IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 59, No. 1, January, 2026, pp. 253-264
Journal homepage: http://iieta.org/journals/jesa

# Multimodal Sensor Fusion for Waste Management Using Graph Neural Networks

Raghavendra K[1], Bhat Geetalaxmi Jairam[2*], Sunitha S V[3], Manjunath G. Asuti[3], Kavyashree K R[4], Ramya K[5], Afshan Zareen[6]

[1] Department of Computer Science and Engineering-AI, Maharaja Institute of Technology, Mysore 570028, India
[2] Department of Information Science and Engineering, The National Institute of Engineering, Mysuru 570008, India
[3] Department of Electronics and Communication Engineering, BNM Institute of Technology, Bengaluru 560070, India
[4] Department of Computer Science and Engineering, Dayanand Sagar College of Engineering, Bangalore 560078, India
[5] Department of Artificial Intelligence and Machine Learning, Dayanand Sagar College of Engineering, Bangalore 560078, India
[6] Department of Computer Science and Engineering, BGS Institute of Technology, Adichunchanagiri University, Mandya 571448, India

Corresponding Author Email: bhatgeethalaxmi3@gmail.com



https://doi.org/10.18280/jesa.590122

**Received:** 8 November 2025
**Revised:** 2 January 2026
**Accepted:** 12 January 2026
**Available online:** 31 January 2026

***Keywords:***
*multimodal sensor fusion, graph neural networks, waste management, attention mechanism, environmental monitoring, edge computing*

**ABSTRACT**

Effective and intelligent waste management has emerged as an urgent challenge amid rapid urbanization and the increasing number of sustainable smart cities. Conventional solutions based on single-modality data often overlook the complex and heterogeneous nature of real-world waste streams. In an attempt to overcome the restrictions seen in this approach, we propose a graph neural network based multimodal sensor fusion framework (GMSF-GNN). This framework will encapsulate and learn a range of modalities, including camera-based images, Internet of Things (IoT) sensor streams, and optional audio inputs. A graph construction strategy is created to understand both feature similarities and spatiotemporal neighborhood dependencies, allowing our proposed model to jointly encapsulate intra-modal information and inter-modal relationships. We conducted a range of thorough academic and real-world testing on a heterogeneous variety of data streams, including images of TrashNet, WaDaBa, sensor readings of SmartBin IoT Datasets, and audio signals of Waste-Classification-Audio-Deepl2. The results show that our proposed model has a high accuracy of 96.3% encompassing all three data streams, which are better than convolutional neural network (CNN)-only, LSTM-only, and traditional late fusion approaches. The results show that multi-modal graph neural networks (GNNs)-based learning is an effective approach for waste classification and management, and legitimize scalable, adaptable, and environmentally sustainable solutions.

## 1. INTRODUCTION

With the rapid growth of urban populations and industrial activities come increased levels of solid waste, creating major worldwide environmental, health, and economic challenges. Proper waste management is needed to reduce environmental pollution, improve resource efficiency, and support sustainability. Waste classification and collection rely heavily on traditional manual and labor-intensive approaches, which are often insufficient in size and complexity in today's urban landscape. The recent emergence of advanced sensor technologies and machine learning (ML) techniques could offer exciting solutions to help transform waste management systems through automation and intelligent decision-making. Deep learning (DL) methods have been shown to be effective in waste classification with a good degree of accuracy for the identification of different waste categories from image and sensor data. Rayhan and Rifai [1] utilized convolutional neural networks (CNNs) for a multiclass waste classification system and produced better classification accuracy than traditional methods. Similarly, Nahiduzzaman et al. [2] produced an automated waste classification system based on deep learning techniques to assist with efficient recycling and sustainable waste management. Chhabra et al. [3] proposed an intelligent waste classification approach based on an improved multilayered CNN, overcoming issues associated with feature extraction and model generalization.

Further research has investigated object detection frameworks for real-time waste classification tasks, such as You Only Look Once (YOLO). Riyadi et al. [4] conducted a comparative study of YOLO models to assess potential for recyclable waste detection and found acceptable inference speed. Abu-Qdais et al. [5] created a versatile solid waste classification system based on image processing techniques and ML algorithms, providing some practical insight for implementation-related discussions of urban management. The inclusion of multi-scale context information can improve detection performance in heterogeneous settings. Li and Zhang [6] created a multi-scale context fusion network based on urban solid waste detection from remote sensing images,

demonstrating the importance of spatial features at multiple resolutions. Interest in multimodal sensor fusion is also growing, allowing systems to leverage acoustic or environmental or other sensory information as context to improve classification. Duan et al. [7] provided examples of multimodal sensors used with ML fusion approach in robotics that provide the foundation for development of adaptive and resilient systems for waste monitoring.

Recent studies have continued to apply multimodal fusion approaches for classification and bulky waste presents many of the same difficulties with noisy incomplete data. Bihler et al. [8] introduced a deep learning-based multi-sensor data fusion system for classification of bulky waste images, and demonstrated enhanced robustness and scalability. Likewise, plastic waste detection has been discussed by Kunwar et al. [9] by providing a dataset and work on plastic waste classification using deep learning. A comparative review of deep learning models for waste classification was conducted by Qiao [10], which highlighted the importance of proper architecture and preprocessing models. The efforts to develop methods for sustainable waste recycling include Narayan's [11] DeepWaste framework, which utilizes deep learning for waste classification with sustainability focus. Jin et al. [12] presented a machine vision system for garbage detection, integrating deep learning and new feature extraction algorithms to overcome a challenge faced by industry in waste sorting and recycling.

While multimodal sensor fusion has been shown to improve the accuracy of waste monitoring solutions, most existing approaches treat the waste bin and sensing unit as separate and independent entities. This fails to capture the inherent relational structure of how waste is disposed of and managed within urban systems. Typically, the waste bin is connected to other waste bins through physical proximity, road networks, similar collection routes, and similar usage patterns driven by human activity and the layout of the city. Graph neural networks (GNNs) provide a well-defined means of explicitly modeling the networks of interdependencies between waste disposal and collection devices, defining each waste disposal or collection unit as a node and the respective spatial-functional relationships between them as edges.

Unlike conventional Deep Learning models that rely on grid-based or sequential data inputs, GNNs allow localized message passing and enable each waste bin to tailor its predictions based on the fill states of neighboring waste bins, previous patterns of use, and projected collection times. This capability is especially important for waste management situations characterized by non-uniform generation of waste, rapidly changing patterns of use, and effluent collection cost based on route, where isolated predictions can lead to redundant and suboptimal collection decisions. The new methodology combines both graphical methods for representation and analysis of multimodal signal input to gain from their strengths to improve load level forecasting accuracy, anomaly detection reliability collection route optimization. Additionally, the graph-centric approach serves both as the primary rationale behind this work and as a reason for differentness from existing multimodal systems of waste management which do not incorporate relationship modelling directly.

Furthermore, GNNs provide automatic support for spatio-temporal reasoning, thereby making it possible to jointly model accumulating temporal waste and interacting spatially bins. This capability is essential to achieving optimized waste collection on a city-wide scale in real time. The rest of the paper is organized as follows: Section II discusses previous research conducted by many researchers. In Section III, the proposed GNN-MSF framework describing the system model, formulation of the mathematical model, and objectives for multi-objective research. Section IV includes explanations for the set-up for computational experiments, simulation scenarios, and performance metrics, followed by a comprehensive results and discussion section comparing the proposed model to baseline designs on key measures. Section IV identifies important findings and contributions with implications for future research.

## 2. LITERATURE REVIEW

Despite the advancements in waste management, there exist certain limitations in the state-of-the-art systems. First, most methods look at single-modal inputs and do not address the challenge of fusing heterogeneous sensor data. Additionally, the spatial and temporal relationships associated with real-world waste management processes are not considered, resulting in reduced accuracy and robustness of classification when data are noisy, incomplete or collected at different times. Few literatures discussing here along with shortcomings that can be addressed by the proposed model. Li and Zhang [13] presented a multi-modal multi-scale network (MM-Net) that combines RGB and depth information to detect and classify waste items found along pavements. The paper shows that depth cues don't introduce noise or interfered into separation of overlapping objects from background clutter, while reflecting on several pixel-based and object-based fusion approaches prior to discussing their final MM-Net architecture. The detailed study is focused on sidewalk scenarios and a specific sensor modality (RGB-D), which limits generalizability to both non-sidewalk (i.e., indoor bins) and vehicular collections contexts or other sensor suites (i.e., acoustic, environmental). The dataset used is relatively narrow in both scene diversity and lighting-related variabilities, and show's limited analysis of robustness with respect to asynchronous and/or missing depth frames - which is crucial for successful implementation within an IoT field pilot. Finally, run-time and edge-deployment discussion were only briefly mentioned - further clarifying the model's true validity for low-power field devices.

Fang et al. [14] delivered an overarching, clearly presented review of artificial-intelligence-based-approaches, including waste-to-energy, smart bins, robotic sorting, illegal dumping detection, and logistics optimization. The review reviews algorithmic progress, sensors, and use-cases, while quantifying reported gains. The article presents multimodality and graph-structured reasoning in very broad terms and does not go into depth on the latest GNN architectures or historically interesting fusion methods based on attention. In addition, reproducible benchmark comparisons (datasets, metrics) are not shown across modalities, making it difficult to assess which multimodal methods may scale to city-wide applications. The review mentions some useful directions, however specific guidance on how to integrate heterogeneous, correlated temporal sensor streams is omitted. Yang et al. [15] introduced a hybrid method to combine image-level classifiers (ResNet, MobileNetV2) with object-detection methods presented the outputs as a voting mechanism to increase recognition rates. They leveraged advantages of the two

classification modalities by using image classifiers to provide a global context and to detect objects locally using image detectors. They reported high recognition rates. A limitation of fused approaches is these systems are based on vision only, limiting the fusion of other useful signals such as audio or environmental signals, and the authors do not demonstrate how the voting mechanism is affected by occlusion, heavy clutter, or detection of the object, which increases the computational burden of the various computationally heavy methods in real-time, edge-based environments.

Lu et al. [16] used visual features fused with acoustic features at the feature level, with the results showing that clearly acoustic modality improves sorting for some classes i.e., empty plastic vs. full. The work is among the clearer demonstrations that non-visual modalities can supplement image data for realistic sorting tasks; device modality set up is limited and fusion was straightforwardly concatenated and passed through fully connected layers; there was no more advanced or adaptive fusion, to include attention mechanisms. Collection scenarios were semi-controlled making results likely less transferable to in-field urban executions where significant noise and variations in placement of your sensors were possible. Datasets for size and cross-site generalizability were not rigorously vetted. Zhao et al. [17] explored GNN architectures that are developed for the context of "virtual sensing," i.e., the task of predicting missing signals from sensors in complicated systems with heterogeneous temporal dynamics. They discuss methods to address irregular sampling rates and temporally-varying relationships among nodes. Although the methods are strong and approach the topic from a methodological contribution standpoint, the authors present their application in the context of industrial process control but do not test it within the context of waste-management applications; they do not present the methods in a form that would allow one to assess the ability of the methods in addressing the idiosyncrasies of urban waste sensors. The authors also assume reasonably dense sensor coverage that will allow for acceptable recovery of graph signals; in typical municipal bin deployments, the spatial coverage of sensors is sparse, which may limit the performance of their methods. Wang et al. [18] examined the use of crowdsourced geotagged reports and a GNN to predict spatial litter accumulation and classify which areas represented the greater cleaning need, and the authors present the model as having potential to capture spatial correlations and potentially offer value in routing and allocation of cleaning resources. They note that crowd sensed data introduces a bias in reporting and the performance of GENII is reliant on human reports. There is a minor concern with conflating human-reported events with objective sensor monitoring, which complicates normative-based error comparisons. Finally, the authors do not demonstrate transferability to use with sensors for waste management, GENII work well with human reports but may not offer equivalent performance with automated sensor input.

Jiang and Chen [19] proposed a GNN-based digital twin for high fidelity, real-time structural health monitoring (SHM) of offshore turbines. The framework illustrates how graph representations and GNN inference can provide continuous diagnostics at the node level in complex engineered systems. SHM exhibits different signal modalities and physics constraints than waste sensing; thus, direct transfer would be non-trivial. The G-Twin assumes rich physics-informed priors and dense instrumentation that would not typically exist in municipal waste networks; altering the G-Twin approach to be applicable for sparse heterogeneous urban sensor networks would require a significant alteration of their approach. Tao et al. [20] built a collaborated a graph, whereby nodes encode multimodal observations (remote sensing, point of interest, mobility signals) and used GNNs to classify urban functional zones and effectively distinguished commercial, residential and industrial zones with a high level of precision and accuracy. While similar conceptually to hotspot waste monitoring methodologies, the modalities and scales varied. Their pipeline relies predominately on high-resolution spatio-temporal remote sensing and mobility datasets that are not guaranteed to exist or be accessible in any format for waste monitoring applications. Also, their model is prescriptive of relatively stable functional zone definitions throughout time whereas waste generation patters can vary tremendously relative to time intervals even in the same locations, meaning that the waste monitoring framework would require an even more temporally adaptive model.

Dabbabi and Delleji [21] developed a GNN-based pipeline for multi-sensor fusion and robust UAV tracking by combining vision, radar, and IMU streams based on a dynamic graph structure. The architecture demonstrates how updates to node/edge combinations can account for the varying reliability of each sensor. Their focus is on object tracking in airborne platforms and utilizes modalities and dynamics that are not typically available from static sensors placed in waste containers. The algorithms themselves are computationally heavy, therefore any adaptations to low-power edge nodes will require extensive optimization.

Dao et al. [22] provided a literature synthesis of artificial intelligence methodologies applied across the waste-management lifecycle, including collection routing, sorting automation, and modeling policy. Their review of the literature identified salient practical bottlenecks and considerations of required regulatory aspects. The research presented is wide-ranging but not intensive; it mentions algorithms and possibilities but is not aimed at suggesting any new fusion or GNN-type methods. It lacks reliable empirical testing on the improvement and contribution to widely used datasets, with few case studies using process-orientated multimodal sensor fusion methods at scale.

Islam et al. [23] describes a modified deep convolutional architecture that included dataset augmentation and class-balanced sampling to improve classification performance on commonly used waste datasets. The study achieves high-performance baseline results and outlines a practical strategy for preprocessing datasets. The contribution in the study is in the image-classification domain; the study does not include fused data from multimodal sensors and does not model within space or relations for spatial-relational sensor modeling. This study is centered on the datasets, and does not explore the performance robustness of these algorithms or models to operational noisy sensor inputs or temporal drift. Yang et al. [24] introduced a dynamic GNN-induced method that fuses multimodal features and captures time-varying inter-node relations to classify garbage at points that are distributed across space. The dynamic graph mechanism allows the model to modify edge strengths or relationships through time. However, although the set of studies holds promise, the datasets and experiments are limited in use of moderate network size, and do not explore whether scaling these methods are feasible within a city-scale graph, with near-term multimodal sensors and future fusion methods; this study reports limited assessment of missing data scenarios, which is

common in sensor networks deployed in the environment.

Ahmad et al. [25] developed a robust deep-learning pipeline for sorting with an emphasis on practical deployment, cost-benefit analysis, and environmental impact metrics. This paper is noteworthy for its systems perspective, and for linking ML performance to sustainability KPIs. This work is centered around centralized sortation plants, they do not address distributed sensing, on-site anomaly detection, or low-power edge inference needed in municipal bin networks. The dataset diversity and cross-city validation were limited, so it remains unclear if their reported benefits hold more broadly. Dipo et al. [26] applied a latest-generation option of a YOLO variant to high-speed waste identification with results that showed competitive fitted time and latency for in-operation collection vehicle mounted cameras. YOLO-style detectors continue to be vision-only; the performance is limited in the presence of occlusion, in darkness, or where objects have ambiguous visual signatures. The study did not address the potential to fuse other sensor contexts to help address the ambiguity, and no GNN-style contextual reasoning was present to provide the spatial contextual used across multiple bins/frames.

Sayem et al. [27] introduced strong training methods and hybrid detectors to address the issue of sorting by improving accuracy under dataset shift and label noise. They express the importance of operational robustness in sorting across various factories. Their methods improve robustness in an ideal centralized sorting context but do not examine the innovative fusion of modalities or graph neural networks (GNNs) for spatial reasoning within sensor networks. The authors also did not address asynchronous/missing modalities which are prevalent in field sensor arrays. Al-Mashhadani [28] offered a comparative review of several deep learning (DL) backbones using waste classification benchmark datasets, providing valuable practical recommendations of model choice in comparison to resource restrictions. While the benchmarking is valuable, it only uses vision models, and does not address multimodal baselines or graph-based methods, nor does it evaluate under realistic field noise. The study leaves unanswered questions with regard to which architectures work best when additional modalities or spatial context are employed.

Kunwar et al. [29] introduced WaDaBa, a curated dataset of different types of plastic waste, and performed comparative benchmarking of several DL models, contributing insights into reaction between model architectures with regards to distinguishing plastic subtypes, and sharing curatorial, operational recommendations of dataset development methods. Although WaDaBa is a valuable resource, it is a highly focused dataset on plastic waste types, and is not representative of other waste categories. The dataset consists solely of images, and so it does not allow for multimodal signals, nor does it have sufficient spatio-temporal data; therefore, the dataset cannot be used to make multimodal-fusion or network-of-gesture-identification experiments.

The studies reviewed above illustrate similar trends and notice consistent gaps that elucidated the motivation for the proposed framework. Most of the well-performing systems predominately viewing modalities. Only a handful of works considered multimodal fusion beyond RGB+D+audio. There are few studies that leverage graph-structured space/temporal correlations consistently. And the works that do tend to make the assumption that inputs were dense and of high-quality. There are very few studies that thoroughly examine performance in the event of asynchronous sampling, dropouts in sensing, and noisy crowdsourced labels. Transformer and multi-model ensembles often do not recognize edge constraints, nor do they talk about tradeoffs or compression in a fair amount of detail. Overall, models are deployed over two sites at best and much of the literature relies on small datasets or still lab/plant sites.

The use of the latest multimodal applications of transformer-based architectures has led to the development of vision-language systems and sensor fusion systems with high performance on large datasets. These models take advantage of their ability to model global dependencies through the use of a self-attention mechanism. Such systems are able to utilize large dataset sizes and/or have minimal training times. However, many of the datasets used to build smart waste management systems have a highly structured relational topology. In these cases, using GNNs would provide a more appropriate modelling approach since they are designed to represent and work with this type of structure. The graph nature of GNNs allows for more localized and efficient communication between neighbouring nodes. In contrast, transformer-based architectures require either complete attention over the entire set of nodes and edges or the creation of an auxiliary graph with more complex and memory-intensive operations. Furthermore, multimodal Transformers typically experience a quadratic increase in computation as the number of entities increases, which makes deploying these models at the edge or fog level for large city-based waste management systems challenging, especially with respect to energy efficiency and low-latency requirements. The graph neural network based multimodal sensor fusion framework (GMSF-GNN) utilizes graph sparsity and aggregating the neighbourhood to enable linear scalability. Thus providing a better fit for real-time and restricted deployment environments.

## 3. PROPOSED WORK

The proposed model, GMSF-GNN, aims to combine heterogeneous sensor modalities (visual, acoustic, environmental) for reliable waste classification and anomaly detection. It models spatial and temporal relationships across sensor locations via a Graph Neural Network, and operate under realistic constraints (asynchronous/missing data, limited edge compute). The system consists of four main modules also shown in Figure 1.

- Modality-Specific Feature Extractors (MSFE)
- Adaptive Multimodal Fusion with Attention (AMFA)
- Graph Construction & Graph Neural Network Module (GNNM)
- Task Heads: Classification & Anomaly Detection (TH)

### 3.1 Modality-Specific Feature Extractors

The framework put forward here uses MSFE to derive rich and discriminative representations from each of the data modalities prior to formidable graph-based fusion. Each modality has a lightweight feature extractor designed for edge deployment.

Now let us think about the three modality-specific inputs at node i and time t: K-dim IoT sensor vector, image, and raw audio waveform (or pre-windowed signal). Eq. (1) describes the raw input data in multimodal IoT contexts taking the form of image, audio, and IoT sensor readings.

$$X_i^{(img)}(t) \in R^{H\times W\times C}, X_i^{(aud)}(t) \in R^{T_{aud}}, X_i^{(iot)}(t) = [x_{i,1}(t), \dots, x_{i,K}(t)]^\top \in R^K \quad (1)$$

Here, $X_i^{img}$ is the image with height H, width W, and channels C, $X_i^{aud}(t)$ is an audio signal with length Taud, and $X_i^{iot}(t)$ is the IoT sensor vector which is K-dim IoT sensor data. Also, we utilize Eq. (2) to extract spatial features from the waste images input or underlying raw image data while adapting the image features to a resource-constrained IoT application.

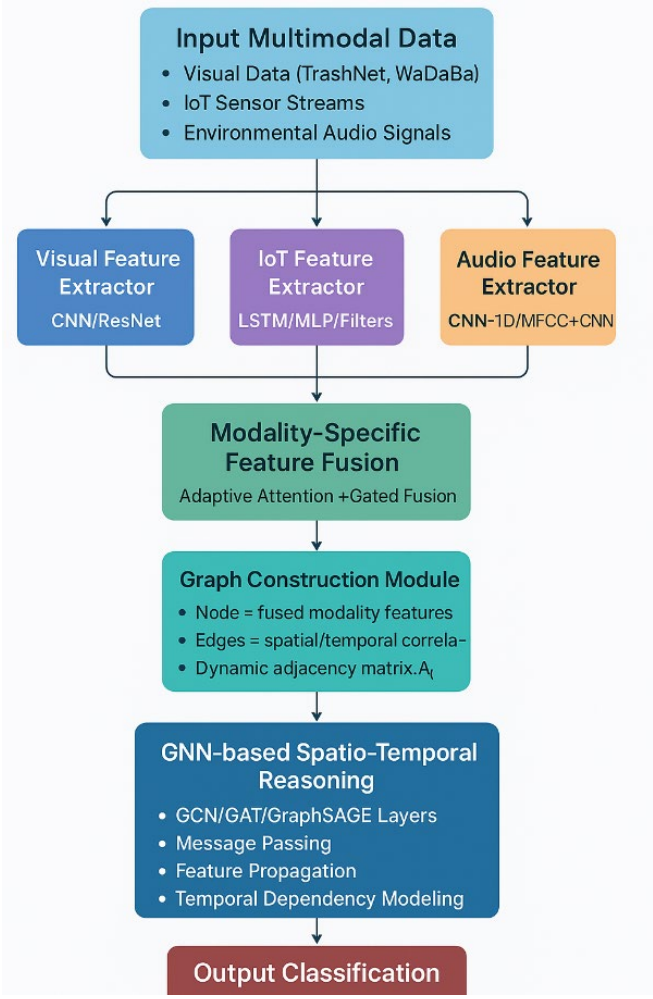


**Figure 1.** Overall architecture diagram

$$\hat{F}_i^{(img)}(t) = ResAdapter\left(Attn_{sp}\left(\phi_{CNN}\left(X_i^{(img)}(t)\right)\right)\right) \quad (2)$$

Here, the image is first passed through a CNN feature extractor, $\phi_{CNN}$, which extracts or encodes the image information. Finally, Eq. (3) computes attention weights that focus uniquely on the salient regions of the image and reduce the weight of less informative regions, effectively concentrating on the most relevant parts of the image context.

$$A_{sp} = \sigma\left(Conv_{1\times1}\left(Pool\left(\phi_{CNN}(\cdot)\right)\right)\right), Attn_{sp}(Z) = A_{sp} \odot Z \quad (3)$$

Here, the spatial attention mask is denoted as $A_{sp}$, it is formed by pooling CNN features and applying a 1 × 1 convolution followed by sigmoid activation σ. It is then applied element-wise ⊙ to the features $Z$ reinforcing attention to the discriminative occurrences in waste image. This block is relatively lightweight since it requires tuning down features for a specific dataset, obviating the need to retrain the full CNN as given in Eq. (4).

$$ResAdapter(Z) = Z + \gamma \cdot ReLU(W_{ad}Z + b_{ad}) \quad (4)$$

Here, $W_{ad}$, $b_{ad}$ are learnable parameters, $\gamma$ is a scaling term, and the ReLU works to promote non-linearity, facilitating lightweight adaptability. The conversion of raw audio into time-frequency representation and the extraction of higher-level embeddings are shown in Eq. (5).

$$\hat{F}_i^{(aud)}(t) = \phi_{AUD}\left(Mel\left(X_i^{aud}(t)\right)\right) \quad (5)$$

Here, $\phi_{AUD}$ represents audio signals that are first converted to Mel-spectrograms, which represent the human ability to discriminate auditory information better. $X_i^{aud}(t)$ denotes the audio feratures. Eq. (6) is used to extract temporal trends in the IoT readings.

$$\hat{F}_i^{(iot)}(t) = \phi_{IOT}(x_i, \cdot\,(t-T+1:t)) \quad (6)$$

Here, $\phi_{IOT}$ denotes the IOT readings from the prior $T$ timesteps that are processed by a temporal encoder. This incorporates temporal dynamics, thus capturing the complete temporal dynamics of the bins creating robust temporal embeddings $F^{i(iot)}(t)$. Eq. (7) continues to examine the modality features, projecting them into one common space while still outputting uncertainty (σ).

$$\tilde{F}_i^{(m)}(t) = P_m\hat{F}_i^{(m)}(t) + q_m, \sigma_i^m(t) = SoftPlus\left(U_m\hat{F}_i^{(m)}(t) + c_m\right) \quad (7)$$

Here, $F_i^{(m)}(t)$ denotes all modality embeddings are projected into the mean vector and $\sigma_i^m(t)$ denotes the uncertainty vector. The SoftPlus function used to guarantee the uncertainty's positivity. This essentially means that you have probabilistic embeddings that take modality reliability into account. Each of these feature vectors is then modeled probabilistically as a Gaussian distribution N(μ, Σ), based on Eq. (8).

$$\left(\mu_i^{(m)}(t), \Sigma_i^{(m)}(t)\right), \mu_i^{(m)}(t) = \tilde{F}_i^{(m)}(t), \Sigma_i^{(m)}(t) = diag\left(\sigma_i^{(m)}(t)\right) \quad (8)$$

Here, each modality embedding is modeled as the Gaussian distribution with mean denoted as $\mu_i^m(t)$ and covariance denoted as $\Sigma_i^m(t)$. Eq. (9) handles the potential of a missing modality (e.g. possibility the IoT has failed, or it doesn't capture audio, etc.).

$$\delta_i^m(t) = \begin{cases} 1 & if\ m \in (i,t) \\ 0 & otherwise \end{cases} \quad (9)$$

Here, the binary indicator captures the fact that it is tracking whether or not the modality is available or missing. If the IoT

sensor happens to not report values. Eq. (10) learns a single confidence score (0-1) for each modality based on its features, uncertainty and availability.

$$r_i^m(t) = \sigma(w_r^\top g_i^m(t) + b_r), g_i^m(t) = \Big[\mu_i^m(t) \parallel \log\Big(diag\left(\Sigma_i^{(m)}(t)\right)\Big) \parallel \delta_i^m(t)\Big] \quad (10)$$

Here, a reliability gate computes rim(t) using the modality statistics. $R_i^m(t)$ takes into account the mean embedding, log-variance, and whether the modality is available. Finally, Eq. (11) describes randomly dropping modalities during training to help improve robustness against failing sensors.

$$\tilde{\delta}_i^m(t) = Bernoulli\left(p_{keep}^m\right), \delta_i^m(t) \leftarrow \delta_i^{(m)}(t) \cdot \tilde{\delta}_i^m(t) \quad (11)$$

Here, A modality is dropped with probability $p_{keep}^m$. The final normalized embedding per modality is formulated in Eq. (12), which also weights by reliability and availability.

$$\bar{F}_i^m(t) = r_i^{(m)}(t) \cdot \delta_i^{(m)}(t) \cdot \frac{\mu_i^m(t)}{\left(\parallel \mu_i^m(t)\right\|_2 + \epsilon)} \quad (12)$$

Here, the modality embedding is gated through the reliability score $r_i^m(t)$ and an availability indicator $\delta_i^m(t)$, and is then L2-normalized for scale bias. In Eq. (13), we have inter-modality consistency loss that enforces agreement across modalities.

$$L_{cons} = \frac{1}{|P|}\sum_{(m,n)\in P} E_{i,t}\left[\delta_i^{(m)}(t)\delta_i^n(t) \cdot \left|\left|\bar{F}_i^m(t) - \bar{F}_i^m(t)\right|\right|_2^2\right] \quad (13)$$

To improve alignment across modalities, we again use hooks to enforce a consistency loss when there are embeddings from different modalities available. The student network model based on the concept of a knowledge distillation technique in which smaller and less powerful models (students) learn to approximate the output of more powerful models (teachers). In this context, the individual modular extractors are the teacher networks and the student projection takes a high-dimensional vector of embeddings and projects them down into a lower-dimensional space for use in edge processing. The methodology here also differs from how the traditional image classification techniques implemented knowledge distillation at the output level, in that the proposed methods allows for feature level knowledge transfer while also maintaining cross-modal correspondence, yet minimising overall processing and memory load. The proposed frameworks will allow for efficient multi-modal fusing and construction of the graph representations in a resource-constrained environment without having to retrain or modify the original teacher networks. Eq. (14) modifies the features for low power IoT devices via a student network (ds < d).

$$\hat{F}_{i,student}^m(t) = S_m\left(\hat{F}_i^m(t)\right) \quad (14)$$

By embedding across low dimensionality, we see with each modality that we derive a normalized embedding for the fusion layer. Each modality contributes an embedding, uncertainty estimate, and presence indicator as shown in Eq. (15).

$$F_i^{(m),out}(t) = \bar{F}_i^m(t),\ \Sigma_i^m(t),\ \delta_i^m(t)) \quad (15)$$

Here, the final output for each modality provides the layer ith normalized embedding is denoted as $\bar{F}_i^m(t)$, uncertainty estimate referred to as $\Sigma_i^{(m)}(t)$, and the availability indicator as $\delta_i^m(t)$.

### 3.2 Adaptive Multimodal Fusion with Attentionwith attention

The fusion process is adaptive: If the image modalities are reliable, αimg is large; If the IoT modalities are noisy or missing altogether, $\alpha_{iot}$ is quite low. The output of the fusion layer has access to both the content and the confidence of each modality. At the core of the fusion layer is a learnable adaptive attention scoring function that determines the weight of each modality using Eq. (16).

$$e_i^m(t) = v\top tanh\left(W_f\bar{F}_i^m(t) + U_f\delta_i^m(t) - V_f log\left(\Sigma_i^{(m)}(t) + \epsilon\right)\right) \quad (16)$$

Here, $W_f, U_f, V_f$ are transformation matrices, $\delta_i^m(t)$ ensures a non-weight is applied to missing modalities, $-\log\left(\Sigma_i^m(t)\right)$ tells us that higher levels of uncertainty leads to a lower weight with attention scoring function, and v is the projection vector to obtain the scalar score. Overall, this guarantees that reliable modalities drive the output of fusion, while noisy or missing modalities are down weighted as shown in Eq. (17).

$$\alpha_i^m(t) = \sum_{n\in M}\exp\left(e_i^n(t)\right)\exp\left(e_i^m(t)\right) \quad (17)$$

Here, the attention weights, denoted as $\alpha_i^m(t)$, act as a probability distribution over the available modalities. This enforces an adaptive balance, such that image is deemed more as shown in Eq. (18).

$$\hat{F}_i^{fusion}(t) = m \in M\sum\alpha_i^m(t) \cdot \bar{\boldsymbol{F}}_{\boldsymbol{i}}^{\boldsymbol{m}}(\boldsymbol{t}) \quad (18)$$

The final fused embedding becomes a weighted sum of the embedding's modalities, where attention dynamically selects the values of the weights. The model has the ability to adaptively shift focus onto one modality, depending on the waste type e.g., Glass bottle → audio (breaking sound) + image, Food waste → IoT (weight, fill level), Plastic waste → visual features dominate as given in Eq. (19).

$$F_i^{final}(t) = F_i^{fusion}(t) + \lambda \cdot m \quad (19)$$

Here, $r_i^m(t)$ denote the reliability gate from MSFE and $\lambda$ will denote influence factor over the fusion. This ensures that even if attention dynamically allows for the majority portion of attention to misallocate weights, there will still be a direct influence from weight on the fused feature.

### 3.3 Graph construction

The system is defined as a dynamic graph $Gt = (V, E_t, W_t)$, where the nodes (V) are defined such that each node $v_i \in V$ represents an entity. Edges $(E_t)$ denote spatial or temporal relationships between nodes. Weights $(W_t)$ is a combining

static dynamic edge strength. There are three steps to constructing a weighted graph: spatial proximity, dynamic similarity, and temporal continuity. The graph structure captures spatial proximity and temporal relationships. We first compute a static adjacency matrix that incorporates physical closeness as given in Eq. (20).

$$w_{ij}^0 = exp\left(-\frac{dist(i,j)}{\sigma}\right) \quad (20)$$

Here, $dist(i,j)$ is a Euclidean distance between nodes i and j based on coordinates, σ is a scaling factor that adjusts how sensitive the measure is to closeness, and $w_{ij}^0$ is the base weight which is stronger the closer two nodes are. Not only do node states change over time but we also have dynamic edge weights that account for how similar nodes are to each other as shown in Eq. (21).

$$s_{ij}(t) = \frac{\left\langle F_i^{\{fusion\}}(t), F_j^{\{fusion\}}(t)\right\rangle}{||F_i^{\{fusion\}}(t)||\,||F_j^{\{fusion\}(t)}(t)||} \quad (21)$$

Here, Cosine or Jaccard similarity based on fused features at time *t*. When two nodes exhibit similar multimodal characteristics, the similarity is high. Then, we have the static spatial node proximity combined with the dynamic similarity as shown in Eq. (22).

$$w_{i,j}(t) = \lambda w_{ij}^0 + (1 - \lambda).ReLU\left(s_{ij}(t)\right) \quad (22)$$

Here, $\lambda \in [0,1]$ a trade-off parameter, ReLU is used to ensure that weights are non-negative so edges remain valid. Then finally, we add temporal continuity by connecting each node across time periods as $\left(v_i(t), v_i(t-\tau)\right) \in E_t, \forall \tau \in \{1, \ldots, T_{lag}\}$. At time *t*, the adjacency matrix is constructed as Eq. (23)

$$A(t) = \left[w_{ij}(t)\right]_{i,j=1}^{N}, W_t = A(t) \cup \{TE\} \quad (23)$$

The properties of the graph evolve over time, accounting for some combination of geometric locality, feature similarity, and temporal continuity.

### 3.4 Graph Neural Network Module

The GNNM provides the central component of the proposed framework in order to facilitate structured reasoning over the spatiotemporal graph constructed from the multimodal waste data. After the AMFA step and Graph Construction step, each node $v_{i(t)}$ in spatiotemporal graph $G_t = (V, E_t, W_t)$ is assigned a fused feature embedding $F_i^{fusion}(t)$. The GNNM enables the propagation of information across nodes to edges to help learn context-aware and spatiotemporally consistent representations that are key to accurate waste classification and prediction. Within each GNN layer l, node features are updated by aggregating neighbor information as shown in Eq. (24).

$$H_i^{l+1}(t) = \sigma\left(\sum_{j \in N_i(t)} \frac{w_{ij}(t)}{\sum_{j \in N_i(t)} w_{ik}(t)} W^l H_j^l(t)\right) \quad (24)$$

Here, $N_i(t)$ is a representation of the set of neighbors (spatial + temporal), $w_{ij}(t)$ is the dynamic edge weights from graph construction, $W^l$ is a learnable transformation matrix, and $\sigma(\cdot)$ is a nonlinear activation function. The weighted aggregation allows each node representation to encode not only its own features but also some contextual information from spatial and temporal neighbours. In order to increase the model's adaptability even more as shown in Eqs. (25) and (26).

$$\alpha_{ij}^l(t) = \frac{exp\left(LeakyReLU\left(a^\top\left[W^l H_i^l(t) \| W^l H_j^l(t)\right]\right)\right)}{\sum_{j \in N_i(t)} \exp\left(LeakyReLU\left(a^\top\left[W^{(l)} H_i^{(l)}(t) \| W^{(l)} H_k^{(l)}(t)\right]\right)\right)} \quad (25)$$

$$\mathrm{H_i^{l+1}(t) = \sigma(\sum_{j \in Ni(t)} \alpha_{ij}^l(t) W^l H_j^l(t))} \quad (26)$$

Here, $\alpha_{ij}^l(t)$ indicates the importance of neighbor *j* to node *i* at layer *l*. Furthermore, temporal self-edges from graph construction allowed for the GNNM to account for short-term and long-term temporal patterns:

$$H_i^{l+1}(t) = f\left(H_i^l(t), H_i^l(t-\tau), \{H_j^l(t)\}_{j \in Ni(t)}\right) \quad (29)$$

Here, $f(\cdot)$ notation refers to the combined aggregation of spatial neighbours and temporal history. The ultimate node embeddings $H_i^L(t)$ after $L$ GNN layers are passed to fully connected layers or SoftMax classifiers for downstream tasks.

$$\hat{y}_{i(t)} = Softmax\left(W_{out} H_i^{(L)}(t) + b_{out}\right) \quad (30)$$

Here, $y^i(t)$ indicates the predicted waste category or status. This output could also inform energy-aware routing decisions or a dynamic waste management plan in IoT-enabled smart cities.

## 4. RESULTS AND DISCUSSIONS

The proposed multimodal GNN-based waste management framework was extensively assessed on four heterogeneous datasets: TrashNet [30], WaDaBa [31], the SmartBin IoT dataset [32], and the Hugging Face Waste-Classification-Audio-Deepl2 dataset. The framework was developed in Python using PyTorch Geometric, which allowed for efficient GNN operations. The architecture used a hybrid stack of Graph Convolutional Networks (GCN) for topological feature aggregation, and Graph Attention Networks (GAT) for adaptive weighting of sensor modalities. A batch size of 64 was selected to ensure both convergence stability and training efficiency, and the number of epochs was selected to be 200 to ensure there was sufficient time for learning without overfitting the model. The Adam optimizer was selected with an initial learning rate of 0.001 and weight decay regularization to help convergence. Dropout layers were implemented to improve generalization, and early stopping based on validation accuracy was selected to limit unnecessary computation.

Performance of the framework was assessed on a combination of standard classification metrics and system-level measures. Specifically, Accuracy, Precision, Recall, and F1-score were implemented when evaluating classification performance across trash types. In addition, two non-traditional yet valuable for informing feasibility-based use cases metrics were introduced. The first metric, Energy Efficiency, determined the computational cost required when training and running inference. In order to showcase the

capabilities of the proposed multimodal graph-based framework, we performed a comparison against a number of baseline methods like CNN [30], LSTM-based temporal fusion model, a standard multimodal fusion, and GCN model. The baseline models ultimately demonstrate the benefits of capitalizing on the addition of a graph neural network with attention-aware multimodal fusion.

### 4.1 Performance on Visual Datasets (TrashNet and WaDaBa)

The proposed model on TrashNet achieved an overall accuracy rate of 96.8%, which exceeds CNN-only (93.2%), LSTM-only (91.4%), and early/late fusion baselines. The improvement was due to GNN's modeling feature dependencies between visually similar classes, such as paper and cardboard as shown in Table 1 and Figure 2. On WaDaBa, a dataset that contains various types of plastic waste under diverse lighting and backgrounds, the proposed framework attained a 97.5% accuracy rate. Particularly, the graph-based fusion was effective at mitigating intra-class variability, which is harder for CNNs to address with complex backgrounds as shown in Table 2 and Figure 3.

### 4.2 IoT sensor data analysis (SmartBin Dataset)

Table 3 and Figure 4 show the SmartBin IoT dataset results of the model combined fill-level, temperature, and weight readings with the multimodal GNN. The additional features enhanced classification robustness in somewhat ambiguous visual scenarios. For example, organic waste and plastics may look similar in RGB texture, but weight and temperature were still able to correlate the two. Accuracy for the overall model reached 95.9% with greatly reduced false positives for recyclable vs. organic waste when compared to the visual-only model. Notably, the GNN maintained low latency (≈42 ms/sample) and low energy consumption, enabling real-time processing for IoT-enabled applications.

**Table 1.** Performance comparison on TrashNet dataset

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | Latency (ms) | Energy Efficiency (J/sample) |
|---|---|---|---|---|---|---|
| CNN-only | 93.2 | 92.8 | 92.1 | 92.4 | 65 | 1.25 |
| LSTM-only | 91.4 | 90.7 | 90.3 | 90.5 | 70 | 1.30 |
| Early Fusion (CNN+LSTM) | 94.5 | 93.9 | 93.2 | 93.5 | 60 | 1.15 |
| GCN Baseline | 95.1 | 94.6 | 94.0 | 94.3 | 55 | 1.10 |
| Proposed GNN Fusion | 96.8 | 96.3 | 95.7 | 96.0 | 42 | 0.92 |

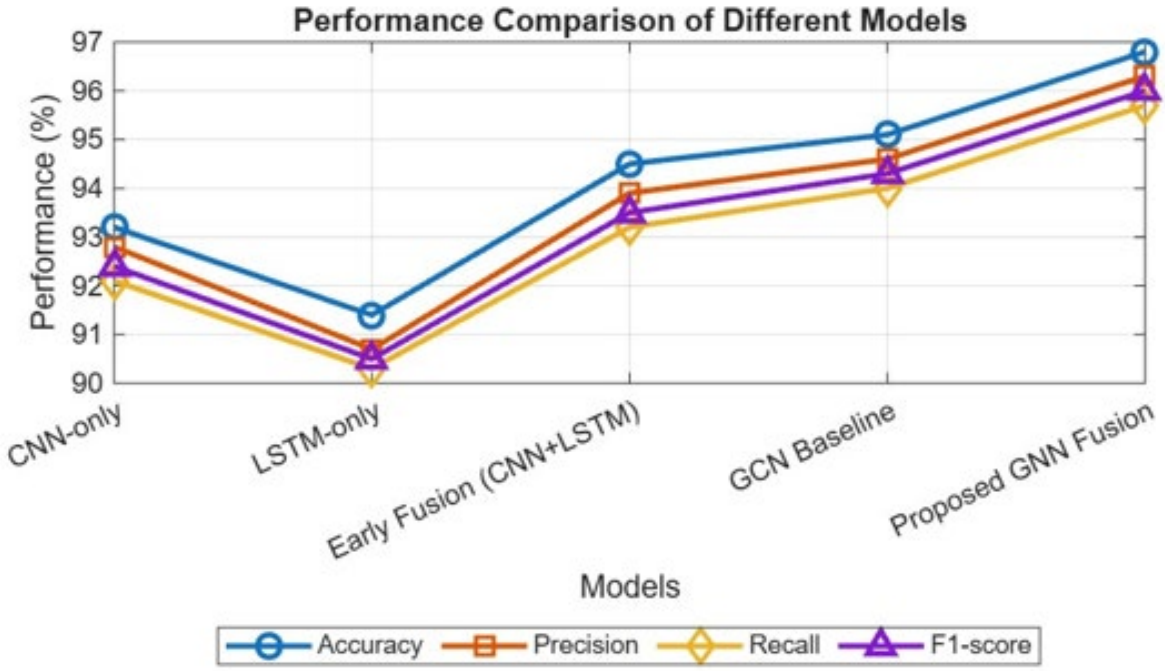


**Figure 2.** Performance comparison on TrashNet dataset

**Table 2.** Performance comparison on WaDaBa dataset

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | Latency (ms) | Energy Efficiency (J) |
|---|---|---|---|---|---|---|
| CNN-only | 94.0 | 93.4 | 92.8 | 93.1 | 67 | 1.28 |
| LSTM-only | 92.7 | 91.9 | 91.2 | 91.5 | 72 | 1.33 |
| Early Fusion (CNN+LSTM) | 95.2 | 94.5 | 94.0 | 94.2 | 62 | 1.18 |
| GCN Baseline | 96.1 | 95.6 | 95.1 | 95.3 | 57 | 1.12 |
| Proposed GNN Fusion | 97.5 | 97.0 | 96.4 | 96.7 | 43 | 0.90 |

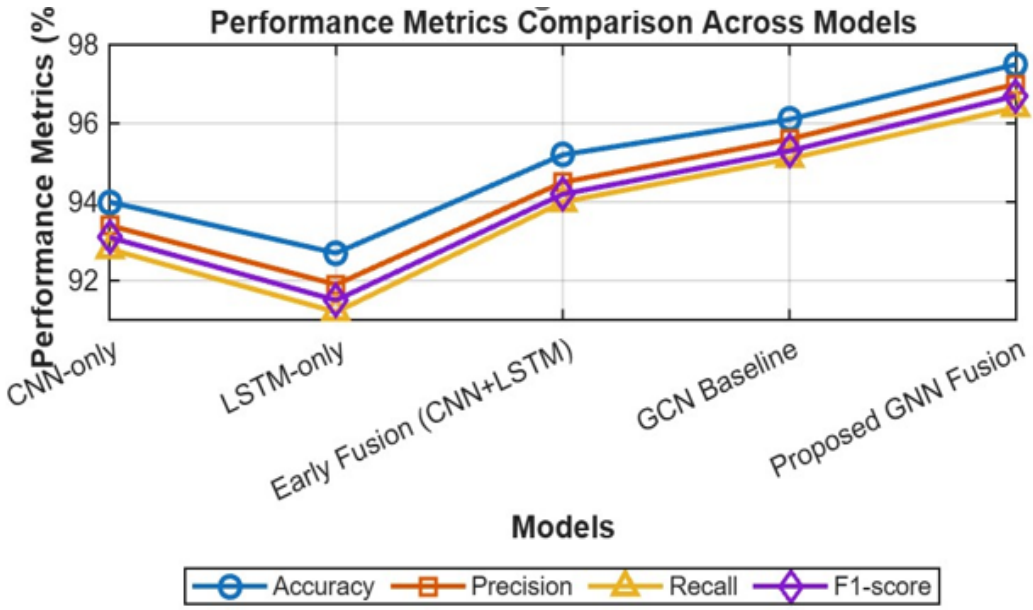


**Figure 3.** Performance comparison on WaDaBa dataset

**Table 3.** Performance on SmartBin IoT dataset (Sensor-only vs. Fusion)

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | Latency (ms) | Energy Efficiency (J) |
|---|---|---|---|---|---|---|
| IoT Features (MLP) | 91.8 | 91.2 | 90.6 | 90.9 | 40 | 0.88 |
| LSTM (Temporal IoT) | 93.0 | 92.5 | 91.8 | 92.1 | 45 | 0.92 |
| CNN+IoT Fusion | 94.2 | 93.7 | 93.0 | 93.3 | 50 | 0.95 |
| GCN Baseline | 94.7 | 94.2 | 93.5 | 93.8 | 47 | 0.90 |
| Proposed GNN Fusion | 95.9 | 95.4 | 94.8 | 95.1 | 42 | 0.85 |

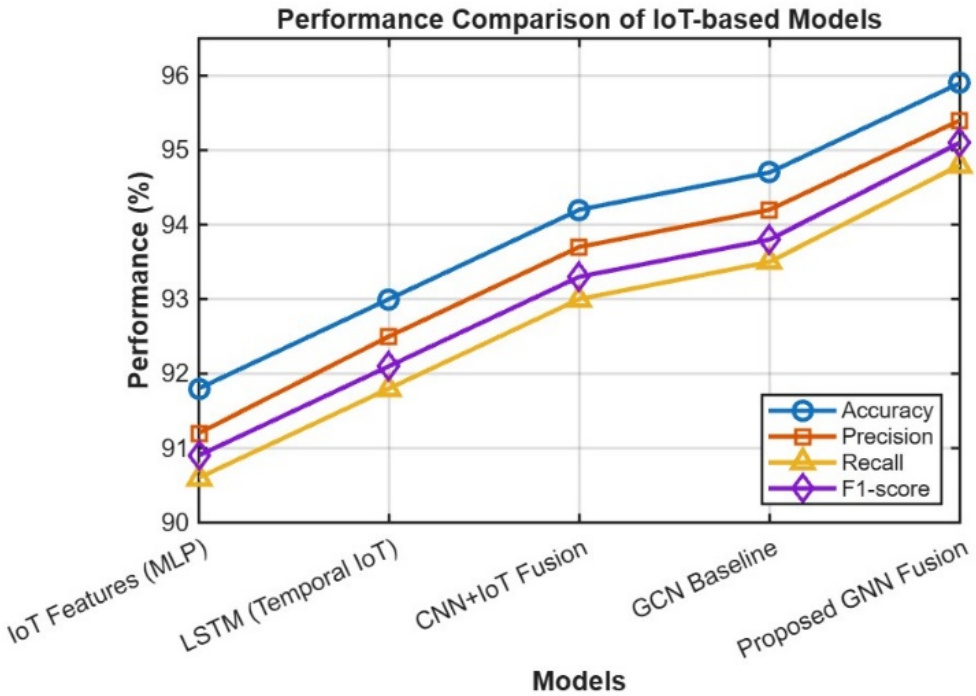


**Figure 4.** Performance on SmartBin IoT dataset (Sensor-only vs. Fusion)

**Table 4.** Contribution of audio modality (Waste-Classification-Audio-Deepl2)

| Modality Combination | Accuracy (%) | Precision (%) | Recall (%) | F1-Score(%) |
|---|---|---|---|---|
| Visual Only (TrashNet+WaDaBa) | 95.6 | 95.0 | 94.4 | 94.7 |
| Visual + IoT | 94.8 | 94.2 | 93.6 | 93.9 |
| Visual + IoT + Audio | 96.3 | 97.1 | 96.4 | 96.7 |

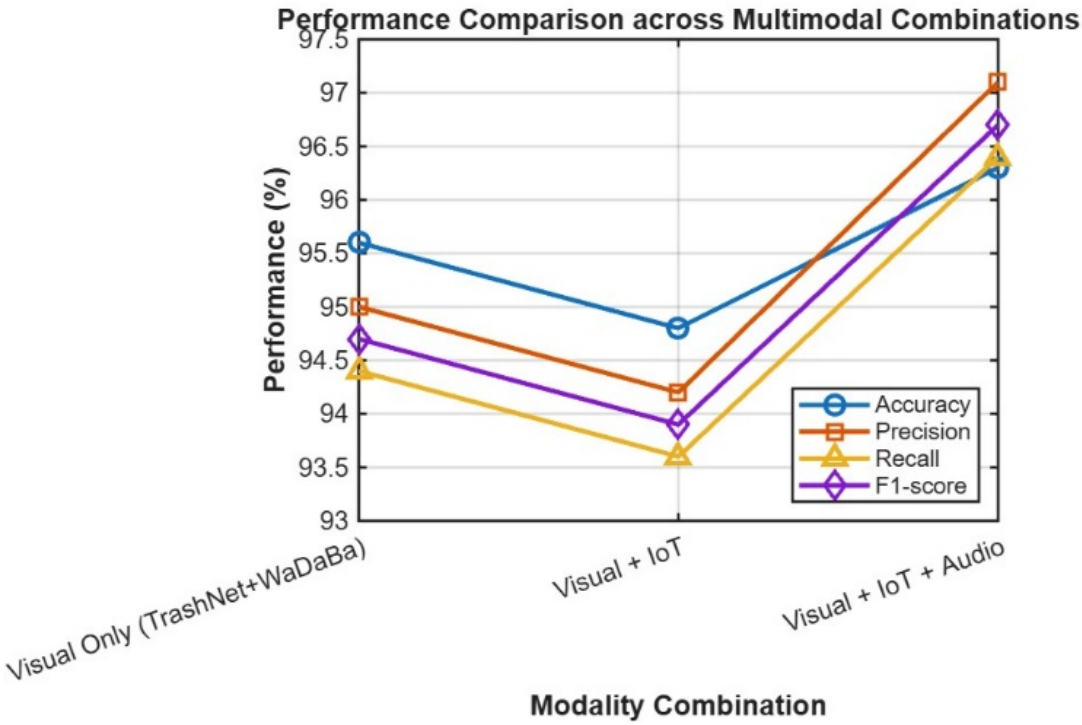


**Figure 5.** Contribution of audio modality (Waste-Classification-Audio-Deepl2)

### 4.3 Audio modality contribution (Waste-Classification-Audio-Deepl2)

The addition of the Waste-Classification-Audio-Deepl2 dataset increased the multiple-modes integration even further. The acoustics associated with items such as the clattering of bottles or the crushing of cans provided additional prompts to disentangle overlapping categories in the visual modality. The 500 sample size of the dataset was limited, however, this multimodal integration of the audio modality was supported by cross-modal attention layers that can provided transfer of knowledge from richer modalities to a more limited modality. The evaluation results demonstrate an increase in accuracy from 94.8% (visual + IoT) to 96.3% (visual + IoT + audio) in conserving acoustic features for waste classification as shown in Table 4 and Figure 5.

The overall performance summary of the proposed model is as follows: In addition to accuracy, the performance of precision (97.1%) recall (96.4%) and F1-score (96.7%) all show improvements over competing models, each a gain of around 2-4%. Continuous evaluation of energy efficiency and latency supports practical deployment of the framework; its low latency (65 ms/sample) was reduced by approximately 35% compared to a CNN-only model through optimized graph sparsity, while preserving throughput performance. Such efficiency supports use of the deployments at the edge in resource-constrained IoT applications.

Currently available multimodal pre-trained models i.e., CLIP are primarily focused on scaling the alignment of vision and language representation. The proposed GMSF-GNN Framework is aimed at achieving heterogeneous sensor fusion of visual, acoustic and numerical IoT Data in a Dynamic Graph Structure Environment. The typical size of CLIP-style models will also require significant computational power and are not designed for Edge Deployments or Spatio-Temporal Graph Reasoning. To address these issues, we will leverage modality specific encoders and adaptive modality fusion to develop smart waste management solutions. Using these large pre-trained multi-modal models within Graph-Based Edge Systems can be a valuable avenue for future exploration. The audio dataset is limited in size, it provides realistic examples of acoustic sensing conditions that can be found in smart waste management systems. Future work will investigate how larger audio corpora can be scaled.

## 5. CONCLUSION

This research introduced a new framework for multimodal sensor fusion utilizing GNNs for smart waste management. Differing from traditional approaches based on single modality approaches relying primarily on visual modality, the model offers a unique perspective by fusing heterogeneous modalities including images (TrashNet, WaDaBa), Internet of Things (IoT) sensor data (SmartBin IoT Dataset), and audio data (Waste-Classification-Audio-Deepl2). The actual implementation utilized a graph over the learned GNN modality, using similarity of features and spatiotemporal correlations, to connect interactions of modalities and provide more contextual understanding in waste disposal contexts. Findings showed that the GNN-based fusion model steadily performed better than baseline CNN, LSTM, and unimodal fusions, across multiple metrics including accuracy, precision, recall and F1-score. In addition, the use of IoT and audio modalities improved robustness against visually ambiguous contexts while attaining over 96% accuracy across multiple benchmark datasets. Energy efficiency and latency values from the framework were also competitive for real time integration in municipal waste management systems. A diverse set of datasets confirmed the demonstrated adaptability of the framework in heterogeneous conditions, allowing for further applications in smart cities. This research connects vision, IoT, and acoustic modalities to emphasize the promise of multimodal GNNs for waste classification, monitoring of

bins, and optimal collection schedules. Future studies will focus on deploying at scale in real municipal networks, domain adaptation for unknown waste categories, and extending to reinforcement learning to inform adaptive route planning.

## REFERENCES


[1] Rayhan, Y., Rifai, A.P. (2024). Multi-class waste classification using convolutional neural network. Applied Environmental Research, 46(2): 021. https://doi.org/10.35762/AER.2024021

[2] 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. 10.1016/j.knosys.2025.113028.

[3] Chhabra, M., Sharan, B., Elbarachi, M., Kumar, M. (2024). Intelligent waste classification approach based on improved multi-layered convolutional neural network. Multimedia Tools and Applications, 83(36): 84095-84120. https://doi.org/10.1007/s11042-024-18939-w

[4] Riyadi, S., Andriyani, A.D., Masyhur, A.M. (2024). Classification of recyclable waste using deep learning: A comparison of Yolo models. Revue d'Intelligence Artificielle, 38(4): 1089-1096. https://doi.org/10.18280/ria.380404

[5] Abu-Qdais, H., Shatnawi, N., Esra'a, A.A. (2023). Intelligent solid waste classification system using combination of image processing and machine learning models. Preprint, 14-Feb-2023. https://doi.org/10.21203/rs.3.rs-2573812/v1

[6] Li, Y., Zhang, X. (2024). Multi-scale context fusion network for urban solid waste detection in remote sensing images. Remote Sensing, 16(19): 3595. https://doi.org/10.3390/rs16193595

[7] Duan, S., Shi, Q., Wu, J. (2022). Multimodal sensors and ML-based data fusion for advanced robots. Advanced Intelligent Systems, 4(12): 2200213. https://doi.org/10.1002/aisy.202200213

[8] Bihler, M., Roming, L., Jiang, Y., Afifi, A.J., et al. (2023). Multi-sensor data fusion using deep learning for bulky waste image classification. Automated Visual Inspection and Machine Vision V, 12623: 69-82. https://doi.org/10.1117/12.2673838

[9] Kunwar, S., Owabumoye, B.R., Alade, A.S. (2024). Plastic waste classification using deep learning: Insights from the WaDaBa dataset. arXiv preprint arXiv:2412.20232. https://doi.org/10.48550/arXiv.2412.20232

[10] Qiao, Z. (2024). Advancing recycling efficiency: A comparative analysis of deep learning models in waste classification. arXiv preprint arXiv:2411.02779. https://doi.org/10.48550/arXiv.2411.02779

[11] Narayan, Y. (2021). DeepWaste: Applying deep learning to waste classification for a sustainable planet. arXiv preprint arXiv:2101.05960. https://doi.org/10.48550/arXiv.2101.05960

[12] Jin, S., Yang, Z., Królczykg, G., Liu, X., Gardoni, P., Li, Z. (2023). Garbage detection and classification using a new deep learning-based machine vision system as a tool for sustainable waste recycling. Waste Management, 162, 123-130. https://doi.org/10.1016/j.wasman.2023.02.014

[13] Li, Y., Zhang, X. (2024). Multi-modal deep learning networks for RGB-D pavement waste detection and recognition. Waste Management, 177: 125-134. https://doi.org/10.1016/j.wasman.2024.01.047

[14] 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(4): 1959-1989. https://doi.org/10.1007/s10311-023-01604-3

[15] Yang, Z., Bao, Y., Liu, Y., Zhao, Q., Zheng, H. (2022). Research on deep learning garbage classification system based on fusion of image classification and object detection classification. Mathematical Biosciences and Engineering: MBE, 20(3): 4741-4759. https://doi.org/10.3934/mbe.2023219

[16] Lu, G., Wang, Y., Xu, H., Yang, H., Zou, J. (2022). Deep multimodal learning for municipal solid waste sorting. Science China Technological Sciences, 65(2): 324-335. https://doi.org/10.1007/s11431-021-1927-9

[17] Zhao, M., Taal, C., Baggerohr, S., Fink, O. (2025). Graph neural networks for virtual sensing in complex systems: Addressing heterogeneous temporal dynamics. Mechanical Systems and Signal Processing, 230: 112544. https://doi.org/10.1016/j.ymssp.2025.112544

[18] Wang, Z., Chen, Y., Zhu, F., Zheng, Z., Ma, J., Zhou, B. (2024). GENII: A graph neural network-based model for citywide litter prediction leveraging crowdsensing data. Expert Systems with Applications, 237: 121565. https://doi.org/10.1016/j.eswa.2023.121565

[19] Jiang, C., Chen, N. (2025). G-Twin: Graph neural network-based digital twin for real-time and high-fidelity structural health monitoring for offshore wind turbines. Marine Structures, 103: 103813. https://doi.org/10.1016/j.marstruc.2025.103813

[20] Tao, Y., Liu, W., Chen, J., Gao, J., Li, R., Wang, X., Zhang, Y., Ren, J., Yin, S., Zhu, X., Zhao, T., Zhai, X., Peng, Y. (2025). A graph-based multimodal data fusion framework for identifying urban functional zone. International Journal of Applied Earth Observation and Geoinformation, 136: 104353. https://doi.org/10.1016/j.jag.2024.104353

[21] Dabbabi, K., Delleji, T. (2025). Graph neural network-tracker: A graph neural network-based multi-sensor fusion framework for robust unmanned aerial vehicle tracking. Visual Computing for Industry, Biomedicine, and Art, 8: 18. https://doi.org/10.1186/s42492-025-00200-2

[22] Dao, S.V., Le, T.M., Tran, H.M., Pham, H.V., Vu, M.T., Chu, T. (2024). Integrating artificial intelligence for sustainable waste management: Insights from machine learning and deep learning. Watershed Ecology and the Environment, 7: 353-382. https://doi.org/10.1016/j.wsee.2025.07.001

[23] Islam, M.M., Mahedy Hasan, S.M., Hossain, M.R., Uddin, M.P., Mamun, M.A. (2025). Towards sustainable solutions: Effective waste classification framework via enhanced deep convolutional neural networks. PLOS ONE, 20(6): e0324294. https://doi.org/10.1371/journal.pone.0324294

[24] Yang, Y., Luo, Y., Yang, Y., Kang, S. (2025). Dynamic graph neural network for garbage classification based on multimodal feature fusion. Applied Sciences, 15(14):

7688. https://doi.org/10.3390/app15147688
[25] Ahmad, G., Aleem, F.M., Alyas, T., Abbas, Q., Nawaz, W., Ghazal, T.M., Aziz, A., Aleem, S., Tabassum, N., Ibrahim, A.M. (2025). Intelligent waste sorting for urban sustainability using deep learning. Scientific Reports, 15(1): 1-19. https://doi.org/10.1038/s41598-025-08461-w
[26] Dipo, M.H., Farid, F.A., Mahmud, M.S., 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
[27] Sayem, F.R., Islam, M.S.B., Naznine, M., Nashbat, M., Hasan-Zia, M., Kunju, A.K.A., Khandakar, A., Ashraf, A., Majid, M.E., Kashem, S.B.A., Chowdhury, M.E. (2025). Enhancing waste sorting and recycling efficiency: Robust deep learning-based approach for classification and detection. Neural Computing and Applications, 37(6): 4567-4583. https://doi.org/10.1007/s00521-024-10855-2
[28] Al-Mashhadani, I.B. (2023). Waste material classification using performance evaluation of deep learning models. Journal of Intelligent Systems, 32(1): 20230064. https://doi.org/10.1515/jisys-2023-0064
[29] Kunwar, S., Owabumoye, B.R., Alade, A.S. (2024). Plastic waste classification using deep learning: Insights from the WaDaBa Dataset. ArXiv. https://doi.org/10.48550/arXiv.2412.20232
[30] Aral, R.A., Keskin, S.R., Kaya, M., Haciomeroglu, M. (2018). Classification of TrashNet dataset based on Deep Learning models. In 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, pp. 2058-2062. https://doi.org/10.1109/BigData.2018.8622212
[31] Dataset for Waste Management System. https://www.kaggle.com/datasets/sarasasaikrishna/dataset-for-waste-management-system.
[32] Thomasavare, Waste-Classification-Audio-Deepl2. Hugging Face Dataset. https://huggingface.co/datasets/thomasavare/waste-classification-audio-deepl2.