Multi-Modal Deep Learning Approach for Waste Management: Integrating Image Classification and Text Mining for Environmental Awareness

Manuscript received January 27, 2026; revised June 3, 2026

Santi Prayudani*
Teknologi Rekayasa Perangkat Lunak
Politeknik Negeri Medan
Medan, Indonesia
santiprayudani@polmed.ac.id

Ainul Hizriadi
Teknologi Informasi
Universitas Sumatera Utara
Medan, Indonesia
ainul.hizriadi@usu.ac.id

Yuyun Yusnida Lase
Teknologi Rekayasa Perangkat Lunak
Politeknik Negeri Medan
Medan, Indonesia
yuyunlase@polmed.ac.id

Abstract— **Environmental degradation caused by inefficient waste management remains a major global challenge, largely due to the limitations of conventional systems that rely on manual waste sorting and limited utilization of heterogeneous data sources. This study proposes a novel multi-modal deep learning framework that integrates visual and textual information to enhance waste classification performance while simultaneously providing insights into environmental awareness. The proposed framework combines convolutional neural networks (CNNs) for waste image classification and a recurrent neural network with long short-term memory (LSTM) architecture for text analysis. Visual and textual feature representations are integrated through a feature-level fusion strategy using vector concatenation before final classification. The image dataset consists of six waste categories, cardboard, glass, metal, paper, plastic, and trash, while the textual dataset contains waste management descriptions, community feedback, and environmental discourse collected from public and field sources. Environmental awareness was assessed through text mining by identifying dominant themes related to recycling practices, waste sorting behavior, environmental responsibility, and public concern regarding pollution and sustainability issues. Experimental results demonstrate that the proposed multimodal framework achieves an accuracy of 88.9% and an F1-score of 0.89, outperforming image-only and text-only models with accuracies of 78.4% and 81.2%, respectively. This corresponds to absolute performance improvements of 10.5% over the image-based model and 7.7% over the text-based model, while reducing the classification error rate by 40.96%. Furthermore, the multimodal model exhibits superior robustness under degraded data conditions, with only a 4.7% reduction in accuracy compared to larger performance declines observed in unimodal approaches. The main contribution of this study lies in the integration of waste image recognition and environmental-awareness extraction within a unified multimodal learning framework, enabling not only accurate waste categorization but also the generation of behavioral and sustainability-related insights that support more intelligent and sustainable waste management systems.**

Keywords—Convolutional Neural Networks, Deep Learning, Environmental Awareness, Multi-Modal Learning, Text Mining, Waste Classification.

I. INTRODUCTION

Waste management remains one of the most critical challenges in achieving sustainable development, as the volume of municipal solid waste continues to increase due to population growth, rapid urbanization, and increasingly complex consumption patterns [1]. Inefficient waste sorting and disposal practices contribute not only to environmental pollution but also to the loss of valuable recyclable resources that could otherwise be recovered through circular economy initiatives [2], [3]. Despite ongoing efforts to improve waste management systems, many operational processes still rely on manual sorting and limited data-driven decision-making, resulting in suboptimal classification accuracy, operational inefficiencies, and reduced sustainability outcomes.

Recent advances in artificial intelligence, particularly deep learning, have significantly improved automated waste classification systems. Convolutional neural networks (CNNs) have demonstrated strong performance in recognizing and categorizing waste materials from images, making them a widely adopted solution for intelligent waste sorting applications [4]. Studies such as those by Ren et al. [5] and Mao et al. reported promising classification performance using image-based deep learning architectures. However, these approaches focus primarily on visual recognition tasks and largely overlook the social, behavioral, and contextual factors that influence waste generation, disposal behavior, and recycling participation. Consequently, image-based systems provide limited insight into the human dimensions of waste management that are essential for long-term environmental sustainability [6] [7].

At the same time, textual data generated through waste management reports, public discussions, social media content, and community feedback contain valuable information regarding public perceptions, attitudes, and behavioral intentions toward environmental issues [8] Natural language processing (NLP) and text mining techniques have increasingly been applied to extract meaningful patterns from such textual information [9]. Existing studies have demonstrated the effectiveness of NLP for sentiment analysis, topic modeling, and environmental discourse analysis. Nevertheless, these studies generally examine environmental communication independently and rarely connect textual insights with operational waste classification systems. As a result, the potential synergy between visual waste recognition and environmental perception analysis remains largely unexplored.

Multi-modal deep learning has emerged as a promising paradigm for integrating heterogeneous data sources to improve predictive performance and contextual understanding [10]. Previous multimodal studies have successfully combined

 EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002036

image and text information in domains such as crisis management, healthcare, and social media analytics [11], [12]. However, the application of multimodal learning in waste management remains limited. Existing waste-related studies primarily focus on improving classification accuracy through visual data, while studies employing text mining mainly investigate environmental discourse without linking the findings to waste categorization processes. To date, few studies have systematically integrated waste image classification and environmental text analysis within a unified framework designed to support both operational waste management and environmental awareness assessment.

In this study, environmental awareness is conceptually defined as the level of public understanding, concern, responsibility, and behavioral intention toward environmental protection and sustainable waste management practices. Methodologically, environmental awareness is derived from textual data through NLP-based thematic analysis, where key indicators are identified from recurring themes related to recycling behavior, waste segregation practices, pollution concerns, environmental responsibility, and sustainability-oriented actions. Through this approach, environmental awareness is treated not merely as an abstract concept but as an observable construct reflected in public environmental discourse and waste-related communication patterns.

Unlike previous studies that focus exclusively on waste image classification or environmental text mining as separate analytical tasks, this research integrates CNN-based visual recognition and LSTM-based textual analysis within a single multimodal deep learning framework using feature-level fusion. The proposed approach enables the simultaneous identification of waste categories and extraction of environmental-awareness indicators from textual information. The contributions of this study are threefold. First, it introduces a unified multimodal framework that combines waste image classification and environmental discourse analysis through the integration of CNN and LSTM architectures. Second, it operationalizes environmental awareness as a measurable construct derived from text mining and incorporates this information into the waste management decision-support process. Third, it empirically demonstrates that multimodal fusion improves classification accuracy, robustness, and reliability compared with unimodal image-based and text-based approaches, thereby providing a stronger foundation for the development of intelligent and sustainable waste management systems.

II. Material and Methods

A. Research Design

This study employed an experimental research design to develop and evaluate a multimodal deep learning framework for intelligent waste management. The proposed framework integrates visual and textual data to improve waste classification performance while simultaneously generating insights related to environmental awareness. The experimental approach was selected to systematically assess the effectiveness of multimodal learning compared with unimodal image-based and text-based models under identical evaluation conditions [13].

The research workflow consisted of five main stages: data collection, including the acquisition of waste images and waste-related textual data; data preprocessing, involving image normalization, augmentation, and text cleaning procedures; development of the multimodal deep learning architecture, comprising a CNN-based image encoder and an LSTM-based text encoder; model training and validation using predefined training and validation datasets; and performance evaluation and analysis using quantitative metrics, including accuracy, precision, recall, and F1-score, as well as qualitative analysis of environmental-awareness themes extracted from textual data [14].

B. Dataset Description and Data Collection

The dataset used in this study consisted of two complementary modalities: waste image data and textual data related to waste management practices and environmental awareness. The image dataset contained 6,000 labeled waste images distributed equally across six categories: cardboard (1,000 images), glass (1,000 images), metal (1,000 images), paper (1,000 images), plastic (1,000 images), and trash (1,000 images). The images were collected from publicly available waste classification repositories and supplemented with field documentation to enhance environmental diversity and representation.

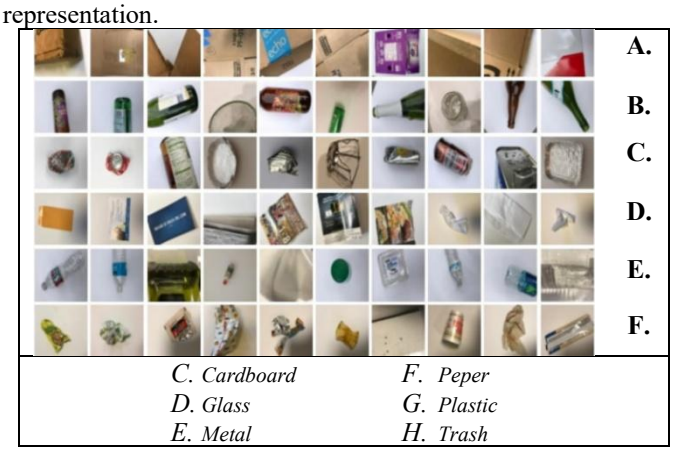


Fig. 1. Examples of Waste Categories in the Image Dataset

To ensure consistency and improve model performance, all images were resized to 224 × 224 pixels, normalized, and subjected to preprocessing procedures including noise reduction, color correction, and pixel-value normalization. To increase data diversity and reduce the risk of overfitting, data augmentation techniques such as random rotation (±15°), horizontal flipping, zooming, and scaling were applied. The complete dataset was divided into training, validation, and testing subsets using a 70:15:15 ratio, resulting in 4,200 training samples, 900 validation samples, and 900 testing samples.

The textual dataset consisted of 6,000 text records collected from waste descriptions, environmental reports, community feedback, and public discussions concerning waste management. Each image was paired with a corresponding textual description or environmental-context statement, resulting in 6,000 paired image–text samples for multimodal learning. This paired-data structure enabled the model to jointly learn visual waste characteristics and contextual information related to environmental behavior and awareness.

Text preprocessing included tokenization, case folding, stop-word removal, and stemming to produce clean and structured textual representations. Subsequently, textual features were extracted using the Term Frequency–Inverse Document Frequency (TF-IDF) method. TF-IDF was selected because it effectively identifies discriminative terms related to waste management and environmental issues while maintaining computational efficiency and interpretability. The

 EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002036

extracted features facilitated the identification of environmental-awareness indicators, including recycling behavior, waste segregation practices, pollution concerns, environmental responsibility, and sustainability-oriented actions.

For image modeling, a Convolutional Neural Network (CNN) architecture was employed to extract visual features such as texture, color, and shape patterns of waste objects. The CNN consisted of three convolutional blocks with 32, 64, and 128 filters, respectively, using 3 × 3 kernels, Rectified Linear Unit (ReLU) activation functions, and 2 × 2 max-pooling layers. A dropout layer with a rate of 0.5 was applied before the fully connected layer to reduce overfitting.

For text modeling, a Long Short-Term Memory (LSTM) network with 128 hidden units and a maximum sequence length of 200 tokens was used to capture contextual dependencies within textual data. The combination of TF-IDF and LSTM was adopted to leverage both term-importance weighting and sequential context learning, thereby improving semantic representation and environmental-discourse analysis.

The visual and textual feature vectors generated by the CNN and LSTM submodels were integrated using a feature-level fusion strategy through vector concatenation. This approach was selected because it preserves modality-specific information while enabling the model to learn interactions between visual and textual representations before the final classification stage. Compared with decision-level fusion, feature-level fusion allows richer multimodal feature learning and more effective joint representation of waste-related information [15].

Model training was performed using the Adam optimizer with a learning rate of 0.001, batch size of 32, and a maximum of 50 epochs. Early stopping with a patience value of 10 epochs was implemented to prevent overfitting. The categorical cross-entropy loss function was employed for multi-class classification. All experiments were conducted in a GPU-based computing environment to accelerate training and evaluation processes.

Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. In addition, the proposed multimodal model was compared with image-only and text-only baseline models to assess the effectiveness of multimodal integration in improving waste classification performance, robustness, and environmental-awareness analysis.

C. Detail Arsitektur Model Multi-Modal

The proposed multimodal deep learning framework is designed to jointly learn visual and textual representations for intelligent waste classification and environmental-awareness analysis. As illustrated in Fig. 2, the architecture consists of three main components: (1) a CNN-based image encoder for visual feature extraction, (2) an LSTM-based text encoder for textual feature learning, and (3) a feature-level fusion module for integrating multimodal information before final classification.

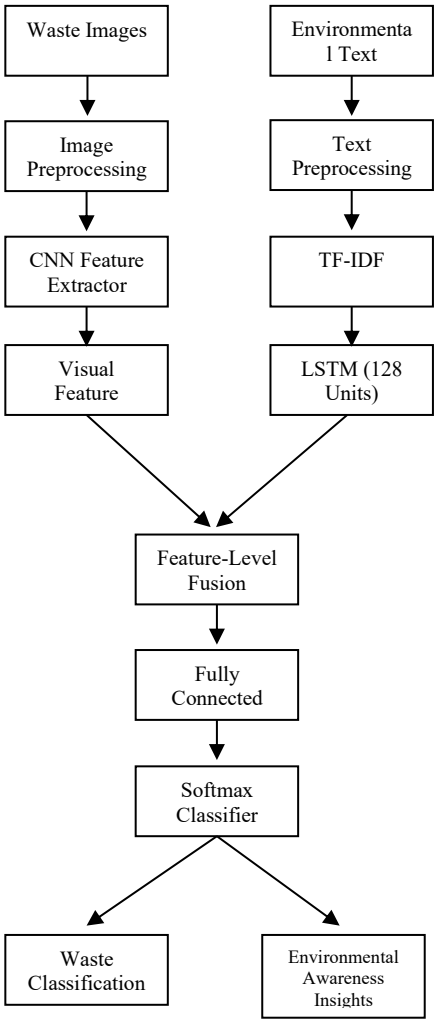


Fig. 2. Proposed multimodal deep learning framework integrating CNN-based image feature extraction and LSTM-based textual feature learning through feature-level fusion for waste classification and environmental-awareness analysis.

The image branch processes waste images through a Convolutional Neural Network (CNN) architecture composed of three convolutional blocks with 32, 64, and 128 filters, respectively. Each convolutional layer employs a 3 × 3 kernel followed by Rectified Linear Unit (ReLU) activation and 2 × 2 max-pooling operations. This hierarchical structure enables the extraction of increasingly complex visual features, ranging from low-level patterns such as edges and textures to higher-level object characteristics associated with different waste categories. A dropout layer with a rate of 0.5 is applied before the fully connected layer to improve model generalization and reduce overfitting. The output of the CNN encoder is represented as a fixed-dimensional visual feature vector.

The text branch processes environmental text data through a sequence of preprocessing stages, including tokenization, case folding, stop-word removal, and stemming. The processed text is then transformed into numerical representations using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. TF-IDF was selected because it effectively emphasizes informative and environmentally relevant terms while maintaining computational efficiency and interpretability. The resulting feature vectors are subsequently processed by a Long Short-Term Memory (LSTM) network with 128 hidden units and a maximum sequence length of 200 tokens. The LSTM encoder captures

contextual and sequential relationships among environmental terms and produces a textual feature vector representing environmental discourse and awareness-related information.

To integrate information from both modalities, the visual and textual feature vectors are combined through a feature-level fusion strategy using vector concatenation. This approach was chosen because it preserves modality-specific information while enabling the model to learn interactions between visual and textual representations before classification. Compared with decision-level fusion, feature-level fusion provides richer joint representations and facilitates more effective multimodal learning.

The fused multimodal representation is subsequently processed by a fully connected layer followed by a Softmax classifier to generate the final waste-category prediction. Through this mechanism, the model can exploit complementary information from both modalities, allowing textual context to support classification when visual information is ambiguous and vice versa. This multimodal interaction improves prediction robustness, classification consistency, and overall system reliability.

The combination of CNN-based visual feature extraction, LSTM-based textual feature learning, and feature-level fusion provides a unified framework capable of simultaneously leveraging object-level waste characteristics and contextual environmental information. Consequently, the proposed architecture offers a more comprehensive representation of waste-related data than conventional unimodal approaches and serves as the foundation for the intelligent waste management system proposed in this study.

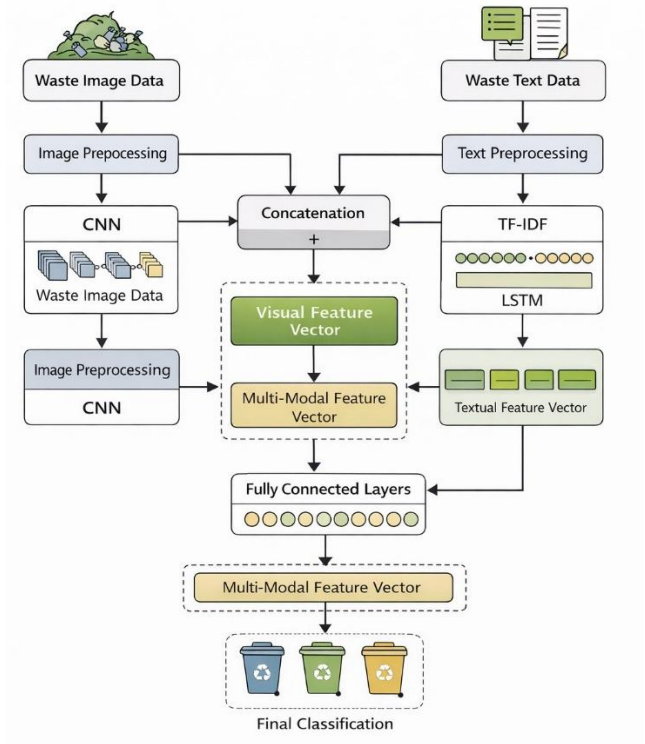


Fig. 3. Multi-modal deep learning architecture integrating CNN-based image features and LSTM-based textual features for waste classification.

III. RESULTS

This section presents empirical results from the application of a multi-modal deep learning approach that integrates convolutional neural network-based (CNN)-based image classification and natural language processing-based text analysis (NLP) to support waste management systems and increase environmental awareness. The results of the study focused on the performance of the classification model, the contribution of each data modality, and the effectiveness of multimodal integration in improving the accuracy and reliability of the system.

A. Performance of Image-Based Waste Classification Model

The evaluation results indicate that the CNN-based image classification model achieved strong performance across the six waste categories, demonstrating a balanced relationship between precision and recall values. This balance suggests that the model was able to effectively distinguish among different waste types while maintaining relatively low rates of false-positive and false-negative predictions. Such performance is particularly important in waste management applications, where classification errors may reduce sorting efficiency and recycling effectiveness.

A more detailed examination of the classification outcomes reveals that the CNN successfully learned discriminative visual patterns associated with each waste category. The convolutional layers captured hierarchical image features ranging from low-level characteristics, such as edges and textures, to higher-level representations, including object shape and structural composition. These findings indicate that the extracted visual features were sufficiently informative to differentiate among waste materials with distinct physical properties.

Despite variations in image appearance caused by differences in illumination, object orientation, background complexity, and waste condition, the model maintained consistent classification performance across most categories. This result suggests that the combination of image preprocessing, data augmentation, and hierarchical feature extraction contributed to improved robustness against common sources of visual variability. However, several misclassification cases were still observed between visually similar categories, particularly among materials that share comparable textures or color distributions. This limitation highlights the inherent challenge of relying solely on visual information for waste recognition and further supports the need for integrating complementary textual information within the proposed multimodal framework.

TABLE I. CONFUSION MATRIX OF THE CNN-BASED WASTE CLASSIFICATION MODEL

Actual \ Predicted	Cardboard	Glass	Metal	Paper	Plastic	Trash
Cardboard	**93.5**	1.2	0.8	3.1	0.7	0.7
Glass	1.0	**92.8**	3.4	0.6	1.5	0.7
Metal	0.6	3.2	**93.1**	0.5	1.8	0.8
Paper	3.9	0.7	0.6	**92.9**	0.8	1.1
Plastic	0.9	3.8	1.6	0.7	**92.1**	0.9
Trash	1.4	1.1	1.3	1.6	1.8	**92.8**

The confusion matrix presented in Table I indicates that most predictions are concentrated along the main diagonal, suggesting that the model was able to correctly classify the majority of waste samples across all categories. The strongest classification performance was observed in the cardboard, metal, and trash classes, indicating that these categories possess more distinctive visual characteristics that can be effectively captured by the CNN feature extraction process.

Misclassification cases were primarily observed among categories with similar visual appearances. For example, confusion occurred between the plastic and glass classes, particularly for transparent plastic objects that share

 EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002036

comparable reflectance and transparency properties with glass materials. Likewise, some ambiguity was identified between paper and trash samples, especially when paper waste exhibited physical degradation, contamination, or irregular textures that reduced visual distinctiveness. These findings suggest that classification errors were largely associated with intrinsic similarities in visual features rather than limitations in the learning process itself.

The observed error patterns highlight a common challenge in image-based waste classification, where visually similar materials may generate overlapping feature representations. Nevertheless, the relatively limited number of off-diagonal predictions indicates that the CNN architecture was generally effective in learning discriminative visual features across waste categories. This outcome demonstrates the capability of convolutional feature extraction to capture meaningful patterns related to texture, shape, and material characteristics despite variations in object appearance and environmental conditions.

From a practical perspective, the confusion matrix analysis suggests that image-based classification can provide a robust foundation for automated waste recognition. However, the remaining ambiguities among visually similar categories also reveal the limitations of relying solely on visual information. This observation further supports the integration of complementary textual information in the proposed multimodal framework, where contextual environmental descriptions may help reduce uncertainty and improve classification performance in challenging cases [16].

B. Results of Text Mining and Environmental Awareness Analysis

The NLP analysis demonstrated that the proposed text-processing pipeline was effective in extracting meaningful semantic patterns from waste-management-related texts. The combination of preprocessing techniques and TF-IDF feature extraction reduced linguistic noise and enhanced the representation of environmentally relevant terms, enabling the model to capture key contextual information within the corpus.

The LSTM-based text model successfully differentiated among informative, persuasive, and neutral content categories, indicating its capability to learn contextual relationships within environmental discourse. Thematic analysis further revealed several recurring topics, including recycling practices, waste segregation behavior, environmental and health impacts of plastic waste, and the role of individuals and communities in sustainable waste management.

In addition, the model identified narratives associated with environmental responsibility, pollution concerns, and sustainability-oriented actions. These findings suggest that textual data provide valuable contextual information that complements visual waste classification by capturing behavioral and perceptual aspects that cannot be directly observed from images alone. Overall, the results indicate that the text modality contributes meaningful environmental context to the proposed multimodal framework, supporting a more comprehensive understanding of waste-management issues beyond visual recognition alone.

C. Comparative Analysis of Single-Modal and Multi-Modal Models

A comparison between the unimodal models (image-only and text-only) and the multimodal model indicates that integrating visual and textual information resulted in higher classification performance. The multimodal model consistently achieved better accuracy than both unimodal approaches, suggesting that the complementary information provided by the two modalities contributed to more effective waste classification.

TABLE II. CLASSIFICATION ACCURACY OF UNIMODAL AND MULTIMODAL MODELS

Model	Modalities	Accuracy (%)
Model A	Image only	78.4
Model B	Text only	81.2
Model C	Multimodal	88.9

The increase in the performance of the multimodal model compared to the unimodal approach is calculated based on the absolute accuracy difference as follows:

$$\Delta A_{MM - Image} = A_{MM} - A_{Image} = 88.9\% - 78.4\% = 10.5\% \quad (1)$$

$$\Delta A_{MM - Text} = A_{MM} - A_{Text} = 88.9\% - 81.2\% = 7.7\% \quad (2)$$

Multimodal integration improves accuracy by up to 10.5% absolutely, demonstrating a significant contribution from cross-modality contexts.

TABLE III. ERROR RATE AND ERROR REDUCTION

Model	Accuracy	Error Rate
Unimodal Image	0.784	0.216
Unimodal Text	0.812	0.188
Multimodal	0.889	0.111

$$RER = \frac{0.188 - 0.111}{0.188} \times 100\% = 40.96\% \quad (3)$$

The multimodal approach reduces misclassification by almost 41% compared to the best unimodal models.

TABLE IV. AVERAGE EVALUATION METRICS

Model	Precision	Recall	F1-Score
Unimodal Image	0.77	0.76	0.76
Unimodal Text	0.81	0.81	0.80
Multimodal	0.89	0.89	0.89

The multimodal model shows a better balance of precision and recall, signaling improved prediction consistency between classes.

The quantitative evaluation results showed that the multimodal model achieved higher performance than the unimodal models across all evaluation metrics. As presented in Tables II–IV, the integration of visual and textual features resulted in the highest accuracy (88.9%) and F1-score (0.89), while also reducing the classification error rate compared with the image-only and text-only approaches. These results indicate that combining visual and textual information provides a more informative representation for waste classification than either modality alone.

The improved performance may be attributed to the complementary nature of the two data modalities. Visual features capture the physical characteristics of waste objects, whereas textual features provide contextual information related to waste management and environmental issues. When one modality provides limited information, the other may contribute additional cues that support the classification process [17]. The findings suggest that multimodal feature integration can improve classification performance and prediction consistency by combining complementary information from visual and textual sources within a unified learning framework.

D. Impact of Multi-Modal Learning on System Reliability

The reliability evaluation examined the performance of the proposed multimodal model under different input conditions to assess its stability across varying data quality scenarios. In this study, reliability refers to the ability of the model to maintain relatively consistent classification performance when processing inputs with different levels of visual and textual quality. The results indicate that the multimodal model maintained more stable performance than the unimodal models across the evaluated test scenarios.

Reliability testing was conducted using both normal and degraded data conditions. Image degradation scenarios included blur and reduced exposure levels, while text degradation scenarios involved variations in language style and incomplete textual information [18]. These controlled conditions were designed to simulate common challenges encountered in real-world waste management environments and to evaluate the model's sensitivity to variations in input quality.

TABLE V. MODEL ACCURACY IN VARIOUS TEST SCENARIOS

Model	Normal Condition (%)	Disturbed Conditions (%)	Decreased Accuracy (%)
Unimodal Image	78.4	69.1	−9.3
Unimodal Text	81.2	73.5	−7.7
Multimodal	88.9	84.2	−4.7

The decrease in accuracy (ΔA) is calculated as:

$$\Delta A = A_{\text{normal}} - A_{\text{disturbed}} \quad (4)$$

The results showed that the multimodal model experienced the smallest decline in performance, which was 4.7%, compared to the unimodal image (9.3%) and unimodal text (7.7%) models. Relatively speaking, the multimodal approach is able to reduce performance degradation by 49.46% compared to the unimodal image model, as expressed in Equation (5):

$$\frac{9.3-4.7}{9.3} \times 100\% = 49.46\% \quad (5)$$

The lower reduction in performance observed in the multimodal model suggests that the integration of visual and textual information may help mitigate the effects of noise or information loss in a single modality. By combining complementary features from both data sources, the model was able to maintain more consistent classification performance under the evaluated degradation scenarios. These findings indicate that multimodal feature integration can contribute to improved stability when processing inputs of varying quality [19].

In addition to accuracy, reliability is also analyzed through performance consistency between test scenarios measured using standard deviation accuracy.

TABLE VI. ACCURACY VARIANCE AS AN INDICATOR OF RELIABILITY

Model	Average Accuracy (%)	Standard Deviation (%)
Unimodal Image	73.8	4.9
Unimodal Text	77.4	3.8
Multimodal	86.6	2.3

Lower standard deviation values in multimodal models indicate that the performance distribution is more stable and does not fluctuate to changes in input conditions. A decrease in variance is as follows:

$$\frac{4.9-2.3}{4.9} \times 100\% = 53.06\% \quad (6)$$

compared to the unimodal model, the image indicates a substantial increase in reliability.

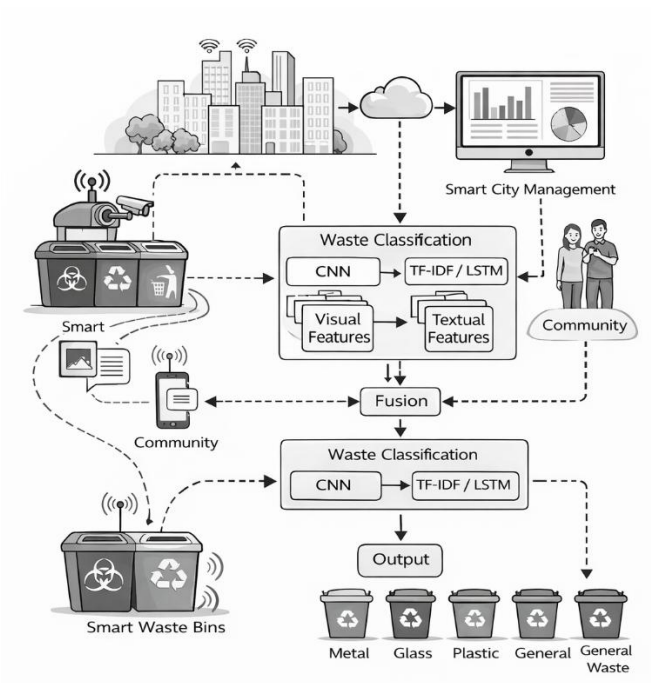


Fig. 3. Conceptual illustration of reliability evaluation in the proposed multimodal learning system under input degradation and uncertainty in real-world waste management scenarios.

In real-world simulation scenarios, some of the data were tested under incomplete conditions, such as very short descriptive texts or images with limited visual information. Under these conditions, the unimodal models experienced a greater decline in classification performance due to their reliance on a single source of information. In contrast, the multimodal model maintained more consistent performance because it was able to utilize complementary information from the available modality when the other modality contained limited or degraded information [20].

Quantitatively, the classification success rate on incomplete data increased from 74.6% for the best-performing unimodal model to 83.1% for the multimodal model, representing an absolute improvement of 8.5%. This result indicates that the integration of visual and textual features can support the classification process under heterogeneous data conditions that may not be adequately handled by a single modality.

The reliability evaluation further suggests that multimodal feature integration can reduce the sensitivity of the model to variations in input quality. Since visual and textual data provide different types of information, the combined representation enables the model to access a broader contextual basis for classification decisions. Consequently, performance degradation under incomplete data conditions was less pronounced in the multimodal model than in the unimodal approaches [21].

These findings have potential implications for applications involving heterogeneous waste-related information, where image quality and textual descriptions may vary across users and operational settings. However, the current evaluation was

 EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002036

conducted under controlled degradation scenarios, and therefore the results should be interpreted as evidence of the model's performance under the tested conditions rather than as a direct representation of all real-world environments.

Overall, the quantitative and qualitative results indicate that multimodal feature integration contributes to more stable classification performance across different input conditions. By combining visual and textual information within a unified framework, the proposed model was able to maintain higher classification performance than the unimodal models in the evaluated scenarios. Nevertheless, further studies involving larger datasets, more diverse degradation settings, and real operational environments are needed to validate the generalizability of these findings.

E. Limitations and Future Work

Although the proposed multimodal deep learning framework demonstrated improved classification performance compared with the unimodal approaches, several limitations should be considered when interpreting the findings and designing future research.

First, the dataset used in this study remains limited in terms of scale and contextual diversity. Although image and text data were curated and augmented to improve representation, the dataset may not fully capture the wide range of environmental conditions, waste characteristics, and waste-management practices encountered across different geographic regions and socio-cultural contexts [22]. Consequently, further validation using larger and more diverse datasets is required to assess the generalizability of the proposed framework.

Second, the textual dataset was primarily collected in a single language, which may restrict the model's ability to capture linguistic variation and environmental discourse in multilingual settings. Future studies may explore multilingual datasets and advanced language representations, including multilingual embeddings and transformer-based language models, to improve the adaptability of the text-analysis component across different language environments [23].

Third, the current system was evaluated in an offline experimental setting and was not designed for real-time processing. While the results provide evidence of the effectiveness of multimodal learning under the evaluated conditions, additional research is needed to investigate the performance of the framework in real-time applications, such as smart waste-reporting systems, environmental monitoring platforms, and other operational decision-support environments .

Fourth, model training and evaluation were conducted using a GPU-based computing environment. The feasibility of deploying the proposed framework on resource-constrained platforms, such as edge devices and Internet of Things (IoT) systems, was not investigated in this study. Future work may examine model compression, pruning, and quantization techniques to improve computational efficiency and support deployment in low-resource environments.

In addition, this study employed a relatively simple feature-level fusion strategy based on vector concatenation. Although the results indicate that this approach is effective for integrating visual and textual information, alternative fusion mechanisms, such as attention-based fusion, transformer-based multimodal architectures, or cross-modal representation learning, may provide richer interactions between modalities and warrant further investigation.

The findings of this study provide evidence that multimodal learning can support waste classification by integrating complementary visual and textual information. However, further research involving larger datasets, multilingual environments, real-time implementation, and advanced multimodal architectures is necessary to evaluate the scalability and broader applicability of the proposed framework.

IV. CONCLUSION

This study evaluates a multi-modal deep learning approach that integrates image classification based on convolutional neural networks (CNN) and natural language processing (NLP)-based text analysis to support artificial intelligence-based waste management. The results of the experiment showed that the integration of visual and textual data consistently improved system performance compared to the unimodal approach.

The multimodal model achieved an increase in accuracy of up to 10.5% compared to an image-based model and 7.7% compared to a text-based model, accompanied by a reduction in the relative error rate of 40.96% and an improvement in the balance of precision, recall, and F1-score. In addition, the multimodal approach showed higher reliability, characterized by a decrease in performance degradation of up to 49.46% under degraded data conditions and lower performance variance compared to the unimodal model.

These findings confirm that cross-modality integration results in richer, more stable, and adaptive feature representations to data variations and uncertainties, making them more suitable for real-world operational conditions. Nonetheless, further research is still needed to test the model's generalization at a broader scale and application context. Overall, the proposed multimodal approach has a strong empirical basis for the development of reliable and sustainable smart waste management systems.

ACKNOWLEDGMENT

The author would like to thank all parties who have contributed to the implementation of this research. This research is supported by the availability of computing resources and research facilities that allow the development and evaluation of the proposed multimodal deep learning framework. The authors also appreciate the use of publicly available datasets and open source libraries that are very helpful in the process of data collection, model development, and experimental analysis. In addition, the author expressed his appreciation to his peers and reviewers for the constructive input provided so as to improve the quality and clarity of this research.

REFERENCES

[1] T. Kumari and A. S. Raghubanshi, "Waste management practices in the developing nations: challenges and opportunities," *Waste Manag. Resour. Recycl. Dev. world*, pp. 773–797, 2023.

[2] A. Lakhouit, "Revolutionizing urban solid waste management with AI and IoT: a review of smart solutions for waste collection, sorting, and recycling," *Results Eng.*, p. 104018, 2025.

[3] A. K. Rathoure, "Revolutionizing waste management with advancements in sorting and processing technologies," in *Municipal Solid Waste Management and Recycling Technologies*, IGI Global, 2024, pp. 387–406.

[4] P. K. Gupta, V. Shree, L. Hiremath, and S. Rajendran, "The use of modern technology in smart waste management and recycling: artificial intelligence and machine learning," in *Recent advances in computational intelligence*, Springer, 2019, pp. 173–188.

[5] C. Ren, H. Jung, S. Lee, and D. Jeong, "Coastal waste detection based on deep convolutional neural networks," *Sensors*, vol. 21, no. 21, p.

7269, 2021.
[6] T. H. Yan, S. N. M. Azam, Z. M. Sani, and A. Azizan, "Accuracy study of image classification for reverse vending machine waste segregation using convolutional neural network.," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, 2024.
[7] W.-L. Mao, W.-C. Chen, C.-T. Wang, and Y.-H. Lin, "Recycling waste classification using optimized convolutional neural network," *Resour. Conserv. Recycl.*, vol. 164, p. 105132, 2021.
[8] S. Alim, *Komunikasi Lingkungan: Konsep Kunci dan Studi Kasus Terkini di Asia dan Indonesia*. Universitas Brawijaya Press, 2024.
[9] R. Egger and E. Gokce, "Natural language processing (NLP): An introduction: making sense of textual data," in *Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications*, Springer, 2022, pp. 307–334.
[10] T. Sadiq and C. W. Omlin, "Sensing in Smart Cities: A Multimodal Machine Learning Perspective," *Smart Cities*, vol. 9, no. 1, p. 3, 2025.
[11] T. Diwan and J. V Tembhurne, "Sentiment analysis: a convolutional neural networks perspective," *Multimed. Tools Appl.*, vol. 81, no. 30, pp. 44405–44429, 2022.
[12] A. Rosewelt and A. Renjit, "Semantic analysis-based relevant data retrieval model using feature selection, summarization and CNN," *Soft Comput.*, vol. 24, no. 22, pp. 16983–17000, 2020.
[13] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, "Multi-modality helps in crisis management: An attention-based deep learning approach of leveraging text for image classification," *Expert Syst. Appl.*, vol. 195, p. 116626, 2022.
[14] I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "Multi-model fusion framework using deep learning for visual-textual sentiment classification," *Comput. Mater. Contin.*, vol. 76, no. 2, pp. 2145–2177, 2023.
[15] J. Yang and Q. Long, "A modification of adaptive moment estimation (adam) for machine learning," *J. Ind. Manag. Optim.*, vol. 20, no. 7, pp. 2516–2540, 2024.
[16] D. V. Yevle and P. S. Mann, "Artificial Intelligence-Based Waste Management: A Review of Classification, Techniques, Issues, and Challenges," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 15, no. 2, p. e70025, 2025.
[17] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 10944–10956, 2021.
[18] S. Li, Z. Chen, Q. Liu, W. Shi, and K. Li, "Modeling and analysis of performance degradation data for reliability assessment: A review," *IEEE Access*, vol. 8, pp. 74648–74678, 2020.
[19] U. Ullah *et al.*, "A review of multi-modal learning from the text-guided visual processing viewpoint," *Sensors*, vol. 22, no. 18, p. 6816, 2022.
[20] Z. Lin, S. Yu, Z. Kuang, D. Pathak, and D. Ramanan, "Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19325–19337.
[21] Z. Han, F. Yang, J. Huang, C. Zhang, and J. Yao, "Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20707–20717.
[22] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *J. big Data*, vol. 8, no. 1, p. 101, 2021.
[23] S. Zhu *et al.*, "Multilingual Large Language Models: A Systematic Survey," *arXiv Prepr. arXiv2411.11072*, 2024.

 EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002036