Pol. J. Environ. Stud. Vol. XX, No. X (XXXX), 1-9
DOI: *10.15244/pjoes/237574* ONLINE PUBLICATION DATE: 2026-11-12

Original Research

Automated Biomedical Waste Segregation Using Deep Learning: A Duplicate-Aware Classification Benchmark and Disposal-Safety Assessment

Selin Vulga Işık, Atakan Işık*

Başkent University Biomedical Engineering Department, Başkent University Engineering Faculty B-309, Etimesgut/ANKARA

Received: 11 June 2026
Accepted: 20 September 2026

Abstract

Improper segregation of biomedical waste allows infectious and otherwise hazardous material to enter general waste streams, with consequences for public health and the environment. Reliable automatic classification at the point of generation could support safer segregation and disposal. Deep learning has been applied to this task, but the public image datasets used for training and evaluation are commonly distributed in a pre-augmented form, in which each source photograph appears as many transformed copies. A conventional random train–test split then places near-identical copies in both sets, so the reported accuracy partly reflects memorisation rather than generalisation. We study automatic classification of 23 biomedical waste categories (15,600 background-removed images) with explicit attention to this issue. Because the dataset carries no ground-truth labelling of duplicates, a leakage-free partition cannot be certified; instead, visually near-identical images are grouped from image content using ORB keypoints with RANSAC geometric verification and kept within a single split, and this controlled estimate is compared against the fully mixed random split, which is known to be leaked, as an optimistic reference. Across four transfer-learning backbones, ConvNeXt-Tiny performs best, reaching a macro-F1 of 0.847±0.007 under the controlled split against 0.958±0.003 under the random split, so naive evaluation is optimistic by about eleven points. Mapping the categories onto World Health Organization hazard streams yields a disposal-bin accuracy of 0.892, with 1.76% of hazardous items misrouted to general waste. The study offers a benchmark and a more realistic basis for reporting automatic biomedical-waste classification.

Keywords: biomedical waste segregation, deep learning, waste classification, disposal safety, evaluation reliability

*e-mail: atakani@baskent.edu.tr
Tel.: +90-537-733-1690
0000-0001-5433-4442

© 2026 This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

Introduction

Health-care activities generate large and rising volumes of waste, a small but dangerous fraction of which can transmit infection or cause physical and chemical injury. The World Health Organization estimates that about 85% of health-care waste is general, non-hazardous material, while the remaining 15% is hazardous, comprising infectious, pathological, sharps, pharmaceutical, or chemical waste [1]. When this hazardous fraction is not separated at the point of generation, it contaminates the whole waste stream, raises the cost and difficulty of treatment, and exposes waste handlers, patients, and the wider environment to infectious and toxic hazards [2]. Segregation at source is therefore the first and most consequential control step in health-care waste management, and the accuracy of that initial sorting decision determines how safely the downstream system can operate.

In most facilities, segregation still relies on manual sorting by clinical staff. Manual sorting is labour-intensive, varies between workers, and is difficult to audit, and a single misplaced item (a used scalpel dropped into general waste, for example) can move a sharps hazard into a stream that is never handled as hazardous. These limitations have motivated interest in automatic visual sorting, in which a camera and a classifier identify each item and direct it to the correct container. Convolutional neural networks suit this task because they learn discriminative visual features directly from images, and they have been applied successfully to general municipal and recyclable waste across categories such as plastic, glass, paper, and metal [3].

The same approach has been extended to biomedical waste. Recent studies train convolutional networks, often through transfer learning from large pretrained backbones, to distinguish categories such as syringes, vials, gloves, masks, and bandages, and they report classification accuracies between roughly 88% and 97% [4]. Taken at face value, these figures suggest that visual recognition of medical waste is close to solved. We argue that this impression is partly an artefact of how the underlying datasets are constructed and evaluated.

Public biomedical waste image datasets are typically small and are distributed in a pre-augmented form, in which each source photograph is expanded into several variants through flips, rotations, scaling, and photometric changes. When such a dataset is divided into training and test sets by drawing images at random, augmented copies of the same source photograph are almost certain to fall on both sides of the split. The classifier is then evaluated on near-duplicates of images it has already seen, and the reported accuracy reflects memorisation as much as generalisation. This failure mode (data leakage through near-duplicate images) is well documented in other imaging domains: it inflates test accuracy in deep-learning classification of optical coherence tomography scans [5], and large-scale audits have found the majority of test images duplicated in the training set of widely used benchmarks [6]. To our knowledge, its effect has not been examined for biomedical waste classification, even though the pre-augmented structure of these datasets makes them especially susceptible.

This study addresses that gap. Working with a public biomedical waste image dataset of 23 categories, we first show that the images are heavily duplicated by augmentation, and we build an evaluation protocol that accounts for it. Because the dataset provides no ground-truth labelling of which images share a common source, a leakage-free partition cannot be certified. Instead, we recover groups of visually near-identical images directly from image content using ORB keypoints with RANSAC geometric verification [7, 8], and we partition the data so that every image assigned to one group stays within a single split. This grouping is a content-based control rather than a validated separation; we therefore treat the conventional random split, in which augmented copies are known to be shared across the sets, as a fully leaked reference, and we read the difference between the two protocols as a measure of how optimistic random splitting is. Under this design, we re-examine how much of the previously reported performance survives once augmentation leakage is controlled, we compare modern classification backbones on equal terms, and we evaluate the outcome in disposal-oriented rather than purely statistical terms.

The contributions of this work are as follows. (i) We introduce a content-based deduplication and group-aware splitting procedure for pre-augmented biomedical waste images, and we show that it is needed by exposing the duplication that a random split conceals. (ii) Under identical training settings, we quantify the optimism of naive evaluation: a random split overstates macro-F1 by about eleven percentage points relative to the group-aware split. (iii) We benchmark four transfer-learning backbones (ConvNeXt, ResNet, Swin Transformer, and EfficientNet) under the group-aware protocol, with results reported across several random seeds. (iv) We map the 23 fine-grained categories onto the hazard streams defined by the WHO health-care waste guidance and report disposal-bin accuracy together with the rate at which hazardous items are misrouted to the general-waste stream, a quantity that bears directly on handler and environmental safety. The remainder of the paper describes the dataset and methods, presents these results, and discusses their implications for the design and reporting of automatic biomedical-waste classification systems.

Materials and Methods

Dataset

We used a publicly available biomedical waste image dataset containing 23 categories of clinical and related waste: sharp instruments (scalpels, several types of

surgical scissors, forceps, hemostats, and tweezers), used syringes, intact and broken ampoules, medicine and vaccine vials, syrup and disinfectant bottles, intravenous bottles, expired tablets, blood-soaked bandages, used masks, gloves and medical paper, human organs, general organic waste, water bottles, and a mixed "uncategorized or overlapping" class. The dataset is distributed in a pre-augmented form; the public release carries the export signature of the Roboflow platform, and each source photograph appears as several transformed copies. Two parallel versions are provided, one with the original photographs and one in which the background has been removed. We used the background-removed version, so that classification depends on the object rather than on incidental background cues. This subset contains 15,600 images, with 640 in each of 22 categories and 1,520 in the mixed category. Because the background-removed images are stored with a transparent alpha channel, every image was composited onto a constant black background and converted to three channels before use [9] (Fig .1).

Near-Duplicate Grouping and Data Partitioning

The pre-augmented structure means that a random train–test split would place transformed copies of the same photograph in different subsets, leaking information from training into testing. As the dataset records no link between an image and its source photograph, we recovered this grouping from image content. For every image we detected up to 300 ORB keypoints and their binary descriptors [7]. Each pair of images within a category was compared by matching descriptors with a nearest-neighbour ratio test (threshold 0.75), and the surviving matches were filtered with a RANSAC estimate of the homography between the two images (reprojection threshold 5 pixels) [8]. The number of geometrically consistent inlier matches, $m(i, j)$, measures how strongly the two images correspond. Two images were linked when this count reached a threshold τ:

$$e(i,j) = 1 if m(i,j) \geq \tau, otherwise\ 0 \quad (1)$$

with τ = 18. Taking images as vertices and the links of Eq. (1) as edges, each connected component of the resulting graph was treated as one group of near-identical images. Grouping was performed within each category, so images of different waste types were never merged. This procedure reduced the 15,600 images to 3,878 groups. The grouping is a content-based control rather than a validated separation: without ground-truth

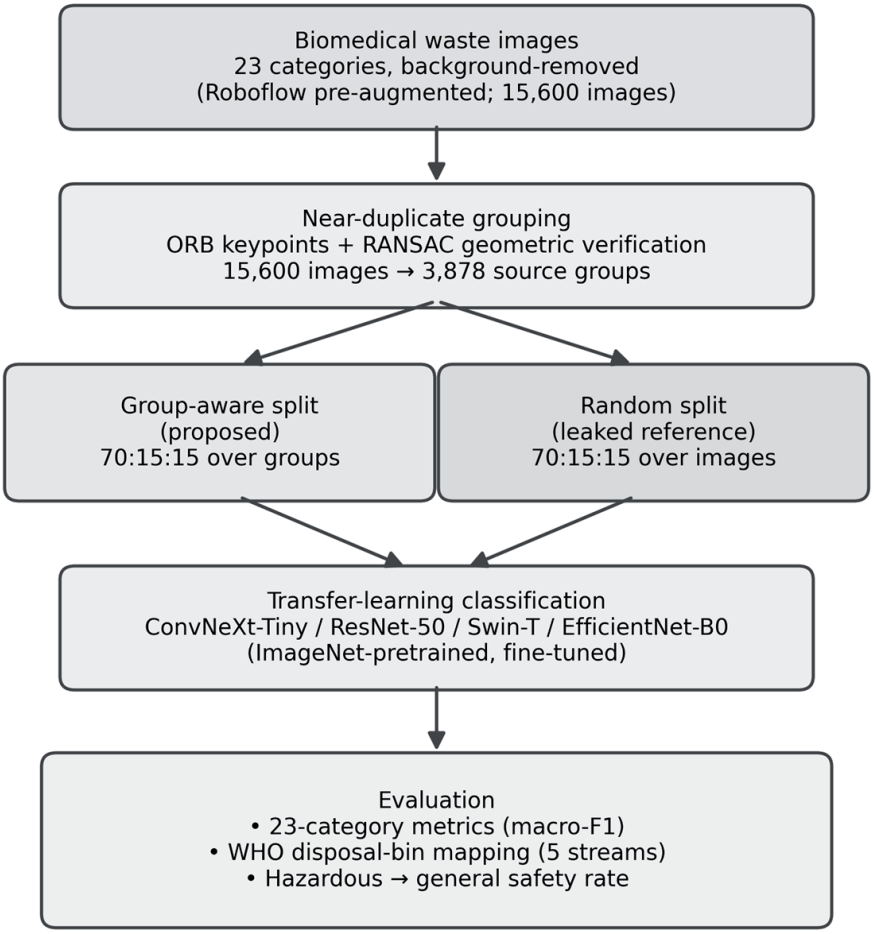


Fig. 1. Proposed pipeline for experiments.

duplicate labels, its precision cannot be measured, and it is conservative by design, because merging two distinct items only costs split diversity, whereas missing a true duplicate would reintroduce leakage.

The data were divided into training, validation, and test subsets in a target ratio of 70:15:15. Splitting was performed over groups rather than images, with a greedy class-stratified assignment that brought the per-category image counts close to the target while keeping every group wholly within one subset. Because a few groups are large and cannot be divided, the realised image proportions were approximately 81% for training and 9.5% each for validation and test. This group-aware split is used throughout the study. For comparison, we also produced a conventional split, in which the same images were assigned at random with stratification by category and no grouping; this split is known to contain augmentation leakage and serves as an optimistic reference.

Classification Models and Training

All classifiers were built by transfer learning from networks pretrained on ImageNet [10]. Four backbones spanning current convolutional and transformer designs were compared: ConvNeXt-Tiny [11], ResNet-50 [12], the tiny Swin Transformer [13], and EfficientNet-B0 [14]. Each network's classification head was replaced by a linear layer with 23 outputs, and the whole network was fine-tuned. Images were resized so that the longer side measured 224 pixels and were padded to a 224×224 square on the same black background, which preserves the aspect ratio of elongated instruments. Training images were augmented with random horizontal flips and small changes in brightness, contrast, and saturation; validation and test images were not augmented. Networks were optimised with AdamW [15] at a learning rate of 3×10^{-4} and a weight decay of 10^{-4}, under a cosine schedule over 15 epochs with mixed-precision arithmetic. The state giving the highest validation macro-F1 was retained, and training stopped early when this score did not improve for four consecutive epochs. Each configuration was trained three times with different random seeds, and we report the mean and standard deviation across seeds. The experiments were implemented in PyTorch [16] with the timm library [17] on a single graphics processing unit.

Evaluation

Performance was summarised by per-category precision and recall, their harmonic mean (the F1 score), the unweighted mean of the F1 scores over categories (macro-F1), and overall accuracy. For a category c with TP_c true positives, FP_c false positives, and FN_c false negatives,

$$P_c = TP_c/(TP_c + FP_c) \tag{2}$$

$$R_c = TP_c/(TP_c + FN_c) \tag{3}$$

$$F1_c = 2P_cR_c/(P_c + R_c) \tag{4}$$

and, for C categories and N test images,

$$macro - F1 = (1/C)\Sigma(c = 1toC)F1_c \tag{5}$$

$$Accuracy = (1/N)\Sigma(c = 1toC)TP_c \tag{6}$$

Macro-F1 was the primary measure, because it weights every category equally and is therefore sensitive to the minority categories. To express performance in operational terms, the 23 categories were mapped onto the health-care waste streams of the WHO guidance [WHO 2014]: sharps, infectious and pathological waste, pharmaceutical and chemical waste, and general waste, with the mixed category kept separate. Predictions were re-scored at this level, and we also report the share of genuinely hazardous items assigned to the general-waste stream,

$$R_{unsafe} = |i: b(y_i) \in H \wedge b(\hat{y}_i) = general|/|i: b(y_i) \in H| \tag{7}$$

where $b(\cdot)$ maps a category to its hazard stream, H is the set of hazardous streams, and y_i and $\hat{y}_i$ are the true and predicted categories of image i. This is the most consequential error type, in which a hazardous item is sent to a stream that receives no hazardous-waste treatment. Ninety-five percent confidence intervals for the headline scores were obtained by bootstrap resampling of the test set.

Results and Discussion

Duplication and Group Structure

Content-based grouping reduced the 15,600 images to 3,878 source groups, an average of about four images per group. The amount of duplication varied widely between categories: the surgical-instrument classes collapsed into very few groups (for example, seven for scalpels and nine for one type of scissors), whereas visually diverse classes such as expired tablets and water bottles produced several hundred groups each. This pattern reflects both genuine differences in how many distinct photographs each category contains and the limits of the matching procedure, which links instruments of near-identical shape more readily than textured, varied objects. Because the grouping is content-based and unvalidated, these counts are best read as a conservative lower bound on the number of independent images; even so, they show that the nominal 640 images per category correspond to far fewer independent photographs. Fig. 2 shows the confusion matrix of 23 groups predicted by the model.

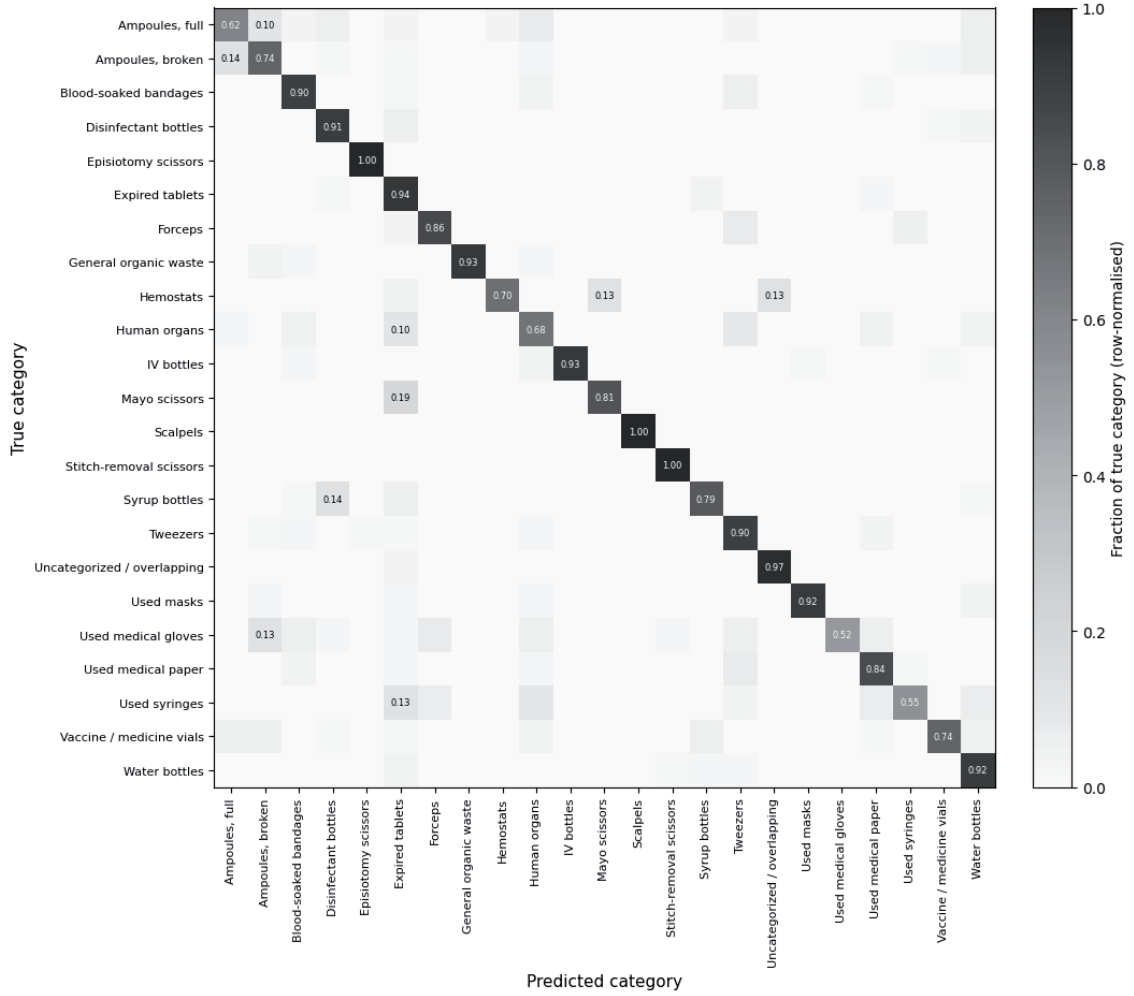


Fig. 2. Confusion matrix of the 23 groups predicted by the model.

Effect of Data Leakage on Reported Performance

Table 2 contrasts the two protocols for the proposed model. Under the random split, ConvNeXt-Tiny reached a macro-F1 of 0.958±0.003; under the group-aware split it reached 0.847±0.007, with the same contrast in accuracy (0.960 against 0.838). The random protocol therefore overstates macro-F1 by about eleven percentage points and accuracy by about twelve. The standard deviations across seeds are small (no more than 0.010), so the gap is far larger than the training noise; a bootstrap 95% confidence interval for the group-aware macro-F1 is 0.80-0.85, well below the random-split value. Because the random split places augmented copies of the same photograph in both the training and the test set, its score reflects partial memorisation, and the difference between the two protocols is a direct measure of the optimism that uncontrolled splitting introduces on this dataset.

Comparison of Classification Backbones

Table 1 compares four backbones under the group-aware split. ConvNeXt-Tiny gave the best result on both accuracy and macro-F1 and was adopted as the proposed model. The ranking of the other three depends

Table 1. Classification performance of four backbones under the group-aware split (single training run).

Model	Accuracy	Macro-F1	Weighted-F1	Params (M)
ConvNeXt-Tiny	0.831	0.831	0.830	27.8
EfficientNet-B0	0.806	0.791	0.806	4.0
ResNet-50	0.793	0.808	0.792	23.6
Swin-Tiny	0.776	0.793	0.777	27.5

Table 2. Effect of the evaluation protocol on the proposed model (ConvNeXt-Tiny).

Split protocol	Accuracy	Macro-F1
Group-aware (controlled)	0.838±0.010	0.847±0.007
Random (leaked, reference)	0.960±0.003	0.958±0.003
Difference (optimism)	+0.122	+0.111

Note: mean±standard deviation over three random seeds.

on the measure: EfficientNet-B0 was second on accuracy (0.806) but last on macro-F1 (0.791), while ResNet-50 showed the reverse (0.793 accuracy, 0.808 macro-F1). EfficientNet-B0 reached its accuracy with only 4.0 million parameters, about one-seventh of ConvNeXt-Tiny, which is attractive for on-site deployment, although the lower macro-F1 indicates weaker performance on the minority categories. The tiny Swin Transformer was the weakest of the four, consistent with the larger data requirements of vision transformers relative to the limited number of independent images available here.

Per-Category Performance

Table 3 reports per-category precision, recall, and F1 for the proposed model. Performance ranged from an F1 of 0.58 for intact ampoules, the weakest category, to above 0.95 for several visually distinctive classes

Table 3. Per-category classification performance of the proposed model (ConvNeXt-Tiny, group-aware split).

Category	Precision	Recall	F1	n
Ampoules, full	0.545	0.615	0.578	39
Ampoules, broken	0.772	0.740	0.755	96
Blood-soaked bandages	0.827	0.896	0.860	96
Disinfectant bottles	0.821	0.906	0.861	96
Episiotomy scissors	0.938	1.000	0.968	15
Expired tablets	0.682	0.938	0.789	96
Forceps	0.857	0.857	0.857	42
General organic waste	1.000	0.927	0.962	96
Hemostats	0.941	0.696	0.800	23
Human organs	0.707	0.677	0.691	96
IV bottles	1.000	0.927	0.962	96
Mayo scissors	0.812	0.812	0.812	16
Scalpels	1.000	1.000	1.000	6
Stitch-removal scissors	0.667	1.000	0.800	4
Syrup bottles	0.884	0.792	0.835	96
Tweezers	0.741	0.896	0.811	96
Uncategorized / overlapping	0.923	0.973	0.947	37
Used masks	0.983	0.919	0.950	62
Used medical gloves	1.000	0.519	0.683	54
Used medical paper	0.835	0.844	0.839	96
Used syringes	0.810	0.548	0.654	31
Vaccine / medicine vials	0.947	0.740	0.830	96
Water bottles	0.800	0.917	0.854	96
Macro average	0.847	0.832	0.831	1481

Table 4. Disposal-bin performance under the WHO health-care waste streams (proposed model).

Hazard stream	Precision	Recall	F1	n
Sharps	0.837	0.845	0.841	329
Infectious and pathological	0.938	0.878	0.907	500
Pharmaceutical and chemical	0.883	0.924	0.903	423
General	0.889	0.922	0.905	192
Unclassified	0.923	0.973	0.947	37

Note: Overall bin accuracy 0.892; macro-F1 0.901; hazardous-to-general misrouting 1.76%.

such as intravenous bottles, general organic waste, and used masks. Errors concentrated in a few visually similar groups: intact ampoules, broken ampoules, and medicine or vaccine vials were frequently confused with one another, and the thin metal instruments (forceps, hemostats, tweezers) were also mutually confused. Two categories combined high precision with low recall: used medical gloves (precision 1.00, recall 0.52) and used syringes (precision 0.81, recall 0.55), so items of these types were more often labelled as something else than mislabelled in turn. The instrument categories that deduplication had collapsed into very few groups, such as scalpels and stitch-removal scissors, reached high F1 on only four to six test images, so their individual scores are statistically unreliable and should be read together with the disposal-bin results.

Disposal-Bin Performance and Safety

Mapping the 23 categories onto the WHO hazard streams gave a coarser but operationally more meaningful view (Table 4). At this level the proposed model reached an accuracy of 0.892 and a macro-F1 of 0.901, several points above the 23-category figures, because most category errors fall within a single disposal stream and so do not change the bin decision. The sharps stream was the hardest (F1 0.841), reflecting the difficulty of the instrument categories. Most important for safety, only 1.76% of genuinely hazardous items were assigned to the general-waste stream, and the large majority of bin-crossing errors stayed among the hazardous streams, where the handling consequences are smaller. Even at the accuracy reported here, the system therefore rarely commits the most dangerous type of sorting error, in which a hazardous item is sent to a stream that receives no hazardous-waste treatment.

Implications and Limitations

The size of the gap between the two protocols bears directly on how published results should be read. Accuracies of 95% and above have been reported for biomedical waste classification on small, augmented datasets [4, 18-20], and our random-split figure (0.958) falls in the same range; once leakage is controlled, the same model on the same data reaches 0.847. This does not show that earlier studies are wrong, but it indicates that scores obtained from random splits of pre-augmented datasets are likely optimistic and are not a reliable guide to performance on genuinely new images. Group-aware evaluation, or an equivalent control for near-duplicate images, should therefore become a routine part of reporting in this area. At the same time, the disposal-bin and safety results are encouraging: at the level that governs how waste is actually handled, the model is correct about nine times in ten and rarely sends a hazardous item to general waste, which suggests that automatic visual sorting is worth developing further as an aid to segregation, provided it is evaluated honestly.

Several limitations qualify these findings. The grouping on which the controlled split rests is recovered from image content and cannot be validated, because the dataset gives no record of which images share a source; the procedure is conservative, but its precision is unknown, and it is sensitive to texture, merging visually similar instruments too readily while leaving low-texture categories under-grouped. The study uses a single public dataset whose images are already background-removed and were augmented by an external tool, so the absolute accuracies need not transfer to raw photographs or to other waste streams. The deduplication also reduces some categories to very few independent groups, which makes their per-category and per-split estimates unstable, as seen in the small instrument classes. Finally, the mapping from categories to WHO streams is one reasonable scheme among several, and a different national classification could shift the bin-level and safety figures. These limitations point to clear next steps: validating the grouping against labelled duplicates, testing on raw images and additional datasets, and enlarging the under-represented categories.

Conclusions

We examined automatic visual classification of biomedical waste under an evaluation protocol designed to remove the optimism that pre-augmented datasets introduce. Grouping near-identical images from their content with ORB and RANSAC, and then splitting over groups rather than over individual images, lowered the macro-F1 of the best model from 0.958 under

a conventional random split to 0.847, a difference of about eleven points that reflects near-duplicate leakage rather than any change in the data or the model. ConvNeXt-Tiny was the strongest of the four backbones tested, and when its predictions were mapped onto the WHO health-care waste streams, the system reached a disposal-bin accuracy of 0.892 and sent only 1.76% of hazardous items to the general-waste stream.

Two conclusions follow. First, automatic sorting is a credible aid to safe segregation: at the level that determines how waste is handled, the model is reliable and rarely commits the most dangerous error. Second, and more generally, accuracy reported from random splits of augmented waste datasets should be treated with caution, because much of it can arise from memorised duplicates; controlling for near-duplicate images, through group-aware splitting or an equivalent check, is needed before such figures can be trusted. As the grouping used here is content-based and could not be validated against ground truth, we regard the controlled score as a realistic rather than an exact estimate. Validating the grouping against labelled duplicates, testing on raw images and additional datasets, and strengthening the weakest categories are the main priorities for future work.

Conflict of Interest

The authors declare no conflict of interest.

AI Usage Disclosure Statement

The authors utilized Claude (Anthropic) for language editing, stylistic refinement, and structural organization of the text. Following the use of this tool, the authors reviewed and edited all content as needed and take full responsibility for the content of the publication.

References

1. WORLD HEALTH ORGANIZATION. Safe Management of Wastes from Health-Care Activities, 2nd ed.; Editors: Chartier Y., Emmanuel J., Pieper U., Prüss-Ustün A., Rushbrook P., Stringer R., Townend W., Wilburn A., Zghondi R. World Health Organization: Geneva, Switzerland, pp. 326, **2014**.
2. KENNY C., PRIYADARSHINI A. Review of Current Healthcare Waste Management Methods and Their Effect on Global Health. Healthcare, **9** (3), 284, **2021**.
3. CHU Y., HUANG C., XIE X., TAN B., KAMAL S., XIONG X. Multilayer Hybrid Deep-Learning Method for Waste Classification and Recycling. Computational Intelligence and Neuroscience, **2018** (1), 5060857, **2018**.
4. ZHOU H., YU X., ALHASKAWI A., DONG Y., WANG Z., JIN Q., HU X., LIU Z., KOTA V.G., ABDULLA M.H.A.H., EZZI S.H.A., QI B., LI J., WANG B., FANG J., LU H. A Deep Learning Approach for Medical Waste Classification. Scientific Reports, **12** (1), 2159, **2022**.
5. TAMPU I.E., EKLUND A., HAJ-HOSSEINI N. Inflation of Test Accuracy Due to Data Leakage in Deep Learning-Based Classification of OCT Images. Scientific Data, **9** (1), 580, **2022**.
6. ADIMOOLAM Y.K., POULLIS C., AVERKIOU M. Data Leakage Detection and De-Duplication in Large Scale Geospatial Image Datasets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Nashville, TN, USA, pp. 72-81, **2026**.
7. RUBLEE E., RABAUD V., KONOLIGE K., BRADSKI G. ORB: An Efficient Alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision; IEEE: Barcelona, Spain, pp. 2564-2571, **2011**.
8. FISCHLER M.A., BOLLES R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM, **24** (6), 381, **1981**.
9. Biomedical Waste Classification Images Dataset. Available online: https://www.kaggle.com/datasets/mario78/biomedical-waste-classification-images-dataset (accessed on 14 June 2026).
10. DENG J., DONG W., SOCHER R., LI L.-J., LI K., LI F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Miami, FL, USA, pp. 248-255, **2009**.
11. LIU Z., MAO H., WU C.-Y., FEICHTENHOFER C., DARRELL T., XIE S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: New Orleans, LA, USA, pp. 11966-11976, **2022**.
12. HE K., ZHANG X., REN S., SUN J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Las Vegas, NV, USA, pp. 770-778, **2016**.
13. LIU Z., LIN Y., CAO Y., HU H., WEI Y., ZHANG Z., LIN S., GUO B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: Montreal, QC, Canada, pp. 9992-10002, **2021**.
14. TAN M., LE Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICML); PMLR: Long Beach, CA, USA, **97**, pp. 6105-6114, **2019**.
15. LOSHCHILOV I., HUTTER F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR); New Orleans, LA, USA, **2019**.
16. PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KÖPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J., CHINTALA S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32 (NeurIPS); Curran Associates: Red Hook, NY, USA, pp. 8024-8035, **2019**.
17. WIGHTMAN R., RAW N., SOARE A., ARORA A., HA C., REICH C., GUAN F., KACZMARZYK J., MRT23,

MIKE, SEEFUN, CONTRASTIVE, RIZIN M., KIM H., KERTÉSZ C., MEHTA D., CUCURULL G., SINGH K., HANKYUL, UCHIDA Y. Rwightman/pytorch-image-models: v0.8.10dev0 Release (Versions v0.8.10dev0) [Computer software]. Zenodo, **2023**.

18. VERMA M., KUMAR A., KUMAR S. Medical Waste Classification Using Deep Learning and Convolutional Neural Networks. In Proceedings of the 2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI); IEEE: Gwalior, India, pp. 1-5, **2022**.
19. KUNWAR S., RAI P. Health Care Waste Classification Using Deep Learning Aligned with Nepal's Bin Color Guidelines. arXiv, 2025. Available online: https://arxiv.org/abs/2508.07450 (accessed on 14 June 2026).
20. MOKTAR M.H.B., HAJJAJ S.S.H., MOHAMED H. Medical Waste Detection and Classification Through YOLO Algorithms. In Robot Intelligence Technology and Applications 8 (RiTA 2023); Lecture Notes in Networks and Systems, Springer Nature Switzerland: Cham, Switzerland, pp. 22-33, **2024**.