Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16531
e-ISSN : 2541-2019
p-ISSN : 2541-044X

# Incremental Effects of Augmentation Strategies on Pre-Augmented Biomedical Waste Detection Using YOLOv11n

**Jevri Tri Ardiansah[1)*], Ahmad Kholish Fauzan Shobiry[2)], Anik Nur Handayani[3)], Mohammad Muzayyin Amrulloh[4)], Taiga Haruta[5)]**
[1,2,3,4)] Universitas Negeri Malang, Indonesia, [5)] Kurume Institute of Technology, Japan
[1)]jevri.ardiansah.ft@um.ac.id, [2)]ahmad.kholish.2505348@students.um.ac.id, [3)]aniknur.ft@um.ac.id, [4)]muzayyin.amrulloh.ft@um.ac.id, [5)]haruta@kurume-it.ac.jp

**Submitted**: June 20, 2026 | **Accepted**: July 14, 2026 | **Published**: July 21, 2026

**Abstract:** Biomedical waste carries infectious and hazardous risk that makes accurate automated sorting valuable, and object detection offers a path toward automation, yet published studies rarely measure how much online data augmentation contributes once training images carry offline augmentation and mean Average Precision alone can conceal how augmentation shifts the balance between missed detections and false alarms.This study measures the incremental effect of Mosaic, MixUp, and Copy-Paste augmentation on YOLOv11n trained for biomedical waste detection, and verifies whether each augmentation executes as configured. We designed a 2×2×2 factorial ablation across eight configurations, trained YOLOv11n three times per configuration with different random seeds on a 14-class biomedical waste dataset and evaluated each run's best checkpoint on a held-out test set using mean Average Precision, precision, recall, and a corrected confusion matrix retaining the background class. We verified framework behavior through prediction-level identity, loss-level comparison, and instrumented tracing. Configurations combining Mosaic and MixUp raised mean Average Precision by 0.0053 over baseline, MixUp alone raised recall from 0.8950 to 0.9071, and Mosaic with MixUp raised precision to 0.9650. False negatives outnumbered false positives three to one across every configuration, identifying class-versus-background rather than inter-class confusion as the dominant failure mode. Copy-Paste showed zero measurable effect once verified, consistent with its restriction to segmentation tasks in the underlying framework. Online augmentation produces modest but distinguishable, metric-specific gains on pre-augmented biomedical waste data, and verifying framework behavior before attributing results to an augmentation strategy is necessary for reliable reporting.

**Keywords:** biomedical waste, confusion matrix, data augmentation, object detection, YOLOv11n

## INTRODUCTION

Healthcare systems worldwide generate waste alongside a mounting global disposal burden across all waste categories, projected to reach 3.4 billion tons by 2050 (Aberger et al., 2025). A fraction of that stream carries infectious or hazardous properties specific to clinical environments: used syringes, contaminated gauze, glass vials, and other sharps and biological materials that pose direct transmission risk when mishandled (Menekşe & Camgöz Akdağ, 2023; Yazdani, Tavana, Pamučar, & Chatterjee, 2020). Fault tree analyses of hospital waste streams identify segregation as the highest-risk stage in the disposal chain, with missing containers and inconsistent bin labeling driving most safety failures (Carnero, 2020).

Segregation still relies overwhelmingly on manual sorting in most facilities, and innovation in that process has stagnated even as waste volumes climb (Aberger et al., 2025). Manual sorting exposes workers to repeated contact with hazardous material and does not scale with rising healthcare demand. Computer vision has been proposed to automate this bottleneck, with reported sorting accuracies for general waste streams ranging from 72.8% to 99.95% depending on material type and model choice (Fang et al., 2023; Lu & Chen, 2022; Nahiduzzaman et al., 2025). Within the clinical domain specifically, deep learning classifiers have reached

*Jevri Tri Ardiansah

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16531
e-ISSN : 2541-2019
p-ISSN : 2541-044X

comparably strong results: a ResNeXt-based model achieved 97.2% accuracy across eight medical waste categories (Zhou et al., 2022), and a YOLOv5-based detector reached 95.06% accuracy on hospital waste aligned with regional bin-color guidelines (Bruno et al., 2023; Kunwar & Rai, 2025). IoT-integrated frameworks combining convolutional networks with smart bins have also been proposed to reduce manual intervention in hazardous waste handling (Kumar, Ali, Kumar, Assaf, & Ilyas, 2025; Nasir, Hossain, Islam, Islam, & Islam, 2024).

These results demonstrate feasibility, not methodological completeness. A bibliometric analysis of medical waste research spanning nearly five decades shows a sharp rise in publications after 2000 but limited depth in evaluation methodology across that growth (Cirstea et al., 2025). Two gaps stand out against the broader object detection literature. Image augmentation strategies such as Mosaic, MixUp, and Copy-Paste have no single best configuration; effectiveness depends on dataset-specific factors including class balance, object scale, and scene composition, yet most applied waste-detection studies apply a fixed augmentation recipe without ablating its components (Yang et al., 2023; Yu et al., 2024). Separately, mean Average Precision, the metric nearly all of these studies rely on, is a ranking-based measure that does not distinguish false positives from false negatives at the operating point that matters for deployment (Padilla, Netto, & da Silva, 2020).

No existing study, to our knowledge, combines a factorial ablation of individual and combined augmentation strategies with a verification protocol confirming that each augmentation genuinely executes as configured, on a biomedical waste dataset that already carries offline augmentation applied upstream. This gap carries direct safety weight in our setting. A missed detection lets a contaminated item bypass sorting, so an augmentation strategy adopted on the strength of an aggregate mAP gain can quietly increase that exact risk if the gain comes from fewer false alarms rather than fewer missed items. A strategy credited with a gain that never actually executed during training compounds the problem: it gives practitioners false confidence in a lever that did nothing. We address this gap through three research questions. RQ1 asks how much each augmentation configuration changes detection performance once online augmentation is layered on training images that already carry offline augmentation. RQ2 asks how that change trades off between precision and recall rather than collapsing into a single aggregate number. RQ3 asks whether each augmentation actually executes as configured, and whether that execution can be verified independently of the metrics it is credited with producing. We answer these questions with a 2×2×2 factorial ablation of Mosaic, MixUp, and Copy-Paste on YOLOv11n, evaluated on a 14-class biomedical waste dataset across three random seeds, with prediction-level and loss-level identity checks confirming augmentation execution. The resulting evidence supports concrete augmentation-selection guidance for biomedical waste detection under class imbalance and safety-critical deployment constraints.

## METHODS

### Research Workflow

The study proceeds through five stages: dataset inspection, factorial experimental design, model training across eight augmentation configurations with three seeds each, test-set evaluation, and a verification stage confirming the internal consistency of framework behavior. Figure 1 illustrates this pipeline end-to-end.

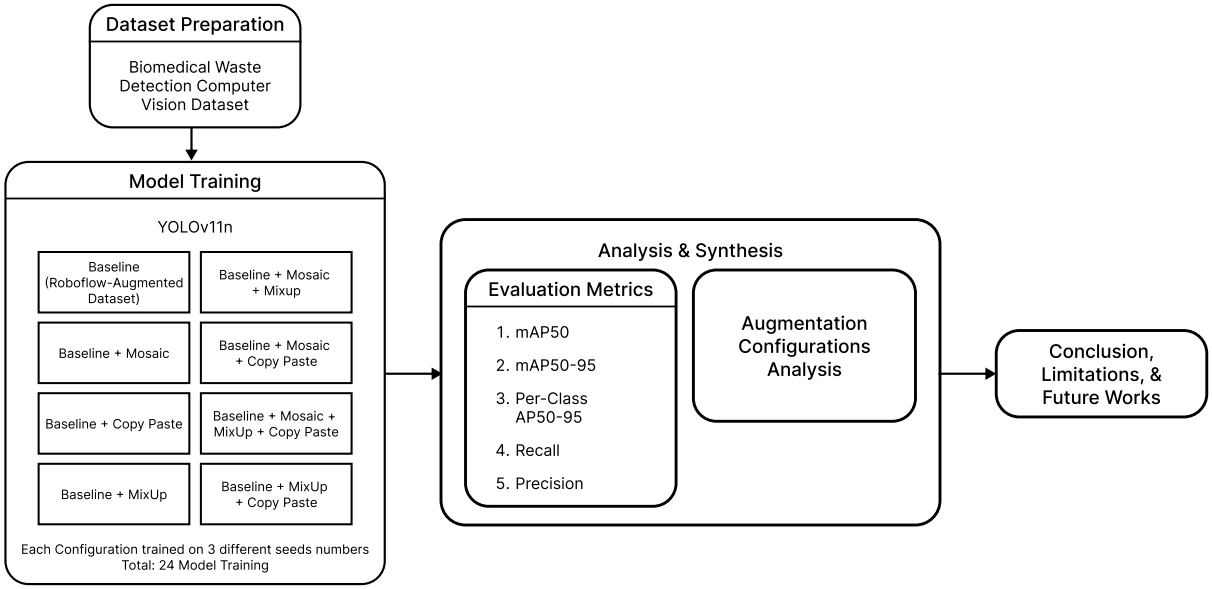


Fig. 1 Research Workflow Diagram

*Jevri Tri Ardiansah

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16531
e-ISSN : 2541-2019
p-ISSN : 2541-044X

**Dataset and Preprocessing**

We use a biomedical waste detection dataset sourced from Roboflow with 7983 training images, 571 validation images, and 570 test images across 14 classes, totaling 15,029 annotated instances (QuratulAin, 2025). Table 1 reports the per-class instance count across all three splits. The dataset carries offline augmentation from the Roboflow platform (flip, rotation, brightness, saturation, blur, noise), fixed and inseparable from the raw images. Figure 2 shows six annotated samples across different classes; Figure 3 shows one class under the baseline, Mosaic, and MixUp settings.

Table 1 Per-class instance counts across train, validation, and test splits

| Class | Train | Valid | Test | Class | Train | Valid | Test |
|---|---|---|---|---|---|---|---|
| **fluid bottle** | 762 | 49 | 55 | **plastic** | 246 | 14 | 24 |
| **gauze** | 1365 | 108 | 97 | **scrub** | 57 | 3 | 6 |
| **glass** | 66 | 2 | 2 | **test tube** | 831 | 64 | 61 |
| **glove** | 2463 | 177 | 156 | **tube** | 978 | 78 | 73 |
| **mask** | 1095 | 78 | 86 | **unused gauze** | 225 | 24 | 18 |
| **needle** | 1323 | 99 | 109 | **used syringe** | 1977 | 108 | 142 |
| **needle cap** | 213 | 21 | 17 | **paper** | 1563 | 95 | 99 |

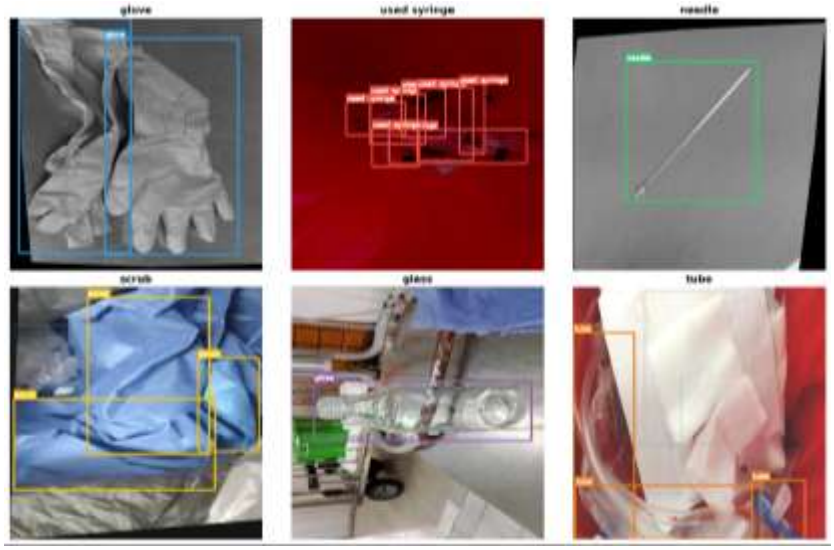


Fig. 2 Six annotated samples

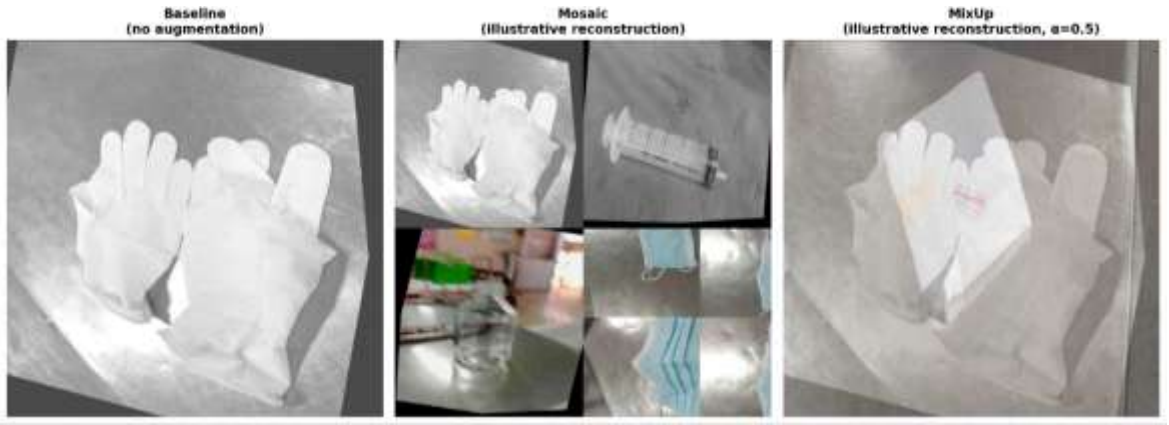


Fig. 3 One class under baseline, Mosaic, and MixUp augmentation

*Jevri Tri Ardiansah

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16531
e-ISSN : 2541-2019
p-ISSN : 2541-044X

**Model Architecture**

We use YOLOv11n, the nano variant of Ultralytics YOLOv11, pretrained on COCO. YOLOv11 replaces earlier C2f blocks with C3k2 modules and adds SPPF and C2PSA for spatial attention, improving feature extraction at low parameter count (Khanam & Hussain, 2024). Recent surveys trace this progression across the YOLO family and its trade-offs between speed and accuracy (Ali & Zhang, 2024). The nano variant totals 2.59 million parameters, fitting our single-GPU budget across 24 runs. The architecture stays fixed throughout so that every observed effect traces to augmentation rather than model capacity.

**Experimental Design**

We adopt a 2×2×2 factorial ablation over three factors, Mosaic, MixUp, and Copy-Paste, producing eight configurations (A through H) listed in Table 2. Factorial designs isolate the individual and combined contribution of each factor within a bounded number of trials (Fostiropoulos & Itti, 2023; Sheikholeslami et al., 2021).

Table 2 Factorial configuration matrix

| Config | Mosaic | MixUp | Copy-Paste | Config | Mosaic | MixUp | Copy-Paste |
|---|---|---|---|---|---|---|---|
| **A: Baseline** | off | off | off | **E: Mosaic + MixUp** | 1.0 | 0.15 | off |
| **B: Mosaic** | 1.0 | off | off | **F: Mosaic + CopyPaste** | 1.0 | off | 0.1 |
| **C: MixUp** | off | 0.15 | off | **G: MixUp + CopyPaste** | off | 0.15 | 0.1 |
| **D: Copy-Paste** | off | off | 0.1 | **H: All** | 1.0 | 0.15 | 0.1 |

**Training Configuration**

Each configuration trains for 100 fixed epochs without early stopping, at image size 640 with automatic batch sizing and the auto-selected AdamW optimizer. We repeat every configuration with three seeds, 42, 123, and 2025, and report mean and standard deviation. A single seed can produce an outlier that misrepresents a configuration's true behavior (Picard, 2023) and reporting one run risks drawing conclusions from noise rather than signal (Gundersen, Shamsaliei, Kjærnli, & Langseth, 2023). Three seeds keep the 24-run compute budget tractable while giving a defensible variance estimate.

**Evaluation and Verification Protocol**

We evaluate the best.pt checkpoint of each run on the held-out test set, touched exactly once. Detection quality rests on Intersection over Union between a predicted box and its ground truth, from which Precision and Recall follow using true positives (TP), false positives (FP), and false negatives (FN). We state each metric in closed form below rather than deferring to external documentation, following recent calls for explicit, reproducible metric definitions in YOLO-based studies (Rana & Vaidya, 2026):

$$IoU = \frac{B_{pred} \cap B_{gt}}{B_{pred} \cup B_{gt}} \quad (1)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

Average Precision integrates Precision over Recall for one class and mean Average Precision averages AP across all N classes. mAP50 fixes the IoU threshold at 0.50; mAP50-95 averages mAP across ten thresholds from 0.50 to 0.95, following the COCO convention and rewarding tighter localization (Rezatofighi et al., 2019):

$$AP = \int_0^1 P(R)dR \quad (4)$$

*Jevri Tri Ardiansah

e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \qquad (5)$$

During analysis we observed that certain configuration pairs produced numerically identical results, prompting a verification step beyond standard evaluation. We checked this at two levels: comparing raw predictions (box coordinates and confidence) between paired configurations on a fixed sample of test images and comparing recorded training loss at every epoch between the same pairs. We also instrumented the relevant Ultralytics augmentation class with a call counter to confirm whether it executes during training. This protocol serves value verification only, not a novel evaluation methodology; comparable instrumentation and differential output checks are established practice in software testing for deep learning libraries (Wang, Lutellier, Qian, Pham, & Tan, 2022; Wei, Deng, Yang, & Zhang, 2022; X. Zhang et al., 2025), where silent inconsistencies propagate undetected through training pipelines when outputs are not directly compared.

**Computational Environment**

All experiments ran on a single workstation (Intel i7-13700KF, 16GB RAM, NVIDIA RTX 3060 with 12GB VRAM) under Windows 11, using Python 3.10.16, PyTorch 2.9.0 with CUDA 12.6, and Ultralytics 8.3.223.

## RESULTS

We organize the results as objective reporting across four analytical dimensions before the discussion. We first present the aggregate detection metrics, followed by per-class AP performance and confusion matrix outcomes together with the associated failure modes. We then report the verification of Copy-Paste behavior before interpreting these observations in the context of the broader literature.

**Overall Performance on the Test Set**

Table 3 reports aggregate test-set metrics across the eight configurations. Each configuration was trained three times with random seeds 42, 123, and 2025, and evaluated using the best.pt checkpoint selected on the validation split. Values are reported as mean and standard deviation over the three seeds.

Table 3 Aggregate detection metrics on the test set (mean ± std, n=3 seeds)

| Configuration | mAP50 | mAP50-95 | Precision | Recall |
|---|---|---|---|---|
| **A: Baseline** | 0.9488 ± 0.0008 | 0.7348 ± 0.0038 | 0.9012 ± 0.0177 | 0.9384 ± 0.0190 |
| **B: Mosaic** | 0.9528 ± 0.0063 | 0.7392 ± 0.0053 | 0.9191 ± 0.0153 | 0.9182 ± 0.0101 |
| **C: MixUp** | 0.9510 ± 0.0041 | 0.7336 ± 0.0098 | 0.8998 ± 0.0125 | 0.9327 ± 0.0221 |
| **D: Copy-Paste** | 0.9521 ± 0.0042 | 0.7353 ± 0.0020 | 0.9158 ± 0.0182 | 0.9315 ± 0.0193 |
| **E: Mosaic+MixUp** | 0.9586 ± 0.0071 | 0.7401 ± 0.0040 | 0.9172 ± 0.0109 | 0.9364 ± 0.0193 |
| **F: Mosaic+CopyPaste** | 0.9528 ± 0.0063 | 0.7392 ± 0.0053 | 0.9191 ± 0.0153 | 0.9182 ± 0.0101 |
| **G: MixUp+CopyPaste** | 0.9510 ± 0.0041 | 0.7336 ± 0.0098 | 0.8998 ± 0.0125 | 0.9327 ± 0.0221 |
| **H: All** | 0.9586 ± 0.0071 | 0.7401 ± 0.0040 | 0.9172 ± 0.0109 | 0.9364 ± 0.0193 |

Configurations E and H reach the highest mAP50-95 at 0.7401, gaining 0.0053 over the baseline. Configuration C posts the lowest mAP50-95 at 0.7336, sitting 0.0012 below the baseline. All configurations fall within a 0.0065 span on mAP50-95 and 0.0098 span on mAP50. Configuration D achieves the lowest standard deviation on mAP50-95 at 0.0020, approximately half of the baseline value. The 95% confidence intervals of most configuration pairs overlap given the n=3 seed budget, so we treat these differences as indicative and preserve the standard deviations throughout the analysis.

**Per-Class AP Analysis**

Figure 4 depicts the change in per-class AP50-95 relative to the baseline for the seven augmented configurations. The dataset exhibits substantial class imbalance, with glove (2463 training instances) and used syringe (1977) at the majority end and scrub (57) and glass (66) at the minority end.

The largest single-class gain in the study comes from scrub under E, at +0.0499 above the baseline. Configuration C achieves +0.0365 on the same class. Both configurations include MixUp. Glass moves in the opposite direction. Every augmented configuration reduces its AP50-95, with C dropping it by 0.0428, E by 0.0301, and D by 0.0271. Tube benefits from every configuration, with gains ranging from +0.0133 under D to

*Jevri Tri Ardiansah

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16531
e-ISSN : 2541-2019
p-ISSN : 2541-044X

+0.0221 under E. Majority classes such as glove, used syringe, and needle exhibit small changes that fall within the noise floor observed across seeds.

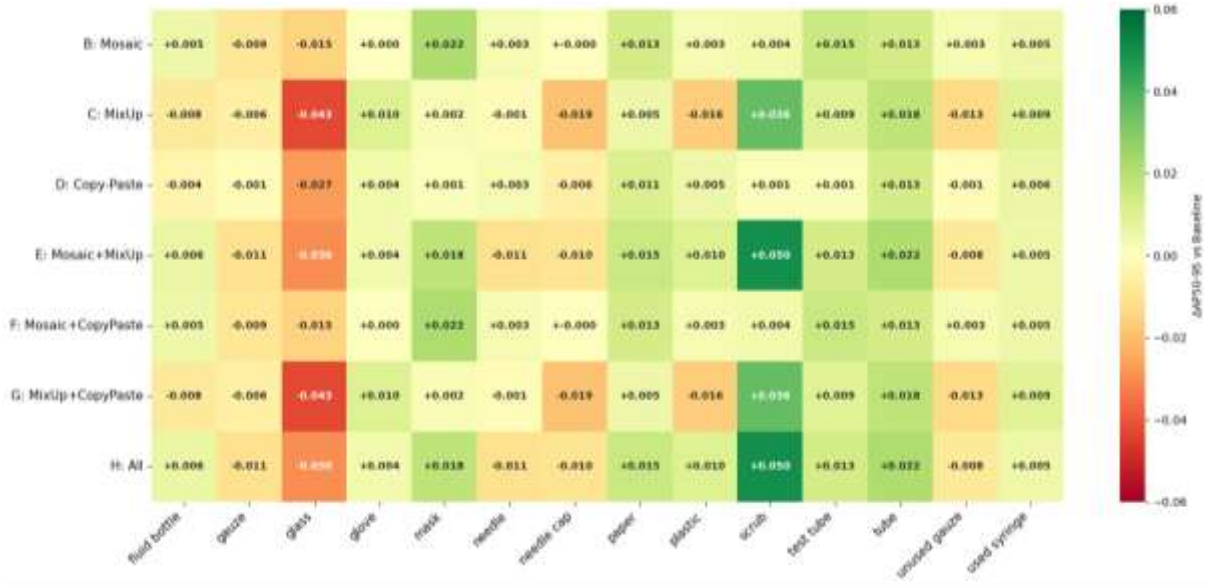


Fig. 4 ΔAP50-95 per class relative to baseline. Green cells mark improvement, red cells mark degradation

The identity between F and B, between G and C, and between H and E is visible in the row-wise pattern of Figure 4. All 14 per-class ΔAP values match to the third decimal place across each pair. The section of Framework Verification documents the verification of this identity at prediction and loss levels.

**Confusion Matrix and Failure Mode Analysis**

Table 4 reports total true positives, false positives, and false negatives derived from the confusion matrix. The confusion matrix reveals that the standard nc × nc slice, which excludes background, hides the dominant failure mode. Figure 5 presents the full confusion matrices for A, C, and E, with the last row showing false positives per class and the last column showing missed detections per class. Configurations F, G, and H are omitted from Table 4 and Figure 5 because they are numerically identical to B, C, and E respectively, as established in the preceding analysis and verified in detail under Framework Verification below. Figure 5 further narrows to three representative configurations (baseline, MixUp, and Mosaic+MixUp) for visual clarity; B and D follow patterns already captured numerically in Table 4.

Table 4 Detection outcomes from the corrected confusion matrix (test set, seed 42)

| Configuration | TP | FP | FN | Precision | Recall |
|---|---|---|---|---|---|
| **A: Baseline** | 904 | 40 | 106 | 0.9576 | 0.8950 |
| **B: Mosaic** | 907 | 36 | 106 | 0.9618 | 0.8954 |
| **C: MixUp** | **908** | **36** | **93** | **0.9619** | **0.9071** |
| **D: Copy-Paste** | 908 | 36 | 103 | 0.9619 | 0.8981 |
| **E: Mosaic+MixUp** | **911** | **33** | **108** | **0.9650** | **0.8940** |

False negatives outnumber false positives across every configuration by roughly three to one. The bidirectional confusion rates for six visually similar class pairs, gauze and unused gauze, needle and needle cap, needle and used syringe, needle cap and used syringe, tube and test tube, plastic and fluid bottle, all measure 0.0000 across every configuration and every seed. The dominant failure mode is not misclassification between similar classes but rather missed detection against the background, concentrated in paper (18 to 25 percent FN), tube (30 to 40 percent FN), and scrub (0 to 33 percent FN across configurations). Recall reaches its highest value under C (0.9071) and precision reaches its highest value under E (0.9650).

*Jevri Tri Ardiansah

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16531
e-ISSN : 2541-2019
p-ISSN : 2541-044X

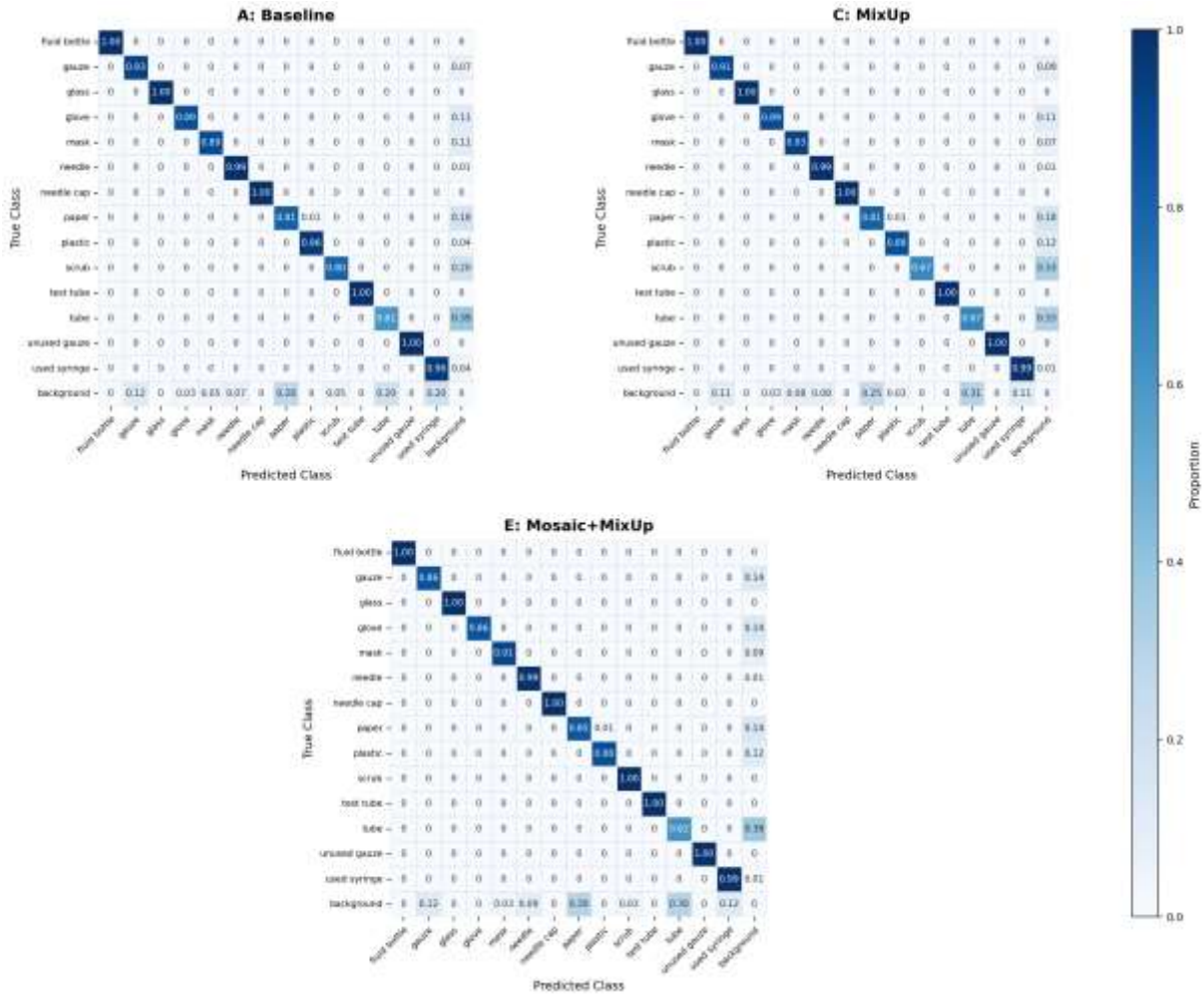


Fig. 5 Confusion matrices for baseline, MixUp, and Mosaic+MixUp. The final row aggregates FP contributions and the final column aggregates missed detections per class

**Framework Verification: Copy-Paste Behavior on Detection Tasks**

Configurations F, G, and H produce numerically identical aggregate metrics to configurations B, C, and E respectively. The identity holds for mAP50, mAP50-95, precision, recall, and per-class AP to the fourth decimal place across every seed. We verified this observation at three independent levels. Table 5 reports prediction-level and loss-level identity checks.

Table 5 Prediction-level and loss-level identity verification

| Pair | Prediction Exact Match | Max Δ box_loss | Max Δ cls_loss | Max Δ dfl_loss |
|---|---|---|---|---|
| **B vs F** | 20/20 (100%) | 0.00e+00 | 0.00e+00 | 0.00e+00 |
| **C vs G** | 20/20 (100%) | 0.00e+00 | 0.00e+00 | 0.00e+00 |
| **E vs H** | 20/20 (100%) | 0.00e+00 | 0.00e+00 | 0.00e+00 |

Prediction outputs for 20 test images match exactly at floating-point precision between each configuration pair. Training loss trajectories match at every epoch for all three loss components. We instrumented the CopyPaste class from ultralytics.data.augment with a call counter and ran a five-epoch diagnostic training with copy_paste=0.1 and copy_paste_mode='flip', matching the exact settings used for configuration D. The counter registered zero invocations.

One asymmetry in the data merits attention. Configuration D differs from the baseline A by 0.0005 on mAP50-95 and 0.0033 on mAP50, while F equals B, G equals C, and H equals E to full precision. If the Copy-Paste transform were entirely inert, D should also equal A. We attribute the small A-D residual to sources of non-determinism unrelated to the transform itself: the AutoBatch procedure selected batch sizes of 20 or 21 across runs depending on GPU memory state, and the parameter parsing step consumed a different amount of random state

*Jevri Tri Ardiansah

e-ISSN : 2541-2019
p-ISSN : 2541-044X

when *copy_paste=0.1* was declared versus when it was absent. The residual falls well within the standard deviation observed across seeds for both A (0.0038) and D (0.0020), so we treat it as stochastic noise rather than as an effect of Copy-Paste on gradient computation. The exact identity of B versus F, C versus G, and E versus H remains the stronger evidence.

## DISCUSSION

### Incremental Effects on a Pre-Augmented Dataset

Our observed effect sizes cluster within a narrow 0.0065 range on mAP50-95. This is consistent with the observation that our training set already carries offline augmentation applied by the Roboflow platform, including flip, rotation, brightness, saturation, blur, and noise transforms. The framing of our study centers on incremental effect rather than augmentation effect in isolation. Prior work on waste detection with YOLO variants reports mAP in a broadly similar operating range. Lin (Lin, 2021) reached 78.04 percent mAP with 100 epochs on a seven-class waste dataset. Sayem et al. (Sayem et al., 2025) reached mAP50 of 63 percent using GELAN-E on 28 recyclable waste categories. Zhou et al. (Zhou et al., 2022) achieved 97.2 percent accuracy for medical waste classification with ResNeXt on eight categories. Comprehensive reviews of YOLO for medical detection (Ragab et al., 2024; Rašić et al., 2023) and deep learning-based waste sorting (Abdu & Mohd Noor, 2022; Hossen et al., 2024) confirm that gains of a few points on mAP are typical when augmentation is added to already-strong baselines, so our narrow range is neither anomalous nor evidence of poor augmentation choice.

### Class Imbalance and the Structure of Per-Class Gains

Figure 4 shows that scrub, the class with the fewest training instances, receives the largest positive push from configurations that include MixUp. Crasto (Crasto, 2024) reported that Mosaic and MixUp mitigate foreground-foreground class imbalance in single-stage detectors by injecting variability into rare-class samples, a pattern also observed in medical detection contexts (Hassan, Hamad, & Mahar, 2022). Our scrub result at +0.0499 under E and +0.0365 under C aligns with that mechanism. The same argument does not carry over to glass, another rare class (66 instances) that consistently loses AP under every augmented configuration. Glass items in the dataset are predominantly transparent or translucent, so pixel-level operations such as MixUp interpolation may obscure the visual cues that separate glass from background. The two-instance validation set for glass also amplifies noise, a limitation that surveys of small-object detection routinely flag (Nikouei et al., 2025). Tube, the class with the smallest median bounding box area (0.0435), benefits from every configuration. Gao et al. (Gao, Gao, & Wei, 2024) and Wu et al. (Wu & Nie, 2025) reported that combining MixUp and Mosaic improved AP for small targets on VisDrone2021 and Tsinghua Tencent100K, and our tube result reproduces that pattern in the biomedical waste domain.

### Class-versus-Background Confusion as the Dominant Failure Mode

The confusion matrix analysis in the Confusion Matrix and Failure Mode Analysis section reveals that missed detections outnumber false positives by roughly three to one. The initial hypothesis that MixUp would amplify inter-class confusion between visually similar pairs, derived from the pixel-interpolation mechanism proposed by Zhang et al. (H. Zhang, Cisse, Dauphin, & Lopez-Paz, 2018) and adopted for object detection by Bochkovskiy et al. (Bochkovskiy, Wang, & Liao, 2020), is not supported by our data. Bidirectional confusion rates for six visually similar pairs measure zero across every configuration. The classifier discriminates these classes reliably when a detection is produced. The failure occurs earlier in the pipeline, at the class-versus-background boundary. This pattern carries direct clinical implications. Pereira et al. (Pereira & Cordeiro, 2026) showed that in medical image classification, false negatives typically carry substantially higher cost than false positives because missed disease delays treatment. Mohammadi-Seif et al. (Mohammadi-Seif & Baeza-Yates, 2026) developed risk-calibrated learning specifically to suppress critical false negatives in medical AI. Both arguments apply to biomedical waste sorting, where missed detections can allow contaminated items to bypass proper handling.

### Precision-Recall Trade-off and Deployment Considerations

The precision and recall values in Table 3 sit on opposite optima depending on augmentation choice. MixUp alone raises recall to 0.9071 and reduces false negatives from 106 to 93. Mosaic combined with MixUp raises precision to 0.9650 and reduces false positives from 40 to 33. Practitioners must choose according to deployment cost structure rather than default to the highest-mAP configuration. Similar trade-off patterns appear in imbalanced clinical detection, where diffusion-based augmentation was shown to prioritize sensitivity for safety-critical classes (Tabibian, Razmpour, & Saha, 2026). The design of a biomedical waste sorting system therefore involves an operational decision about acceptable false positive versus false negative rates, which our findings can inform but cannot fully answer without domain-specific cost weighting.

*Jevri Tri Ardiansah

e-ISSN : 2541-2019
p-ISSN : 2541-044X

**Nuanced Pattern of the Copy-Paste Contribution**

The pairwise identity of B and F, C and G, and E and H holds to full floating-point precision across all seeds and all classes. The diagnostic run confirmed zero invocations of the CopyPaste transform under detection-only settings. The Ultralytics documentation states that the copy_paste parameter only takes effect for segmentation tasks ("Data Augmentation Using Ultralytics YOLO," 2025). These three lines of evidence converge on a single interpretation: the Copy-Paste transform does not enter the augmentation pipeline for detection-only datasets, even when the parameter is declared. The small A-versus-D residual on aggregate metrics reflects run-to-run variability from AutoBatch selection and random state consumption during parameter parsing. Related work on floating-point non-associativity in GPU training has shown that such small residuals can arise from purely hardware-level non-determinism (Shanmugavelu et al., 2025). Recording the parameter in args.yaml without disclosing that it is inert on the training task represents a silent failure mode, a category that Tambon et al. (Tambon, Nikanjam, An, Khomh, & Antoniol, 2023) identified as particularly dangerous in deep learning frameworks because it produces no error message and no visible misbehavior. Kapoor and Narayanan (Kapoor & Narayanan, 2023) documented that such reproducibility hazards propagate across scientific literature when researchers do not verify framework behavior. Our verification protocol, combining byte-level loss identity checks and instrumented tracing, offers a lightweight template that other researchers can adopt before publishing augmentation results built on Ultralytics or similar frameworks.

**Practical Guidelines from the Combined Evidence**

Bringing together the four results sections and the discussion above, we can draw three practical guidelines for training YOLOv11n on biomedical waste datasets that already carry offline augmentation. When missed detection carries the higher deployment cost, MixUp alone offers the highest recall and the largest reduction in false negatives. When false alarms carry the higher cost, Mosaic combined with MixUp offers the highest precision. When training stability across retraining runs is the primary concern, the low variance of Copy-Paste alone reflects the fact that no augmentation transformation actually runs, so it should be understood as an artifact rather than as a strategy. The 43-fold class imbalance in the dataset amplifies per-class variance for rare classes, a pattern that recent work on waste detection under imbalance has also emphasized (Fotovvatikhah, Ahmedy, Noor, & Munir, 2025; Madhavi et al., 2025).

## CONCLUSION

We examined the incremental effect of three online augmentation strategies, Mosaic, MixUp, and Copy-Paste, on YOLOv11n for biomedical waste detection using a 2×2×2 factorial ablation with three random seeds per configuration. The study contributes on three fronts that share one method: each decomposes an aggregate number that looked sufficient on its own into evidence fine enough to change the practical reading of that number. Our per-class analysis identifies class-versus-background confusion as the dominant failure mode of YOLOv11n on this dataset, with missed detections outnumbering false positives by roughly three to one across every configuration. The visually similar class pairs we initially expected MixUp to affect show zero bidirectional confusion which redirects the augmentation design target toward class-versus-background boundaries rather than toward inter-class variability. Beyond the aggregate mAP view, we document a precision-recall trade-off where MixUp alone maximizes recall (0.9071) and Mosaic combined with MixUp maximizes precision (0.9650), so practitioners can align augmentation choice with the safety asymmetry of their deployment. We also verify and document a silent limitation in Ultralytics YOLOv11 whereby the copy_paste parameter has no effect on detection-only datasets, confirmed through byte-level loss trajectory identity, prediction-level output identity, and instrumented tracing of the transform. The verification protocol we propose adds a lightweight reproducibility safeguard for future augmentation studies.

The n=3 seed budget provides enough statistical power for standard deviations and rough confidence intervals but not for formal hypothesis testing on the small effect sizes we observed. The pre-augmented Roboflow dataset limits the counterfactual we can construct, since we cannot isolate the incremental effect of online augmentation from the compounding effect of offline augmentation already baked into the training images. Our verification of Copy-Paste behavior applies specifically to Ultralytics YOLOv11 with copy_paste_mode='flip' on datasets without polygon masks. The mask-based Copy-Paste method of Ghiasi et al. (Ghiasi et al., 2021) on segmentation-annotated datasets may behave differently. The test set contains only two and three instances for glass and scrub respectively, so per-class AP for these classes is sensitive to individual predictions and should be read as indicative.

Extending the seed count from three to at least ten would enable formal significance testing on the precision-recall trade-off. Replicating the study on a dataset that carries polygon mask annotations would test whether mask-based Copy-Paste behaves differently from the silent no-op we documented and would separate framework limitation from technique limitation. The class-versus-background failure mode we identified motivates augmentation strategies that specifically target the class-background boundary rather than inter-class variability,

*Jevri Tri Ardiansah

e-ISSN : 2541-2019
p-ISSN : 2541-044X

potentially through methods that synthesize background-adjacent hard negatives or that reweight the class-versus-background objective during training.

## REFERENCES

Abdu, H., & Mohd Noor, M. H. (2022). A Survey on Waste Detection and Classification Using Deep Learning. *IEEE Access*, *10*, 128151–128165. https://doi.org/10.1109/ACCESS.2022.3226682

Aberger, J., Shami, S., Häcker, B., Pestana, J., Khodier, K., & Sarc, R. (2025). Prototype of AI-powered assistance system for digitalisation of manual waste sorting. *Waste Management*, *194*, 366–378. https://doi.org/10.1016/j.wasman.2025.01.027

Ali, M. L., & Zhang, Z. (2024). The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. *Computers*, *13*(12), 336. https://doi.org/10.3390/computers13120336

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020, April 23). *YOLOv4: Optimal Speed and Accuracy of Object Detection*. arXiv. https://doi.org/10.48550/arXiv.2004.10934

Bruno, A., Caudai, C., Leone, G. R., Martinelli, M., Moroni, D., & Crotti, F. (2023). Medical Waste Sorting: A Computer Vision Approach For Assisted Primary Sorting. *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 1–5. https://doi.org/10.1109/ICASSPW59220.2023.10193520

Carnero, M. C. (2020). Waste Segregation FMEA Model Integrating Intuitionistic Fuzzy Set and the PAPRIKA Method. *Mathematics*, *8*(8), 1375. https://doi.org/10.3390/math8081375

Cirstea, I., Radu, A.-F., Tit, D. M., Radu, A., Bungau, G., & Negru, P. A. (2025). Bibliometric Analysis of Medical Waste Research Using Python-Driven Algorithm. *Algorithms*, *18*(6), 312. https://doi.org/10.3390/a18060312

Crasto, N. (2024, March 11). *Class Imbalance in Object Detection: An Experimental Diagnosis and Study of Mitigation Strategies*. arXiv. https://doi.org/10.48550/arXiv.2403.07113

Data Augmentation using Ultralytics YOLO. (2025, April 14). Retrieved July 4, 2026, from Ultralytics Docs website: https://docs.ultralytics.com/guides/yolo-data-augmentation

Fang, B., Yu, J., Chen, Z., Osman, A. I., Farghali, M., Ihara, I., … Yap, P.-S. (2023). Artificial intelligence for waste management in smart cities: A review. *Environmental Chemistry Letters*, *21*(4), 1959–1989. https://doi.org/10.1007/s10311-023-01604-3

Fostiropoulos, I., & Itti, L. (2023). ABLATOR: Robust Horizontal-Scaling of Machine Learning Ablation Experiments. *Proceedings of the Second International Conference on Automated Machine Learning*, 19/1-15. PMLR. Retrieved from https://proceedings.mlr.press/v224/fostiropoulos23a.html

Fotovvatikhah, F., Ahmedy, I., Noor, R. M., & Munir, M. U. (2025). A Systematic Review of AI-Based Techniques for Automated Waste Classification. *Sensors*, *25*(10), 3181. https://doi.org/10.3390/s25103181

Gao, S., Gao, M., & Wei, Z. (2024). MCF-YOLOv5: A Small Target Detection Algorithm Based on Multi-Scale Feature Fusion Improved YOLOv5. *Information*, *15*(5), 285. https://doi.org/10.3390/info15050285

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., … Zoph, B. (2021). Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2917–2927. Nashville, TN, USA: IEEE. https://doi.org/10.1109/CVPR46437.2021.00294

Gundersen, O. E., Shamsaliei, S., Kjærnli, H. S., & Langseth, H. (2023). On Reporting Robust and Trustworthy Conclusions from Model Comparison Studies Involving Neural Networks and Randomness. *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*, 37–61. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3589806.3600044

Hassan, N. M., Hamad, S., & Mahar, K. (2022). A Deep Learning Model for Mammography Mass Detection Using Mosaic and Reconstructed Multichannel Images. *Computational Science and Its Applications – ICCSA 2022: 22nd International Conference, Malaga, Spain, July 4–7, 2022, Proceedings, Part I*, 544–559. Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-031-10522-7_37

Hossen, M. M., Ashraf, A., Hasan, M., Majid, M. E., Nashbat, M., Kashem, S. B. A., … Chowdhury, M. E. H. (2024). GCDN-Net: Garbage classifier deep neural network for recyclable urban waste management. *Waste Management*, *174*, 439–450. https://doi.org/10.1016/j.wasman.2023.12.014

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Khanam, R., & Hussain, M. (2024, October 23). *YOLOv11: An Overview of the Key Architectural Enhancements*. arXiv. https://doi.org/10.48550/arXiv.2410.17725

*Jevri Tri Ardiansah

Kumar, A. K., Ali, Y., Kumar, R. R., Assaf, M. H., & Ilyas, S. (2025). Artificial Intelligent and Internet of Things framework for sustainable hazardous waste management in hospitals. *Waste Management*, *203*, 114816. https://doi.org/10.1016/j.wasman.2025.114816

Kunwar, S., & Rai, P. (2025, October 15). *Health Care Waste Classification Using Deep Learning Aligned with Nepal's Bin Color Guidelines*. arXiv. https://doi.org/10.48550/arXiv.2508.07450

Lin, W. (2021). YOLO-Green: A Real-Time Classification and Object Detection Model Optimized for Waste Management. *2021 IEEE International Conference on Big Data (Big Data)*, 51–57. https://doi.org/10.1109/BigData52589.2021.9671821

Lu, W., & Chen, J. (2022). Computer vision for solid waste sorting: A critical review of academic research. *Waste Management*, *142*, 29–43. https://doi.org/10.1016/j.wasman.2022.02.009

Madhavi, B., Mahanty, M., Lin, C.-C., Lakshmi Jagan, B. O., Rai, H. M., Agarwal, S., & Agarwal, N. (2025). SwinConvNeXt: A fused deep learning architecture for Real-time garbage image classification. *Scientific Reports*, *15*(1), 7995. https://doi.org/10.1038/s41598-025-91302-7

Menekşe, A., & Camgöz Akdağ, H. (2023). Medical waste disposal planning for healthcare units using spherical fuzzy CRITIC-WASPAS. *Applied Soft Computing*, *144*, 110480. https://doi.org/10.1016/j.asoc.2023.110480

Mohammadi-Seif, A., & Baeza-Yates, R. (2026, April 14). *Risk-Calibrated Learning: Minimizing Fatal Errors in Medical AI*. arXiv. https://doi.org/10.48550/arXiv.2604.12693

Nahiduzzaman, Md., Ahamed, Md. F., Naznine, M., Karim, Md. J., Kibria, H. B., Ayari, M. A., … Haider, J. (2025). An automated waste classification system using deep learning techniques: Toward efficient waste recycling and environmental sustainability. *Knowledge-Based Systems*, *310*, 113028. https://doi.org/10.1016/j.knosys.2025.113028

Nasir, N., Hossain, M. S., Islam, Md. S., Islam, Md. A., & Islam, M. M. (2024). Transforming Medical Waste Management Through IoT and Machine Learning: A Path Towards Sustainability. In K. Arai (Ed.), *Intelligent Computing* (pp. 556–575). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-62277-9_36

Nikouei, M., Baroutian, B., Nabavi, S., Taraghi, F., Aghaei, A., Sajedi, A., & Moghaddam, M. E. (2025). Small object detection: A comprehensive survey on challenges, techniques and real-world applications. *Intelligent Systems with Applications*, *27*, 200561. https://doi.org/10.1016/j.iswa.2025.200561

Padilla, R., Netto, S. L., & da Silva, E. A. B. (2020). A Survey on Performance Metrics for Object-Detection Algorithms. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130

Pereira, M. R. S., & Cordeiro, F. R. (2026, April 26). *Risk-Aware Robust Learning: Reducing Clinical Risk under Label Noise in Medical Image Classification*. arXiv. https://doi.org/10.48550/arXiv.2604.23875

Picard, D. (2023, May 11). *Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision*. arXiv. https://doi.org/10.48550/arXiv.2109.08203

QuratulAin. (2025, May). Biomedical waste detection dataset [Open Source Dataset]. Retrieved from https://universe.roboflow.com/quratulain/biomedical-waste-detection-nv7fq

Ragab, M. G., Abdulkadir, S. J., Muneer, A., Alqushaibi, A., Sumiea, E. H., Qureshi, R., … Alhussian, H. (2024). A Comprehensive Systematic Review of YOLO for Medical Object Detection (2018 to 2023). *IEEE Access*, *12*, 57815–57836. https://doi.org/10.1109/ACCESS.2024.3386826

Rana, A., & Vaidya, P. (2026). YOLO-based deep learning framework for real-time multi-class plant health monitoring in precision agriculture. *Scientific Reports*, *16*(1), 197. https://doi.org/10.1038/s41598-025-29132-w

Rašić, M., Tropčić, M., Karlović, P., Gabrić, D., Subašić, M., & Knežević, P. (2023). Detection and Segmentation of Radiolucent Lesions in the Lower Jaw on Panoramic Radiographs Using Deep Neural Networks. *Medicina*, *59*(12), 2138. https://doi.org/10.3390/medicina59122138

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 658–666. Long Beach, CA, USA: IEEE. https://doi.org/10.1109/CVPR.2019.00075

Sayem, F. R., Islam, Md. S. B., Naznine, M., Nashbat, M., Hasan-Zia, M., Kunju, A. K. A., … Chowdhury, M. E. H. (2025). Enhancing waste sorting and recycling efficiency: Robust deep learning-based approach for classification and detection. *Neural Computing and Applications*, *37*(6), 4567–4583. https://doi.org/10.1007/s00521-024-10855-2

Shanmugavelu, S., Taillefumier, M., Culver, C., Hernandez, O., Coletti, M., & Sedova, A. (2025). Impacts of floating-point non-associativity on reproducibility for HPC and deep learning applications. *Proceedings of the SC '24 Workshops of the International Conference on High Performance Computing, Network,*

*Jevri Tri Ardiansah

*Storage, and Analysis*, 170–179. Atlanta, GA, USA: IEEE Press. https://doi.org/10.1109/SCW63240.2024.00028

Sheikholeslami, S., Meister, M., Wang, T., Payberah, A. H., Vlassov, V., & Dowling, J. (2021). AutoAblation: Automated Parallel Ablation Studies for Deep Learning. *Proceedings of the 1st Workshop on Machine Learning and Systems*, 55–61. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3437984.3458834

Tabibian, M., Razmpour, T., & Saha, R. (2026). Diffusion models vs. DCGANs for class-imbalanced lung cancer CT classification: A comparative study. *Intelligence-Based Medicine*, *13*, 100336. https://doi.org/10.1016/j.ibmed.2025.100336

Tambon, F., Nikanjam, A., An, L., Khomh, F., & Antoniol, G. (2023). Silent bugs in deep learning frameworks: An empirical study of Keras and TensorFlow. *Empirical Software Engineering*, *29*(1), 10. https://doi.org/10.1007/s10664-023-10389-6

Wang, J., Lutellier, T., Qian, S., Pham, H. V., & Tan, L. (2022). EAGLE: Creating equivalent graphs to test deep learning libraries. *Proceedings of the 44th International Conference on Software Engineering*, 798–810. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3510003.3510165

Wei, A., Deng, Y., Yang, C., & Zhang, L. (2022). Free lunch for testing: Fuzzing deep-learning libraries from open source. *Proceedings of the 44th International Conference on Software Engineering*, 995–1007. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3510003.3510041

Wu, Q., & Nie, X. (2025). Improved YOLOv10: A Real-Time Object Detection Approach in Complex Environments. *Sensors*, *25*(22), 6893. https://doi.org/10.3390/s25226893

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., & Shen, F. (2023, November 5). *Image Data Augmentation for Deep Learning: A Survey*. arXiv. https://doi.org/10.48550/arXiv.2204.08610

Yazdani, M., Tavana, M., Pamučar, D., & Chatterjee, P. (2020). A rough based multi-criteria evaluation method for healthcare waste disposal location decisions. *Computers & Industrial Engineering*, *143*, 106394. https://doi.org/10.1016/j.cie.2020.106394

Yu, B., Li, Z., Cao, Y., Wu, C., Qi, J., & Wu, L. (2024). YOLO-MPAM: Efficient real-time neural networks based on multi-channel feature fusion. *Expert Systems with Applications*, *252*, 124282. https://doi.org/10.1016/j.eswa.2024.124282

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018, February 15). *mixup: Beyond Empirical Risk Minimization*. Presented at the International Conference on Learning Representations. Retrieved from https://openreview.net/forum?id=r1Ddp1-Rb

Zhang, X., Jiang, W., Shen, C., Li, Q., Wang, Q., Lin, C., & Guan, X. (2025). Deep Learning Library Testing: Definition, Methods and Challenges. *ACM Computing Surveys*, *57*(7), 187:1-187:37. https://doi.org/10.1145/3716497

Zhou, H., Yu, X., Alhaskawi, A., Dong, Y., Wang, Z., Jin, Q., … Lu, H. (2022). A deep learning approach for medical waste classification. *Scientific Reports*, *12*(1), 2159. https://doi.org/10.1038/s41598-022-06146-2

*Jevri Tri Ardiansah