International Research Journal of Multidisciplinary Sciences ISSN:AWAITED
VOL-1 Issue -4 April 2025 PP:23-30

Optimizing Waste Management Through Intelligent Machine Learning Systems: Emphasis on Data Preprocessing and Label Engineering

[1]YazdaniHasan,[2]BhawnaKaushik,[3]PriyaGupta

1. yazhassid@gmail.com,NoidaInternationalUniversity

2. priya.gupta@niu.edu.in,NoidaInternationalUniversity

3. bhawna.kaushik@niu.edu.in,NoidaInternationalUniversity

Abstract

Effective waste management has become an urgent priority due to escalating environmental concerns and increasing urbanization. Traditional systems, which often rely on manual classification and operational inefficiencies, are increasingly being supplemented or replaced by intelligent machine learning (ML) models. However, the performance of these models largely depends on the quality of data preprocessing and accurate label generation. This study introduces an innovative approach to developing a machine learning-based waste management system, emphasizing the importance of robust data preprocessing techniques and strategic labeling for supervised learning tasks. We present a comparative evaluation of preprocessing methods, propose a novel labeling schema, and assess their impact across multiple ML algorithms. The results demonstrate significant performance improvements, highlighting the potential for more sustainable, automated, and intelligent waste classification and decision-making systems.

1. Introduction

1.1 Background

Waste management is a critical component of urban sustainability. Cities around the world face increasing challenges related to waste generation, segregation, collection, and recycling. Traditional waste classification methods—manual, labor-intensive, and prone to errors—are insufficient in handling the complexity and volume of modern urban waste.

Machine learning (ML), with its ability to learn from data and make intelligent decisions, has emerged as a promising solution. Applications of ML in waste management include image-based waste classification, anomaly detection in disposal patterns, and optimization of collection routes. However, the effectiveness of these systems is contingent upon the quality and integrity of the data on which they are trained.

1.2 Motivation and Problem Statement

Despite growing interest in ML for waste management, a significant bottleneck exists in the data preparation stages—specifically in preprocessing and labeling. Raw datasets often contain noise, inconsistencies, missing values, and ambiguous labels, all of which negatively impact model accuracy. Moreover, many existing systems lack a standardized protocol for categorizing waste, resulting in poor generalizability and scalability.

This paper aims to fill that gap by:

- Developing a rigorous preprocessing pipeline for waste data (images and sensor data).
- Introducing a well-defined labeling strategy to improve classification accuracy.
- Evaluating the impact of preprocessing and labeling quality on ML model performance.

2. Related Work

2.1 Waste Classification with Machine Learning

Previous studies have utilized convolutional neural networks (CNNs) and traditional classifiers such as Random Forest (RF) and Support Vector Machines (SVM) for waste type prediction. Examples include:

- **TrashNet dataset** applied to CNN architectures like ResNet and VGGNet.
- Sensor-based datasets analyzed using decision trees and ensemble learning.

These models have shown promise but often suffer from overfitting and poor generalization due to inconsistent or imprecise labels.

2.2 Importance of Preprocessing

In ML pipelines, preprocessing involves cleaning, normalizing, transforming, and encoding raw data. For waste datasets, preprocessing steps often include:

- Resizing and normalizing images.
- Filling missing sensor readings.
- Removing irrelevant or redundant features.

Inadequate preprocessing can lead to suboptimal training and model instability.

2.3 Labeling Challenges

Labeling waste data is complex due to overlapping categories (e.g., biodegradable vs. recyclable), label ambiguity (e.g., mixed materials), and lack of domain expertise. Weak labeling often results in misleading training signals, severely affecting performance.

Our research uniquely focuses on enhancing these critical early-stage processes.

3. Methodology

3.1 Dataset Description

We utilize a hybrid dataset consisting of:

- **Image Data**: 3,000+ waste images across 6 classes: plastic, metal, glass, organic, paper, and mixed.
- **Sensor Data**: IoT sensor logs from smart bins measuring parameters like weight, type, odor level, and timestamp.

Each entry is tagged manually and verified by environmental experts.

3.2 Preprocessing Pipeline

3.2.1 For Image Data:

- **Resizing**: All images resized to 224×224 pixels for CNN compatibility.
- **Normalization**: Pixel values scaled between [0,1].
- **Augmentation**: Random flips, rotation, and brightness variation to increase robustness and reduce overfitting.
- **Noise Removal**: Median filtering for blurry or low-resolution images.

3.2.2 For Sensor Data:

- **Missing Value Imputation**: Median strategy for missing weight/odor readings.
- **Feature Scaling**: Min-max normalization applied.
- **Outlier Detection**: Z-score filtering to eliminate anomalies.
- **Time Series Smoothing**: Moving average to reduce temporal volatility.

3.3 Label Engineering Strategy

We introduced a multi-level labeling approach:

- **Level 1**: Broad category (e.g., organic, inorganic).
- **Level 2**: Specific waste type (e.g., plastic bottle, food waste).
- **Level 3**: Disposal type recommendation (e.g., recycle, compost, landfill).

Label verification was performed by two independent reviewers for quality assurance. Ambiguous items were flagged and relabeled using consensus.

3.4 Machine Learning Models

We tested the pipeline across:

- **Traditional ML**: Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN).
- **Deep Learning**: Convolutional Neural Networks (CNN), Transfer Learning using MobileNet and ResNet50.
- **Hybrid Models**: Combined sensor and image input using multi-modal architectures.

Models were evaluated using stratified 80/20 train-test splits and 5-fold cross-validation.

4. Experimental Results

4.1 Performance Metrics

International Research Journal of Multidisciplinary Sciences ISSN:AWAITED
VOL-1 Issue -4 April 2025 PP:23-30

We used the following metrics:

- **Accuracy**: Percentage of correctly classified instances.
- **Precision, Recall, F1-Score**: Evaluated per class to handle class imbalance.
- **Confusion Matrix**: For error analysis.
- **Training Time and Inference Speed**: For deployment feasibility.

4.2 Results Before and After Preprocessing

Model	Accuracy (Raw Data)	Accuracy (Processed Data)
SVM	63.4%	76.2%
RF	68.5%	82.1%
CNN (Custom)	71.2%	87.5%
ResNet50	75.4%	91.6%

Data preprocessing improved model performance by up to **18%** in accuracy. Noise removal and normalization had the highest impact.

4.3 Impact of Label Engineering

We evaluated models with and without hierarchical labeling.

Labeling Strategy	Accuracy	F1-Score
Flat Labeling	81.3%	78.4%
Hierarchical Labeling	89.7%	87.2%

Hierarchical labeling provided richer semantic context, improving model learning and generalization.

4.4 Multi-Modal Input Performance

Combining sensor and image data in a hybrid CNN + MLP model gave the best performance:

- **Accuracy**: 93.5%
- **Precision**: 92.1%
- **Recall**: 94.3%
- **F1-Score**: 93.2%

5. Discussion

5.1 Benefits of Robust Preprocessing

Our study reinforces that thoughtful preprocessing:

- Reduces noise and bias in training data.

- Enhances learning efficiency.
- Leads to faster convergence and improved generalization.

5.2 Importance of Intelligent Labeling

Labels are not just ground truth—they are the foundation of model learning. Poor labels can be more damaging than missing data. Our hierarchical scheme:

- Reduces misclassification.
- Enables multi-task learning (e.g., classification + recommendation).
- Scales better with complex datasets.

5.3 Limitations

- Manual labeling is time-consuming.
- Sensor reliability issues may affect real-time deployment.
- Image-based classification struggles with mixed-material waste.

5.4 Ethical and Environmental Implications

Automated waste classification systems can:

- Reduce human exposure to hazardous waste.
- Improve recycling efficiency.
- Minimize landfill overuse.

However, model misuse or poor deployment could lead to unintended consequences, such as misdirected waste streams.

6. Conclusion and Future Work

This research presents a comprehensive approach to building an intelligent waste management system powered by machine learning, with a strong emphasis on data preprocessing and labeling. Our experiments confirm that these foundational steps significantly affect overall system performance. The integration of hierarchical labels and multi-modal data inputs further enhances the system's robustness and accuracy.

Future Directions:

- Expand dataset with real-time streaming data.
- Use federated learning to train models without centralized data collection.
- Deploy edge AI on smart bins for on-device classification.
- Integrate active learning for continuous label improvement.

References

1. Abbasi M, El Hanandeh A. (2016) Forecasting municipal solid waste generation using artificial intelligence modelling approaches. Waste Management 56: 13–22. [DOI] [PubMed] [Google Scholar]

2. Abbasi M, Rastgoo MN, Nakisa B. (2019) Monthly and seasonal modeling of municipal waste generation using radial basis function neural network. Environmental Progress & Sustainable Energy 38: e13033. [Google Scholar]

3. Abu Qdais H, Shatnawi N. (2019) Assessing and predicting landfill surface temperature using remote sensing and an artificial neural network. International Journal of Remote Sensing 40: 9556–9571. [Google Scholar]

4. Abunama T, Othman F, Younes MK. (2018) Predicting sanitary landfill leachate generation in humid regions using ANFIS modeling. Environmental Monitoring and Assessment 190: 1–15. [DOI] [PubMed] [Google Scholar]

5. Abunama T, Othman F, Ansari M, et al. (2019) Leachate generation rate modeling using artificial intelligence algorithms aided by input optimization method for an MSW landfill. Environmental Science and Pollution Research 26: 3368–3381. [DOI] [PubMed] [Google Scholar]

6. Abushammala MFM, Basri NEA, Elfithri R, et al. (2014) Modeling of methane oxidation in landfill cover soil using an artificial neural network. Journal of the Air & Waste Management Association 64: 150–159. [DOI] [PubMed] [Google Scholar]

7. Adeogba E, Barty P, O'Dwyer E, et al. (2019) Waste-to-resource transformation: Gradient boosting modeling for organic fraction municipal solid waste projection. ACS Sustainable Chemistry & Engineering 7: 10460–10466. [Google Scholar]

8. Akanbi LA, Oyedele AO, Oyedele LO, et al. (2020) Deep learning model for demolition waste prediction in a circular economy. Journal of Cleaner Production 274: 122843. [Google Scholar]

9. Al-Refaie A, Al-Hawadi A, Fraij S. (2020) Optimization models for clustering of solid waste collection process. Engineering Optimization. Epub ahead of print 23 November 2020. DOI: 10.1080/0305215X.2020.1843165. [DOI] [Google Scholar]

10. Antanasijevic D, Pocajt V, Popovic I, et al. (2013) The forecasting of municipal waste generation using artificial neural networks and sustainability indicators. Sustainability science 8: 37–46. [Google Scholar]

11. Arebey M, Hannan MA, Begum RA, et al. (2012) Solid waste bin level detection using gray level co-occurrence matrix feature extraction approach. Journal of Environmental Management 104: 9–18. [DOI] [PubMed] [Google Scholar]

12. Ayeleru OO, Fajimi LI, Oboirien BO, et al. (2021) Forecasting municipal solid waste quantity using artificial neural network and supported vector machine techniques: A case study of Johannesburg, South Africa. Journal of Cleaner Production 289: 125671. [Google Scholar]

13. Azadi S, Karimi-Jashni A. (2016) Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate: A case study of Fars province, Iran. Waste Management 48: 14–23. [DOI] [PubMed] [Google Scholar]

14. Bagheri M, Esfilar R, Golchi MS, et al. (2019) A comparative data mining approach for the prediction of energy recovery potential from various municipal solid waste. Renewable and Sustainable Energy Reviews 116: 109423. [Google Scholar]

15. Bayram A, Kankal M, Ozsahin TS, et al. (2011) Estimation of the carbon to nitrogen (c:n) ratio in compostable solid waste using artificial neural networks. Fresenius Environmental Bulletin 20: 3250–3257. [Google Scholar]

16. Behera SK, Rene ER, Kim MC, et al. (2014) Performance prediction of a RPF-fired boiler using artificial neural networks. International Journal of Energy Research 38: 995–1007. [Google Scholar]

17. Bircanoğlu C, Atay M, Beser F, et al. (2018) RecycleNet: Intelligent waste sorting using deep neural networks. In: 2018 Innovations in intelligent systems and applications (INISTA), Thessaloniki, Greece, 3–5 July 2018, pp.1–7. New York, NY: IEEE. [Google Scholar]

18. Birgen C, Magnanelli E, Carlsson P, et al. (2021) Machine learning based modelling for lower heating value prediction of municipal solid waste. Fuel 283: 118906. [Google Scholar]

19. Boniecki P, Dach J, Pilarski K, et al. (2012) Artificial neural networks for modeling ammonia emissions released from sewage sludge composting. Atmospheric Environment 57: 49–54. [Google Scholar]

20. Bunsan S, Chen W, Chen H, et al. (2013) Modeling the dioxin emission of a municipal solid waste incinerator using neural networks. Chemosphere 92: 258–264. [DOI] [PubMed] [Google Scholar]

21. Carvalho DV, Pereira EM, Cardoso JS. (2019) Machine learning interpretability: A survey on methods and metrics. Electronics 8: 832. [Google Scholar]

22. Ceylan Z. (2020) Estimation of municipal waste generation of Turkey using socio-economic indicators by Bayesian optimization tuned Gaussian process regression. Waste Management & Research 38: 840–850. [DOI] [PubMed] [Google Scholar]

23. Ceylan Z, Bulkan S, Elevli S. (2020) Prediction of medical waste generation using SVR, GM (1,1) and ARIMA models: A case study for megacity Istanbul. Journal of Environmental Health Science and Engineering 18: 687–697. [DOI] [PMC free article] [PubMed] [Google Scholar]

24. Chang NB, Chen WC. (2000) Prediction of PCDDs/PCDFs emissions from municipal incinerators by genetic programming and neural network modeling. Waste Management & Research 18: 341–351. [Google Scholar]

25. Chang Y, Chiao H, Abimannan S, et al. (2020) An LSTM-based aggregated model for air pollution forecasting. Atmospheric Pollution Research 11: 1451–1463. [Google Scholar]

26. Chen J, Lin K. (2008) Diagnosis for monitoring system of municipal solid waste incineration plant. Expert Systems with Applications 34: 247–255. [Google Scholar]

27. Chhay L, Reyad MAH, Suy R, et al. (2018) Municipal solid waste generation in China: Influencing factor analysis and multi-model forecasting. Journal of Material Cycles and Waste Management 20: 1761–1770. [Google Scholar]

28. Chu Y, Huang C, Xie X, et al. (2018) Multilayer hybrid deep-learning method for waste classification and recycling. Available at: https://www.hindawi.com/journals/cin/2018/5060857 (accessed 10 December 2020). [DOI] [PMC free article] [PubMed]

29. Cubillos M. (2020) Multi-site household waste generation forecasting using a deep learning approach. Waste Management 115: 8–14. [DOI] [PubMed] [Google Scholar]

30. Da Paz DHF, Lafayette KPV, Holanda MJDO, et al. (2020) Assessment of environmental impact risks arising from the illegal dumping of construction waste in Brazil. Environment, Development and Sustainability 22: 2289–2304. [Google Scholar]

31. Demirbilek D, Onal AO, Demir V, et al. (2013) Characterization and pollution potential assessment of Tunceli, Turkey municipal solid waste open dumping site leachates. Environmental Monitoring and Assessment 185: 9435–9449. [DOI] [PubMed] [Google Scholar]

32. Díaz MJ, Eugenio ME, López F, et al. (2012) Neural models for optimizing lignocellulosic residues composting process. Waste and Biomass Valorization 3: 319–331. [Google Scholar]

33. Dissanayaka DMSH, Vasanthapriyan S. (2019) Forecast municipal solid waste generation in Sri Lanka. In: 2019 international conference on advancements in computing (ICAC), Malabe, Sri Lanka, 5–7 December 2019, pp.210–215. New York: IEEE. [Google Scholar]

34. Drudi KCR, Drudi R, Martins G, et al. (2019) Statistical model for heating value of municipal solid waste in Brazil based on gravimetric composition. Waste Management 87: 782–790. [DOI] [PubMed] [Google Scholar]

35. Duan N, Li D, Wang P, et al. (2020) Comparative study of municipal solid waste disposal in three Chinese representative cities. Journal of Cleaner Production 254: 120134. [Google Scholar]

36. Edgar E, Dan D, Jason J, et al. (2020) Trash to Treasure: Predicting Landfill Gas Flow to Optimize Electricity Generation. Journal of Information Systems Applied Research 13: 29–39. [Google Scholar]

37. Fallah B, Ng KTW, Hoang LV, et al. (2020) Application of a multi-stage neural network approach for time-series landfill gas modeling with missing data imputation. Waste Management 116: 66–78. [DOI] [PubMed] [Google Scholar]

38. Ferreira JA, Figueiredo MC, Oliveira JA. (2017) Household packaging waste management. In: Gervasi O, Murgante B, Misra S, et al. (eds) Computational Science and Its Applications – ICCSA 2017. Lecture Notes in Computer Science, vol. 10405. Cham, Switzerland: Springer, pp.611–620. [Google Scholar]

39. Gao MJ, Tian JW, Jiang W, et al. (2007) Research of sludge compost maturity degree modeling method based on wavelet neural network for sewage treatment. In: Li K, Fei M, Irwin GW, et al. (eds) Bio-Inspired Computational Intelligence and Applications. LSMS 2007. Lecture Notes in Computer Science, vol. 4688. Berlin, Germany: Springer, pp.608–618. [Google Scholar]

40. Hannan MA, Zaila WA, Arebey M, et al. (2014) Feature extraction using Hough transform for solid waste bin level detection and classification. Environmental Monitoring and Assessment 186: 5381–5391. [DOI] [PubMed] [Google Scholar]