Masand et al., 2021). In a 2026 benchmark comparison on TrashNet, Swin Transformer (88.2%), SE-ResNet (87.1%), and improved ResNeXt-50 (86.9%) outperformed ResNet-18 (85.7%), EfficientDet (83.2%), VGG16 (82.5%), YOLOv8 (81.4%), MobileNetV2 (79.3%), and ShuffleNetV2 (77.6%) (Liu et al., 2026).
Convolutional neural networks (CNNs) have been widely employed for waste image classification because of their ability to automatically extract complex visual features from raw images. Early CNN-based waste classification approaches typically consisted of a limited number of convolutional layers, hidden layers, and a softmax output layer for multi-class classification . More advanced architectures have subsequently been developed, including multi-model cascaded CNN frameworks that integrate networks such as DSSD, YOLOv4, and Faster R-CNN for domestic waste detection and classification. Hybrid multilayer CNN architectures have also demonstrated promising performance, achieving accuracies of up to 92.6% on the TrashNet dataset.Because waste-image datasets are often relatively small, transfer learning has become a widely adopted strategy. In this approach, a CNN is initialized with weights pretrained on ImageNet and subsequently fine-tuned using waste-image datasets. More recently, Vision Transformers (ViTs), which employ self-attention mechanisms rather than conventional convolutional operations, have emerged as competitive alternatives to CNNs, particularly because of their ability to model long-range dependencies and overcome limitations associated with the local receptive fields of convolutional architectures.

B. Transfer-Learning-Based Approaches

Transfer learning dominates the waste-classification literature, and CNN-based architectures with transfer learning show particularly promising results (Fotovvatikhah et al., 2025). Vo et al. proposed DNN-TC, an improvement of the ResNeXt model, yielding 94% accuracy on TrashNet and 98% accuracy on the VN-Trash dataset, outperforming state-of-the-art methods on both (Vo et al., 2019). A combination model fusing three pre-trained CNNs — VGG19, DenseNet169, and NASNetLarge — achieved 96.5% and 94% classification accuracy on two waste image datasets (Huang et al., 2020). On a two-class biodegradable/non-biodegradable formulation of TrashNet, a NASNetMobile-based transfer model reached 97% accuracy (Karthikeyan, 2022). DenseNet169 achieved 93.10% accuracy on a 17,628-image, 11-category household waste dataset (Kunwar, 2024).
On TrashNet specifically, ResNet-50 transfer learning with a multi-class SVM classifier reached 87% accuracy (Khan et al., 2024); Inception-v3 achieved 92.5% on a plastic-model detection dataset (Chhabra et al., 2024); an Inception-ResNet combination produced 88.6% (Chhabra et al., 2024); and an improved multi-layered CNN achieved 87% (Chhabra et al., 2024). Transfer-learning models trained on the TrashBox dataset — which has more images and classes than TrashNet or TACO — demonstrated accuracy of 98.47% (Khan et al., 2024). An EfficientNet-B0 model trained with a two-stage transfer-learning strategy (feature extraction followed by fine-tuning) achieved 94.3% average accuracy (standard deviation ±0.007) under 5-fold cross-validation with only 4.38M parameters, outperforming ResNet50 and DenseNet121 while using the fewest parameters (Manik et al., 2026).

C. Transformer-Based Approaches

Vision transformers address the receptive-field bottleneck of CNNs. A ViT-based architecture achieved 96.98% accuracy on TrashNet, surpassing existing CNN methods, and was deployed as a server-side classifier accessed from a portable device (Huang et al., 2021). A ViT application for classifying 12 types of household garbage achieved 95% accuracy, significantly surpassing VGG16 and MobileNetV2 (Prova, 2024). Hybrid designs that combine convolution and attention are also prominent: Garbage FusionNet integrates ResNet's local feature extraction with ViT's global information capture, achieving 96.54% accuracy on a 10-category dataset of 23,642 images — surpassing standalone ResNet50 and ViT by 1.09 and 4.18 percentage points, respectively (Wang et al., 2024). Transformers also excel in post-collection scenarios: ViT-based models detect contamination in densely cluttered waste scenes with superior accuracy relative to leading object detectors (Mewada et al., 2025), and a vision transformer reached 96.74% accuracy in detecting garbage-container fullness on a public Montevideo dataset (Tanwer et al., 2025).

D. Object-Detection-Based Approaches

For real-time sorting, object detection models localize and classify waste simultaneously. YOLO-Green, designed specifically for trash detection and trained for 100 epochs on a seven-class real-world dataset, achieved a mean average precision (mAP) of 78.04% at 2.72 FPS with a 117 MB model size, exceeding YOLOv4 in both accuracy and efficiency (Lin, 2021). A YOLOv8-based real-time system for household waste, enhanced with five data-augmentation techniques and attention mechanisms, achieved 89.5% mAP over 17 waste types (3,775 images) — a 4.2% improvement over the baseline (Arishi, 2025). YOLOv12-based detection reported 73% precision and 78% mAP in 100 epochs (Dipo et al., 2025). In the segmentation-oriented line of work, a two-stage detector combining EfficientDet-D2 for litter localization and EfficientNet-B2 for seven-category classification reached up to 70% average precision in detection and approximately 75% classification accuracy

(Majchrowska et al., 2021).

Table II — Representative Deep Learning Models For Solid Waste Classification

Model / Approach	Dataset	Accuracy (reported)	XAI used
DNN-TC (improved ResNeXt)	TrashNet / VN-Trash	94% / 98%	No
ScrapNet	TrashNet / custom	98% / 92.87%	No
ViT	TrashNet	96.98%	No
ViT, 12 household classes	custom	95%	No
Garbage FusionNet (ResNet+ViT)	10-class, 23,642 images	96.54%	No
NASNetMobile transfer	TrashNet (2 classes)	97%	No
VGG19+DenseNet169+NASNetLarge	two waste datasets	96.5% / 94%	No
DenseNet169 transfer	11-class, 17,628 images	93.10%	No
EfficientNet-B0, two-stage transfer	multi-class waste	94.3% ± 0.007 (5-fold CV)	Grad-CAM
MobileNetV2 + Grad-CAM + SHAP	benchmark + field images	90.25% (validation)	Grad-CAM, SHAP
DP-CNN + Ensemble ELM	TriCascade	96% / 91% / 85.25% (3 stages)	Multiple XAI methods
Swin Transformer / SE-ResNet	TrashNet	88.2% / 87.1%	No

5. EXPLAINABLE AI IN SOLID WASTE CLASSIFICATION

A. Motivation

The central argument for XAI in waste classification is accountability: Although deep CNNs deliver high accuracy, their limited interpretability may hinder user trust and real-world adoption in municipal and industrial setting. Lack of confidence in model results is one of the main reasons ML methods are not yet widely used in productive business processes (Stadlhofer & Mezhuyev, 2023). In waste management specifically, researchers have argued that explainability makes models accountable and reliable enough for large organizations to deploy them for daily household and industrial waste segregation (Gupta et al., 2023).

B. XAI Methods Applied to Waste Classification

Grad-CAM and Grad-CAM++. Grad-CAM produces coarse localization heatmaps highlighting the regions of a waste image that drive a model's decision (Selvaraju et al., 2019); Grad-CAM++ refines these explanations for multiple object instances (Chattopadhay et al., 2018). In waste classification, Grad-CAM analysis confirmed that an EfficientNet-B0 model's decisions are based on relevant object features rather than background artifacts (Manik et al., 2026), and Grad-CAM heatmaps have been used to verify that a MobileNetV2 classifier focuses on semantically relevant visual features in organic/inorganic waste images (Adib et al., 2025).

SHAP. SHAP provides additive feature-contribution explanations and has been implemented alongside Grad-CAM to quantify which image regions contribute most to waste-class predictions (Adib et al., 2025; Stadlhofer & Mezhuyev, 2023). Combining Grad-CAM and SHAP within one framework demonstrated transparent feature contributions, confirming that lightweight CNN architectures can yield classification systems that are simultaneously accurate, transparent, and accountable (Adib et al., 2025).

LIME. LIME generates locally faithful explanations by perturbing inputs and observing prediction changes and has been applied to waste classification in combination with SHAP and Grad-CAM for eight-category industrial waste segregation (Gupta et al., 2023; Stadlhofer & Mezhuyev, 2023).

Interpretability of constrained models. The DP-CNN-En-ELM framework — a nine-layer, 1.09-million-parameter depth-wise separable CNN paired with an ensemble of pseudoinverse and L1-regularized extreme learning machines — explicitly incorporated multiple XAI methods to explore the interpretability of its three-stage, 36-class hierarchical waste classification, while achieving 96%, 91%, and 85.25% accuracy at the two-class, nine-class, and 36-class stages with an average testing time of 0.00001 seconds (Nahiduzzaman et al., 2025).

6. EVALUATION METRICS

Waste classification studies standardly report accuracy, precision, recall, F1-score, and ROC-AUC (Nahiduzzaman et al., 2025), while detection-oriented systems use mean average precision (mAP), intersection-over-union, and inference throughput (Arishi, 2025; Liu et al., 2026). Comprehensive reporting is essential because datasets with uniform backgrounds such as TrashNet inflate accuracy relative to realistic settings (Chomicki et al., 2025). Several studies underscore that accuracy alone is insufficient: models must also be evaluated on robustness across data partitions (e.g., 5-fold cross-validation) and on interpretability evidence from XAI methods (Manik et al., 2026).

7. CHALLENGES AND RESEARCH GAPS

A. Dataset Bias, Class Imbalance, and Data Quality

Dataset quality remains one of the most significant limitations affecting the reliability of deep-learning-based solid waste classification. Many benchmark datasets contain images captured under controlled conditions, with limited background variation, illumination changes, object occlusion, and contamination. For example, the uniform background of TrashNet facilitates model training but does not adequately represent the heterogeneous conditions encountered in practical waste streams. Such dataset characteristics can introduce background bias and cause models to learn features that are not directly related to the waste category.

Class imbalance is another persistent challenge. Certain waste categories are substantially more represented than others, which can bias model optimization toward majority classes and result in poor recognition of minority categories. This problem becomes more pronounced for realistic datasets containing visually diverse and infrequently occurring waste materials. Although data augmentation, resampling, class-weighting, and synthetic data generation can partially address imbalance, their effectiveness and potential to introduce additional bias require systematic evaluation.

Furthermore, the lack of sufficiently large, diverse, and consistently annotated datasets limits the development of robust models. Existing datasets differ substantially in class definitions, image quality, annotation formats, environmental conditions, and data-collection procedures. Consequently, direct comparison of reported performance across studies is often unreliable.

B. Real-World Generalization and Standardized Evaluation

High performance on benchmark datasets does not necessarily translate into reliable performance in real-world waste-management environments. Waste items may appear partially occluded, damaged, contaminated, overlapping, or mixed with other materials. Changes in lighting, camera position, background, geographic location, waste composition, and collection practices can further produce distribution shifts between training and deployment environments.

The current literature also lacks standardized evaluation protocols. Studies frequently use different datasets, train-test splits, preprocessing procedures, augmentation strategies, performance metrics, and experimental settings. Consequently, a model reporting higher accuracy on one dataset cannot necessarily be considered superior to a model evaluated under different conditions.

Future studies should therefore adopt standardized benchmark datasets and reporting protocols and should evaluate models using independent test sets, cross-dataset validation, cross-domain testing, and real-world field data. In addition to accuracy, precision, recall, F1-score, mAP, inference latency, memory consumption, and energy requirements should be reported where appropriate. Such comprehensive evaluation would improve reproducibility and enable more meaningful comparison across studies.

C. Computational Efficiency and Deployment Constraints

Although deep CNNs, transformers, and hybrid architectures have demonstrated strong classification performance, their computational and memory requirements may limit deployment on edge devices, smart bins, embedded sorting systems, and portable waste-classification platforms. Large models may require substantial processing power, memory, and energy, which can be impractical for continuous real-time operation.

There is therefore a need to balance predictive performance with computational efficiency. Model compression, pruning, quantization, knowledge distillation, lightweight architectures, and efficient attention mechanisms represent promising strategies for reducing computational requirements. Future research should report not only classification accuracy but also model size, parameter count, floating-point operations, inference latency, memory requirements, and energy consumption to provide a more complete assessment of deployment readiness.

D. Limitations of Current XAI Methods

Although XAI methods such as Grad-CAM, Grad-CAM++, SHAP, and LIME improve the interpretability of waste-classification models, their current use has several limitations. Most studies employ post-hoc explanations that are generated after model training and are primarily assessed through qualitative inspection of heatmaps or feature-attribution visualizations. Such explanations do not necessarily demonstrate that the model has learned causally meaningful waste characteristics.

Gradient-based methods such as Grad-CAM may provide coarse spatial explanations and can be sensitive to the selected network layer. Perturbation-based methods such as LIME and SHAP may be computationally expensive for high-dimensional image inputs and can produce explanations that vary according to segmentation or perturbation settings. Furthermore, different XAI methods can produce inconsistent explanations for the same prediction, making it difficult to determine which explanation should be considered reliable.

Another important gap is the absence of standardized quantitative metrics for evaluating explanation quality specifically in waste-classification applications. Few studies investigate whether explanations improve operator trust, decision-making, error detection, or sorting performance. Future research should therefore move beyond visualization-based interpretation toward quantitative evaluation using explanation fidelity, stability, localization accuracy, consistency, and human-centered assessment. Integrating explainability directly into model design and combining complementary explanation techniques may further improve the reliability and practical usefulness of XAI-enabled waste-classification systems.

8. FUTURE DIRECTIONS

Future work should prioritize sustainability-aware frameworks, multi-modal data integration, and resource-efficient system deployment for scalable, real-time applications (Gomathi & Michael, 2026). Priorities identified in the broader MSWM literature include developing robust datasets, fostering cross-sector partnerships, and integrating ML/DL with IoT infrastructure, aligned with the United Nations Sustainable Development Goal 11 (inclusive, safe, resilient, sustainable cities) (Dawar et al., 2025). The most recent systematic reviews propose a structured roadmap organizing challenges and opportunities into short-, mid-, and long-term priorities, integrating model accuracy, system efficiency, and sustainability goals (Fotovvatikhah et al., 2025). For XAI specifically, promising directions include embedding explainability by design into lightweight mobile architectures, standardizing quantitative XAI evaluation for waste applications, and combining gradient-based, perturbation-based, and attention-based explanations to support both human operators and automated sorting robots.Future studies should prioritize the development of large-scale, diverse, geographically distributed, and consistently annotated waste image datasets.Research should move beyond evaluation on isolated benchmark datasets toward cross-dataset and real-world validation. Domain adaptation, domain generalisation, self-supervised learning, continual learning, and synthetic-to-real learning can be investigated to improve model robustness under distribution shifts. Field-based evaluation involving waste-sorting facilities, municipal collection systems, and smart-bin environments would provide stronger evidence of practical applicability.A standardised evaluation framework is required to enable reliable comparison among waste-classification models. Future studies should report classification and detection metrics together with computational and deployment indicators. Independent test sets, cross-validation, cross-dataset evaluation, confidence intervals, and statistical significance testing should be encouraged where appropriate. Standardized reporting would improve reproducibility and reduce the risk of drawing conclusions from dataset-specific performance.

9. Conclusion

This survey reviewed deep-learning approaches for image-based solid waste classification with particular emphasis on Explainable Artificial Intelligence. The literature demonstrates that CNNs, transfer-learning architectures, vision transformers, hybrid models, and object-detection approaches can achieve strong performance on several benchmark datasets. However, reported accuracy is strongly influenced by dataset characteristics, class distribution, experimental settings, and evaluation protocols and therefore should not be interpreted as direct evidence of real-world deployment readiness.The review identifies dataset bias, class imbalance, limited real-world generalization, inconsistent evaluation protocols, computational constraints, and insufficiently rigorous interpretability assessment as major barriers to practical adoption. Although XAI techniques such as Grad-CAM, Grad-CAM++, SHAP, and LIME provide useful insights into model predictions, current approaches remain predominantly post-hoc and qualitative, with limited standardized evaluation of explanation fidelity, stability, and human usefulness.Beyond summarizing existing methods, this survey establishes an integrated perspective connecting data quality, model performance, computational efficiency, real-world robustness, and explainability. Future research should therefore focus on diverse and representative datasets, cross-domain validation, standardized evaluation protocols, resource-efficient architectures, and quantitative and human-centered XAI evaluation. Integrating these directions with multimodal sensing, edge computing, IoT, and explainability-by-design can contribute to the development of reliable, transparent, sustainable, and deployable intelligent waste-management systems.

References

1. Al Adib, M., Siregar, A. A., Raj, B., Hasibuan, R. H. P., Nasution, J., Parapat, A. J., & Rosnelly, R. (2025). Explainable deep learning framework for waste image classification using MobileNetV2 with Grad-CAM and SHAP. Jurnal Komputer Teknologi Informasi Sistem Informasi (JUKTISI), 4(3), 1697–1710. https://doi.org/10.62712/juktisi.v4i3.739
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
3. Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized

gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 839–847). IEEE. https://doi.org/10.1109/WACV.2018.00097
4. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. WIREs Data Mining and Knowledge Discovery, 11(5), e1424. https://doi.org/10.1002/widm.1424
5. Fang, B., Yu, J., Chen, Z., Osman, A. I., Farghali, M., Ihara, I., Hamza, E. H., Rooney, D. W., & Yap, P.-S. (2023). Artificial intelligence for waste management in smart cities: A review. Environmental Chemistry Letters, 21, 1959–1989. https://doi.org/10.1007/s10311-023-01604-3
6. Fotovvatikhah, F., Ahmedy, I., Noor, R. M., & Munir, M. U. (2025). A systematic review of AI-based techniques for automated waste classification. Sensors, 25(10), 3181. https://doi.org/10.3390/s25103181
7. Chomicki, A., Wójcik, F., & Dudycz, H. (2025). Assessing the impact of dataset quality on the performance of artificial intelligence models in automatic waste classification. Procedia Computer Science, 270, 1061–1070. https://doi.org/10.1016/j.procs.2025.09.227
8. Majchrowska, S., Mikołajczyk, A., Ferlin, M., Klawikowska, Z., Plantykow, M. A., Kwasigroch, A., & Majek, K. (2021). Waste detection in Pomerania: Non-profit project for detecting waste in environment. arXiv. https://arxiv.org/abs/2105.06808
9. Masand, A., Chauhan, S., Jangid, M., Kumar, R., & Roy, S. (2021). ScrapNet: An efficient approach to trash classification. IEEE Access, 9, 130947–130958. https://doi.org/10.1109/ACCESS.2021.3111230
10. Proença, P. F., & Simões, P. (2020). TACO: Trash annotations in context for litter detection. arXiv. https://arxiv.org/abs/2003.06975
11. Vo, A. H., Hoang Son, L., Vo, M. T., & Le, T. (2019). A novel framework for trash classification using deep transfer learning. IEEE Access, 7, 178769–178779.
12. Huang, K., Lei, H., Jiao, Z., & Zhong, Z. (2021). Recycling waste classification using vision transformer on portable device. Sustainability, 13(21), 11572. https://doi.org/10.3390/su132111572
13. Karthikeyan, N. (2022). Review of deep transfer learning models for image classification. International Journal of Recent Contributions from Engineering, Science & IT, 10(1), 17–28. https://doi.org/10.3991/ijes.v10i01.29783
14. Lin, W. (2021). YOLO-Green: A real-time classification and object detection model optimized for waste management. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 51–57). IEEE. https://doi.org/10.1109/BigData52589.2021.9671821
15. Dipo, M. H., Farid, F. A., Mahmud, M. S. A., Momtaz, M., Rahman, S., Uddin, J., & Karim, H. A. (2025). Real-time waste detection and classification using YOLOv12-based deep learning model. Digital, 5(2), 19. https://doi.org/10.3390/digital5020019
16. Arishi, A. (2025). Real-time household waste detection and classification for sustainable recycling: A deep learning approach. Sustainability, 17(5), 1902. https://doi.org/10.3390/su17051902
17. Mewada, D., Agnew, C., Grua, E. M., Eising, C., Denny, P., Heffernan, M., Scanlan, A., et al. (2024). Contamination detection from highly cluttered waste scenes using computer vision. IEEE Access, 12, 129434–129446. https://doi.org/10.1109/ACCESS.2024.3456469
18. Stadlhofer, A., & Mezhuyev, V. (2023). Approach to provide interpretability in machine learning models for image classification. Industrial Artificial Intelligence, 1, 10. https://doi.org/10.1007/s44244-023-00009-z
19. Nahiduzzaman, M., Ahamed, M. F., Naznine, M., Karim, M. J., Kibria, H. B., Ayari, M. A., Khandakar, A., Ashraf, A., Ahsan, M., & Haider, J. (2025). An automated waste classification system using deep learning techniques: Toward efficient waste recycling and environmental sustainability. Knowledge-Based Systems, 310, 113028. https://doi.org/10.1016/j.knosys.2025.113028
20. Arthur, M. P., Shoba, S., & Pandey, A. (2024). A survey of smart dustbin systems using the IoT and deep learning. Artificial Intelligence Review, 57, 56. https://doi.org/10.1007/s10462-023-10646-6
21. Kunwar, S., Owabumoye, B. R., & Alade, A. S. (2024). Plastic waste classification using deep learning: Insights from the WaDaBa dataset. arXiv. https://arxiv.org/abs/2412.20232