Applied Research in Science and Technology 6(1): 100–121 2026

Contents lists available at openscie.com
E-ISSN: 2776-7205
Applied Research in Science and Technology
DOI: 10.33292/areste.v6i1.154
Journal homepage: https://areste.org/index.php/oai

MedPackID: An Indonesian Benchmark Dataset for Medicine Packaging and Pill Detection Using YOLOv8n and Faster R-CNN

Qory Hidayati[1,2]*, Hendra Kusuma [1], Muhammad Attamimi[1]

[1] Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology (F-ELECTICS), Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2] Department of Electrical Engineering, Balikpapan State Polytechnic, Balikpapan, Indonesia

*Correspondence: E-mail: qory.hidayati@poltekba.ac.id

ARTICLE INFO

Article History:
Received 13 August 2026
Revised 2 October 2026
Accepted 24 October 2026
Published 27 October 2026

Keywords:
Medicine,
Pharmaceutical image dataset,
Packaging,
Pills,
Visual retrieval.

ABSTRACT

Background: Reliable image-based medicine recognition is necessary to support safer medication handling, reduce identification errors, and improve automated access to pharmaceutical information. However, its performance may degrade when medicine packaging, pills, and printed text vary across acquisition conditions.
Aims: This study aimed to develop MedPackID as a public Indonesian benchmark for medicine packaging and pill detection and to establish reproducible detection and conditional recognition baselines.
Methods: MedPackID contains 1,444 consumer-device images and 1,540 manually reviewed bounding boxes collected at one site using one smartphone under varied illumination, viewpoint, orientation, scale, background clutter, and partial occlusion. YOLOv8n and Faster R-CNN were evaluated for object detection. Detected packaging regions were processed using EasyOCR, text normalization, and fuzzy matching, whereas detected pill regions were evaluated using ResNet-18 embeddings and cosine-similarity retrieval.
Results: On the frozen test partition, YOLOv8n achieved mAP@0.5 of 0.968 and mAP@0.5:0.95 of 0.707, compared with 0.909 and 0.645 for Faster R-CNN. Packaging retrieval achieved 90.98% Top-1 accuracy, while closed-set pill retrieval achieved 85.4% Top-1 accuracy. MedPackID provides a reproducible regional benchmark for pharmaceutical object detection and conditional recognition. The results establish within-dataset reference performance rather than clinical validity or generalization to unseen medicine identities, devices, or acquisition sites.

To cite this article: Hidayati, Q., Kusuma, H., Attamimi, M. (2026). MedPackID: An Indonesian Benchmark Dataset for Medicine Packaging and Pill Detection Using YOLOv8n and Faster R-CNN. *Applied Research in Science and Technology*, 6(1), 100–121.

This article is under a Creative Commons Attribution-ShareAlike 4.0 International (**CC BY-SA 4.0**) License. Creative Commons Attribution-ShareAlike 4.0 International License Copyright ©2026 by author/s

1. Introduction

Medication-related harm remains a major patient-safety concern. The World Health Organization identifies harm associated with medicines and therapeutic options as an important component of preventable harm in medical care (World Health Organization, 2024). Reliable image-based medicine recognition can support safer medication handling and automated access to pharmaceutical information by localizing pharmaceutical objects and providing machine-readable cues for subsequent identification. If medicine packaging or pills cannot be localized and distinguished reliably, downstream text

recognition or visual matching may fail before an identity decision is attempted, limiting the reliability of automated pharmaceutical-information systems.

Existing pharmaceutical computer-vision methods include object detection, scene-text recognition, and visual retrieval. Faster R-CNN represents a two-stage detection paradigm in which a Region Proposal Network generates candidate regions before classification and bounding-box refinement (Ren *et al.,* 2015), while Feature Pyramid Networks improve multi-scale feature representation (Lin *et al.,* 2017). In contrast, YOLO represents a one-stage detection paradigm that performs localization and classification within a unified inference process (Redmon *et al.,* 2016). These approaches are relevant to pharmaceutical images because a single scene may contain relatively large medicine packages together with considerably smaller pills. However, pharmaceutical object detection remains challenging because reflective materials, dense packaging graphics, partial occlusion, small object size, and visual similarity among pills can reduce localization reliability (My *et al.,* 2025). Tan *et al.* (2021) also demonstrated that pharmaceutical detector selection involves a trade-off between detection accuracy and inference speed, supporting the need to evaluate contrasting detection paradigms under a common protocol.

Public pharmaceutical image datasets have primarily emphasized pill-level matching or classification. The NLM Pill Image Recognition Challenge evaluated consumer-query pill images against controlled reference images (Yaniv *et al.,* 2016), while ePillID introduced a low-shot, fine-grained pill-identification benchmark (Usuyama *et al.,* 2020). These datasets provide important resources for pill retrieval research, but they do not jointly represent medicine packaging, individual pills, packaging text, and routine contextual backgrounds within a single acquisition and annotation framework. Packaging-text recognition studies have also demonstrated the value of extracting printed pharmaceutical information (Yue *et al.,* 2021). However, optical character recognition remains sensitive to small fonts, glare, perspective distortion, multilingual layouts, low contrast, and visually crowded packaging designs (Baek, 2019; Shi *et al.*, 2017). As a result, existing datasets and recognition pipelines often address object localization, packaging-text recognition, and pill identification as separate tasks rather than as parts of a unified pharmaceutical-recognition workflow.

These limitations motivate the development of MedPackID, a unified pharmaceutical computer-vision benchmark that combines medicine-packaging and pill detection within a common annotation framework and links each detected object class to an appropriate downstream recognition pathway. Rather than introducing a new object-detection architecture, MedPackID provides a common experimental framework in which established detection models can be evaluated together with packaging-text and pill-retrieval baselines. YOLOv8n and Faster R-CNN were selected as complementary detection baselines because they represent established one-stage and two-stage strategies, respectively. YOLOv8n provides an efficient end-to-end detector implemented in the Ultralytics framework (Ultralytics, 2023), whereas Faster R-CNN with a ResNet-50 FPN backbone provides a region-proposal-based comparison with explicit multi-scale feature processing (Ren, 2015; Lin *et al.*, 2017). The purpose of this comparison is not to claim general architectural superiority, but to characterize how these two established paradigms perform on the same pharmaceutical images, annotations, dataset partitions, and evaluation criteria.

This study introduces MedPackID, a publicly available Indonesian dataset containing 1,444 consumer-device images and 1,540 manually reviewed bounding boxes for medicine packaging and pills (Hidayati *et al.,* 2026). The novelty of MedPackID lies not in a new detection architecture, but in the design of a region-specific benchmark that jointly represents medicine packaging and individual pills, provides a common localization task for both object types, and supports separate OCR-based and embedding-based recognition pathways within the same evaluation framework. This differs from established pill-oriented benchmarks such as the NLM Pill Image Recognition Challenge (Yaniv *et al.,* 2016) and ePillID (Usuyama *et al*., 2020), which primarily focus on pill-level matching or identification rather than joint packaging-and-pill localization followed by class-dependent recognition. MedPackID

also preserves routine visual variation in illumination, viewpoint, orientation, scale, background clutter, and partial occlusion.

The objectives of this study were to: (1) construct and document a reproducible Indonesian benchmark for medicine packaging and pill detection; (2) establish common-protocol object-detection baselines using YOLOv8n and Faster R-CNN; (3) evaluate OCR-assisted packaging retrieval and closed-set visual pill retrieval as class-dependent downstream baselines; and (4) identify technical limitations that should guide future cross-device, identity-disjoint, and open-set evaluation. MedPackID is intended for research, education, and benchmarking and is not presented as a clinically validated medication-identification system.

2. Materials and Methods

MedPackID was developed to support research on the detection and downstream recognition of medicine packaging and pills under varied single-site consumer-device acquisition conditions. Unlike datasets that primarily contain isolated pills or images captured against controlled backgrounds, MedPackID includes pharmaceutical objects photographed in routine environments, with variations in illumination, viewpoint, scale, orientation, background clutter, partial occlusion, and printed packaging content.

The dataset focuses on pharmaceutical products available in Indonesia. This context is relevant because medicine packaging may contain heterogeneous typography, multilingual text, reflective surfaces, dense visual layouts, and different container forms. These characteristics create practical challenges for object detection, optical character recognition, and visual pill retrieval. MedPackID therefore provides a unified dataset in which medicine packaging and individual pills may appear separately or together within the same image. The dataset is intended to support object localization and conditional downstream processing, in which detected packaging regions can be processed through OCR, whereas detected pill regions can be processed through visual feature retrieval.

2.1 Study Setting, Materials, and Image Acquisition

The MedPackID images were collected from 1 to 24 December 2025 at one acquisition location, namely Medica ITS Clinic. Image acquisition was performed using an OPPO Reno12 smartphone to represent practical consumer-device capture conditions. The smartphone camera was used only to acquire still images; object detection was subsequently performed offline on the stored MedPackID images rather than through live-camera inference. Image collection on the clinic premises was conducted with permission from the management of Medica ITS Clinic. The dataset collection involved only medicine packaging and individual pills and did not involve human participants or the collection of patient data. The released images contain no patient-identifiable information, such as patient names, patient identification numbers, or other personal identifiers.

A fixed studio-style acquisition setup was not imposed. Images were captured under variations in illumination, camera position, object orientation, and background appearance. The maximum stored image resolution was 2412 × 1080 pixels. This collection strategy was intended to preserve visual variability commonly encountered when medicine packaging and pills are photographed in routine handling environments.

The resulting images include variations in:

- natural, artificial, and relatively low illumination;
- camera viewpoints, object scales, and camera-to-object distances;
- portrait and landscape image orientations;
- simple and cluttered backgrounds;
- partial occlusion and overlap between pharmaceutical objects; and
- single-object and multi-object scenes.

Each image may contain medicine packaging, individual tablets or capsules, or both object types. The final dataset contains 1,444 images representing 243 medicine identities. Images were retained only when the target pharmaceutical objects were sufficiently visible to support consistent class assignment and bounding-box annotation. Images with severe blur, extreme occlusion, or insufficient object visibility were excluded before annotation.

2.2 Annotation Protocol and Quality Control

All retained images were manually annotated using LabelImg (Tzutalin, 2024) according to a consistent class-definition and bounding-box protocol. Two object classes were defined:

1. Packaging: visible medicine containers or package surfaces, including boxes, blister packs, bottles, and other product containers.
2. Pill: individual tablets or capsules that were visually distinguishable as separate pharmaceutical objects.

Each target object was annotated using a rectangular bounding box enclosing its visible spatial extent. Packaging objects were labeled as a single object when their visible boundaries could be identified consistently. Individual tablets and capsules were labeled separately when they appeared as distinct objects. For blister packs, the blister package was labeled as packaging, whereas tablets or capsules removed from the blister and visible as independent objects were labeled as pill.

Partially occluded objects were annotated when their class and approximate boundaries remained visually identifiable. Objects were excluded when severe occlusion, blur, image cropping, or visual ambiguity prevented reliable class assignment or bounding-box placement. When packaging and pills appeared in the same image, each visible object was annotated independently.

The initial annotations were created by one annotator and subsequently reviewed by one additional reviewer. The review covered all annotated images and examined incorrect class assignments, missing target objects, duplicate bounding boxes, inaccurate object boundaries, and inconsistent treatment of partially occluded objects. Identified errors were corrected through direct re-examination of the original images.

Because the dataset was not independently annotated in full by multiple annotators, inter-annotator agreement and mean inter-annotator bounding-box IoU were not calculated. Annotation quality was instead controlled through a complete secondary review of the initial labels.

The final dataset contains 1,540 bounding-box annotations, comprising 899 packaging instances and 641 pill instances. The annotations are distributed in both YOLO text format and Pascal VOC XML format, enabling their use with common object-detection frameworks, including YOLO and Faster R-CNN.

2.3 Dataset Partitioning, Statistics, and Leakage Audit

MedPackID contains 1,444 images and 1,540 manually annotated object instances. The dataset was divided into predefined training, validation, and test partitions containing 1,010, 216, and 218 images, respectively. The training partition was used for model fitting, the validation partition was used for model monitoring and configuration selection, and the test partition was reserved for final evaluation. The distribution of images is presented in Table 2.

Table 2. Distribution of images across the MedPackID partitions.

Subset	Number of Images	Percentage
Training	1,010	69.94%
Validation	216	14.96%
Test	218	15.10%
Total	1,444	100%

Because individual images may contain more than one pharmaceutical object the number of bounding-box instances differs from the number of images. Table 3 presents the object-level distribution.

The test partition contains 233 annotated objects, comprising 138 packaging instances and 95 pill instances. Across the complete dataset, packaging objects account for 58.38% of all annotations, whereas pill objects account for 41.62%. This naturally observed class distribution was retained without artificial class balancing.

Table 3. Object-instance distribution by class and dataset partition

Subset	Packaging	Pill	Total
Training	632	441	1,073
Validation	129	105	234
Test	138	95	233
Total	899	641	1,540

In addition to the object-instance distribution reported in Table 3, an identity-aware audit was conducted to determine whether medicine identities were shared across the training, validation, and test partitions. Among the 243 medicine identities, 102 identities were exclusive to a single partition, whereas 141 identities occurred in more than one partition. Specifically, 93 identities were shared between the training and validation partitions, 85 between the training and test partitions, and 42 between the validation and test partitions. The audit therefore identified identity-level overlap across the predefined image-level partitions, including between the training and test sets. Consequently, the reported detection results should be interpreted as performance under the current image-level partitioning protocol rather than as an evaluation of generalization to entirely unseen medicine identities.

To distinguish identity overlap from image-level leakage, we additionally conducted a cross-partition duplicate and near-duplicate audit. Exact duplicates were screened using SHA-256 file hashes, while perceptually similar images among shared medicine identities were screened using perceptual hashing and image-embedding similarity, followed by manual verification. No exact cross-partition duplicates were identified. Three image pairs were confirmed as near duplicates across dataset partitions. We therefore repeated the affected evaluation after excluding the corresponding evaluation images. The resulting performance estimates remained comparable to those obtained from the original partition, indicating that the reported results were not materially driven by these near-duplicate cases.

2.4 Dataset Availability

MedPackID is publicly available through Mendeley Data (Hidayati *et al.,* 2026). The current release is identified as Version 1 and is accessible through the following DOI: https://doi.org/10.17632/d23kndp3sv.1

The dataset is released under the Creative Commons Attribution 4.0 International license (CC BY 4.0). This license permits users to copy, redistribute, and adapt the dataset for any purpose, provided that appropriate credit is given to the original authors and source.

The Version 1 repository contains the following files:

- images.zip, containing the MedPackID image collection;
- labels.zip, containing bounding-box annotations in YOLO text format;
- annotations voc.zip, containing bounding-box annotations in Pascal VOC XML format;
- data.yaml, defining the dataset configuration, directory paths, number of classes, and class names used by the YOLO framework;
- labels.txt, providing the human-readable class-label definitions;

- drug candidates.csv, containing structured candidate information used during medicine-name matching;
- drug name dose function.csv, containing medicine names, dosage information, and associated medicine functions used by the downstream recognition process;
- readme.md, providing instructions and general dataset documentation; and
- license.txt, containing the CC BY 4.0 licensing terms.

The machine-readable class configuration defines two object categories: packaging and pill. The released YOLO and Pascal VOC annotations enable the dataset to be used with common object-detection frameworks, including YOLO and Faster R-CNN. The structured CSV files support metadata-assisted medicine matching and are distinct from the manually verified transcription set used to compute the OCR metrics reported in this article. For exact reproduction of CER, WER, and exact-match accuracy, the verified transcription set should accompany the final archival release.

MedPackID is intended for research, education, benchmarking, and the development of pharmaceutical computer-vision methods. The dataset has not been clinically validated and must not be used as the sole basis for medication administration, clinical diagnosis, or unsupervised medication decisions. Medicine names, trademarks, logos, and packaging designs remain the property of their respective owners.

2.5 Detection, Training, and Evaluation Workflow

The research proceeded sequentially from image acquisition and annotation to dataset partitioning, detector training, frozen-test inference, and conditional recognition. Figure 1 summarizes the overall research workflow, whereas Figure 2 provides a more detailed representation of detector training, inference, and class-dependent recognition.

The MedPackID dataset contains 1,444 images representing 243 medicine identities (Hidayati *et al.*, 2026). The dataset includes 1,540 manually reviewed bounding-box annotations, comprising 899 packaging instances and 641 pill instances. The predefined training, validation, and test partitions contain 1,010, 216, and 218 images, respectively. At the object-instance level, the corresponding partitions contain 1,073, 234, and 233 annotations. The frozen test partition contains 138 packaging objects and 95 pill objects.

The training partition was used to estimate model parameters. The validation partition was used to monitor optimization, select the best checkpoint, and configure operational thresholds. The test partition was reserved for final evaluation and was not used for hyperparameter adjustment, checkpoint selection, or threshold optimization. Bounding-box annotations were created and reviewed using LabelImg (Tzutalin, 2024). The annotations were distributed in two formats. YOLO text annotations were used for the YOLOv8n experiments, while Pascal VOC XML annotations were used for Faster R-CNN. Both detectors used the same image partitions, class definitions, and ground-truth objects to maintain a consistent comparison.

Two complementary object-detection paradigms were evaluated. YOLOv8n was implemented using the Ultralytics framework (Ultralytics, 2023) and represents the one-stage YOLO family, in which localization and classification are produced within a unified inference pipeline (Redmon *et al.*, 2016). It was selected as the efficiency-oriented baseline. Faster R-CNN with a ResNet-50 Feature Pyramid Network backbone was selected as the two-stage comparison because it generates candidate regions and subsequently classifies and refines their bounding boxes, while FPN supports multi-scale representation (Ren, 2015; Lin, 2017). This pairing provides a common-dataset comparison between efficient one-stage detection and region-proposal-based two-stage detection rather than an attempt to attribute all performance differences solely to architecture.

Both detectors were trained for 50 epochs. YOLOv8n was initialized using COCO-pretrained weights. Faster R-CNN was implemented using PyTorch and Torchvision (Paszke *et al*, 2019; Torchvision Contributors, 2026) and was initialized using the default pretrained Torchvision detection weights. The original Faster R-CNN classification predictor was replaced with a predictor containing three output classes: background, packaging, and pill. The final experimental settings are summarized in Table 4.

For YOLOv8n, training-time augmentation included HSV variation with hue, saturation, and value gains of 0.015, 0.7, and 0.4, respectively; translation of 0.1; scale variation of 0.5; horizontal flipping with a probability of 0.5; and mosaic augmentation. Mosaic augmentation was disabled during the final 10 epochs. Vertical flipping, rotation, shear, perspective transformation, mixup, and cutmix were not applied

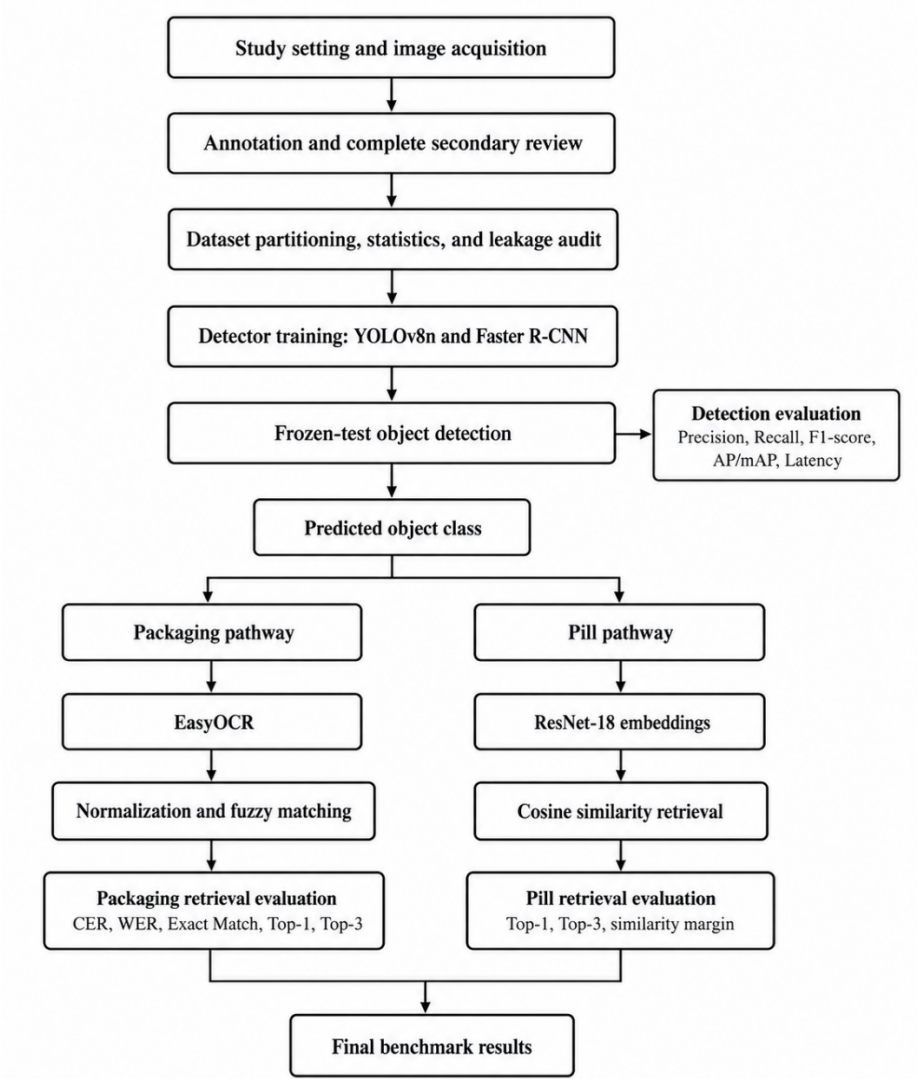


Figure 1. Overall research workflow of MedPackID

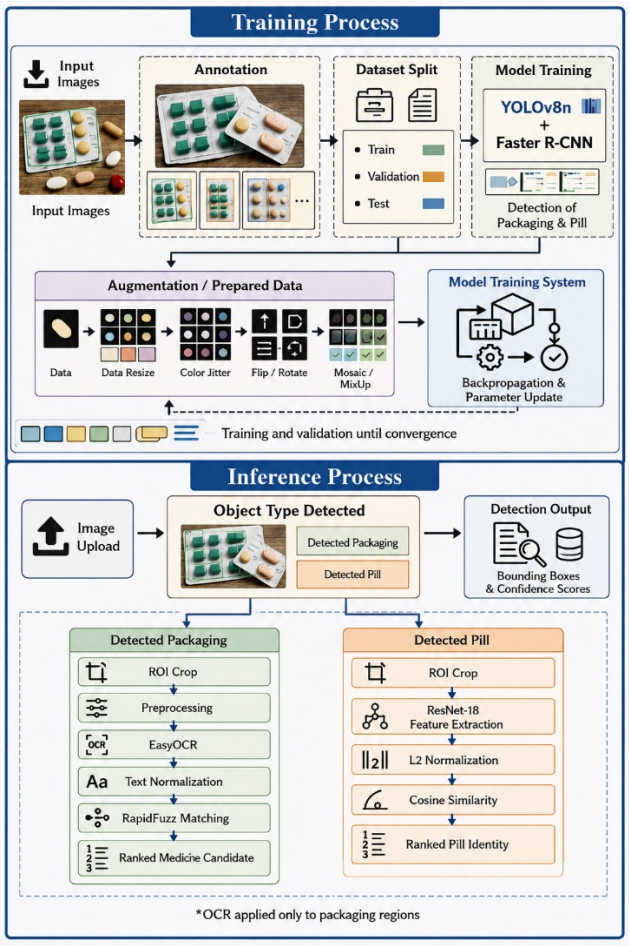


Figure 2. Detector training, inference, and conditional recognition workflow.

Table 4. Training and inference configurations of the evaluated detectors

Configuration	YOLOv8n	Faster R-CNN
Model variant/backbone	YOLOv8n	ResNet-50 FPN
Implementation	Ultralytics	PyTorch/Torchvision
Pretrained weights	COCO-pretrained weights	Torchvision default pretrained weights
Input image size	640 × 640 pixels	Minimum side 640 pixels; maximum side 1,024 pixels
Batch size	16	2
Optimizer	AdamW	AdamW
Initial learning rate	1×10^{-3}	1×10^{-4}
Weight decay	5×10^{-4}	1×10^{-4}
Number of epochs	50	50
Deterministic execution	Enabled	Enabled
Operational confidence threshold	0.25	0.25
NMS IoU threshold	0.70	0.50
Evaluation IoU criteria	0.50 and 0.50:0.95	0.50 and 0.50:0.95
Checkpoint selection	Highest validation mAP@0.5:0.95	Highest validation mAP@0.5:0.95

For Faster R-CNN, training-time augmentation consisted of horizontal flipping with a probability of 0.5 and brightness and contrast factors randomly varied within the interval [0.8, 1.2]. Geometric transformations were applied jointly to the images and their bounding-box coordinates. No augmentation was applied to the validation or test partitions.

For each detector, model performance was calculated on the validation partition after every epoch. The checkpoint with the highest validation mAP@0.5:0.95 was retained. After all configurations had been fixed, the selected checkpoint was evaluated once on the frozen test partition.

Each detector produced a predicted class, bounding-box coordinates, and a confidence score for every candidate object. Object matching and intersection over union followed the standard PASCAL VOC and COCO evaluation conventions (Everingham, 2010; Lin, 2014). For a predicted bounding box Bpred and a corresponding ground-truth bounding box Bgt, IoU was calculated as:

$$\mathrm{IoU} = \frac{|B_{\mathrm{pred}} \cap B_{\mathrm{gt}}|}{|B_{\mathrm{pred}} \cup B_{\mathrm{gt}}|} \tag{1}$$

For threshold-dependent summaries, an operational confidence threshold of 0.25 was applied to both detectors. A prediction was counted as a true positive when its predicted class was correct, its confidence score was at least 0.25, and its IoU with an unmatched ground-truth object was at least 0.50. Each ground-truth object could be matched to at most one prediction. Unmatched predictions were counted as false positives, while unmatched ground-truth objects were counted as false negatives.

Threshold-dependent precision, recall, and F1-score were calculated from true-positive, false-positive, and false-negative counts using standard classification-measure definitions (Sokolova & Lapalme., 2009):

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}, \tag{2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \tag{3}$$

$$\mathrm{F1\text{-}Score} = 2\frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision+Recall}}. \tag{4}$$

Average Precision was calculated from the confidence-ranked precision-recall curve following the COCO evaluation convention, including interpolation at 101 recall levels (Lin *et al.*, 2014). For class and IoU threshold, Average Precision was calculated as:

$$\mathrm{AP}_{i,t} = \frac{1}{101}\sum_{r \in \mathcal{R}} p^*_{i,t}(r), \tag{5}$$

where

$$\mathcal{R} = \{0, 0.01, 0.02, \ldots, 1.00\}, \tag{6}$$

and $p^*_{i,t}(r)$ denotes the maximum precision obtained at a recall level greater than or equal to r.

Mean Average Precision at a fixed IoU threshold t was calculated across the N object classes as:

$$\mathrm{mAP@}t = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_{i,t}, \tag{7}$$

where $N = 2$ for the packaging and pill classes. Overall mAP@0.5:0.95 was calculated by averaging AP across the two object classes and the ten IoU thresholds in:

$$\mathcal{T} = \{0.50, 0.55, 0.60, \ldots, 0.95\}, \tag{8}$$

as

$$\text{mAP@0.5:0.95} = \frac{1}{N|\mathcal{T}|}\sum_{i=1}^{N}\sum_{t\in\mathcal{T}}\text{AP}_{i,t}. \quad (9)$$

Class-specific performance at an IoU threshold of 0.50 was reported as AP@0.5, whereas the average across the packaging and pill classes was reported as mAP@0.5. The mAP@0.5:0.95 metric additionally evaluates localization under progressively stricter bounding-box overlap requirements.

The operational confidence threshold of 0.25 was used for threshold-dependent precision, recall, F1-score, and confusion-matrix construction. It was not used to truncate the ranked confidence outputs used for AP calculation. The class-only confusion matrices included matched detections. Unmatched ground-truth objects and unmatched predictions were reported separately as false negatives and false positives.

Computational efficiency was assessed using a batch size of one on the same workstation for both detectors. Mean latency was expressed in milliseconds per image, and processing speed was calculated as:

$$\text{FPS} = \frac{1000}{\text{Latency}_{\text{ms}}}, \quad (10)$$

where $\text{Latency}_{\text{ms}}$ denotes the mean inference latency per image in milliseconds.

During conditional object-level processing, each accepted detection was cropped from the original image and routed according to its predicted class. Packaging and pills were processed through separate pathways because their recognition cues differ. Packaging recognition primarily depends on printed text, while pill recognition depends on visual characteristics such as shape, color, surface appearance, and visible imprint patterns.

When an object was classified as packaging, its region of interest was rotated by 90° when its height exceeded 1.2 times its width. The region was converted to grayscale and processed using a bilateral filter with a diameter of 9 and sigma values of 75. EasyOCR was used as the OCR implementation and was configured with the English recognition model (JaidedAI, 2020). The selected OCR pipeline is consistent with established scene-text detection and recognition approaches represented by CRAFT and CRNN (Baek, 2019; Shi, 2017).

The extracted text was converted to lowercase, non-alphanumeric characters were replaced with spaces, and repeated whitespace was normalized. A predefined stopword list was subsequently applied, and tokens containing fewer than four characters were excluded. Medicine candidates were ranked using the RapidFuzz weighted-ratio scorer, (RapidFuzz Project, 2020). Candidate matches with scores below 70 on a scale from 0 to 100 were excluded.

Packaging recognition was evaluated at two levels. Raw OCR transcription quality was measured using Character Error Rate, Word Error Rate, and exact-match accuracy. CER and WER were based on Levenshtein edit operations between predicted and manually verified ground-truth transcriptions (Levenshtein, 1966; Drobac & Lindén, 2020). Equations (11) and (12) therefore follow standard edit-distance error-rate definitions, while Equation (13) reports the study-level exact-match proportion.

Character Error Rate was calculated as:

$$\text{CER} = \frac{S_c + D_c + I_c}{N_c}, \quad (11)$$

where S_c, D_c, and I_c denote character-level substitutions, deletions, and insertions, respectively, and N_c denotes the total number of ground-truth characters.

Word Error Rate was calculated as:

$$\text{WER} = \frac{S_w + D_w + I_w}{N_w}, \tag{12}$$

where S_w, D_w, and I_w denote word-level substitutions, deletions, and insertions, respectively, and N_w denotes the total number of ground-truth words. CER and WER were calculated from raw EasyOCR outputs before stopword removal, candidate filtering, or fuzzy matching. Exact-match accuracy measured the proportion of raw OCR outputs that were identical to their corresponding ground-truth transcriptions, following the strict exact-match criterion commonly used in OCR evaluation (Singh & Middleton, 2025). It was calculated as:

$$\text{ExactMatch} = \frac{N_{\text{exact}}}{N_{\text{OCR}}}, \tag{13}$$

where N_{exact} denotes the number of OCR outputs that exactly matched their corresponding ground-truth transcriptions and N_{OCR} denotes the total number of evaluated OCR samples.

Semantic medicine retrieval was evaluated after text normalization and fuzzy matching. Performance was reported using Top-1 and Top-3 accuracy, following the ranked-candidate evaluation principle used in large-scale recognition benchmarks (Russakovsky *et al.*, 2015). Equation (14) expresses the corresponding Top-k accuracy.

$$\text{Top-}K = \frac{1}{Q}\sum_{q=1}^{Q} \mathbb{I}\left(y_q \in \hat{\mathcal{Y}}_{q,K}\right), \tag{14}$$

where Q denotes the number of evaluated queries, y_q denotes the correct medicine identity, $\hat{\mathcal{Y}}_{q,K}$ denotes the set of the K highest-ranked candidates for query q, and $\mathbb{I}(\cdot)$ denotes the indicator function.

When an object was classified as a pill, its cropped region was resized to 224×224 pixels and converted into a tensor. A ResNet-18 model initialized using ImageNet-pretrained weights was used as the visual feature extractor (He *et al.*, 2016; Russakovsky *et al.*, 2015). Its final classification layer was replaced with an identity operation, producing a 512-dimensional visual feature vector. Each feature vector was normalized using the L2 norm before retrieval.

The pill-retrieval evaluation used 83 query images and 458 reference images representing 62 pill identities. Query images were kept separate from their corresponding reference images. Query and reference embeddings were compared using cosine similarity, a standard vector-space similarity measure (Salton *et al.*, 1975); Equation (15) follows this definition. Equation (16) is an operational margin defined in this study as the difference between the highest and second-highest cosine-similarity scores.

For a query embedding q and a reference embedding r, cosine similarity was calculated as:

$$\text{sim}(\mathrm{q}, \mathrm{r}) = \frac{\mathrm{q}^{\mathrm{T}}\mathrm{r}}{\|\mathrm{q}\|_2 \|\mathrm{r}\|_2}. \tag{15}$$

Reference candidates were ranked from the highest to the lowest cosine similarity. The highest-ranked identity was returned as the Top-1 candidate, while the three highest-ranked identities were used for Top-3 evaluation.

For this study, the relative separation between the two highest-ranked candidates was operationally defined as:

$$\Delta_{\text{sim}} = s_{(1)} - s_{(2)}, \tag{16}$$

where $s_{(1)}$ and $s_{(2)}$ denote the highest and second-highest cosine similarity scores, respectively. Cosine similarity and the similarity margin were treated as relative ranking measures and were not interpreted as calibrated probabilities.

The evaluated pill-retrieval baseline returned the reference identity with the highest cosine similarity and did not apply a similarity-based rejection threshold. Accordingly, pill-retrieval performance was reported using Top-1 accuracy, Top-3 accuracy, mean cosine similarity for correct and incorrect predictions, and the mean Top-1-to-Top-2 similarity margin.

The final recognition output was determined using the detected object class and the ranked candidates produced by the corresponding recognition pathway. Packaging candidates below the fuzzy-matching cutoff were treated as unsupported. Detection outputs were rejected when they failed to satisfy the operational detector confidence requirement.

This rejection procedure was implemented as a technical risk-control mechanism to reduce unsupported identification outputs. It does not establish clinical safety and is not intended to replace professional medication verification.

2.6 Computational Environment and Experimental Setup

Experiments were conducted on a 64-bit Windows workstation equipped with an NVIDIA GeForce RTX 5070 Laptop GPU with 8 GB of VRAM, an Intel® Core™ Ultra i7 processor, and 32 GB of system memory. YOLOv8n was implemented using Ultralytics, Faster R-CNN using PyTorch/Torchvision, EasyOCR for packaging-text recognition, and an ImageNet-pretrained ResNet-18 backbone for visual pill retrieval. The predefined MedPackID partitions contained 1,010 training images, 216 validation images, and 218 frozen test images, corresponding to 1,073, 234, and 233 annotated objects. Both detectors were evaluated on the same frozen test images, ground-truth annotations, class definitions, and object-matching rules. The OPPO Reno12 smartphone was not used for live detection; it was used only for dataset image acquisition.

3. Results

3.1 Object-Detection Results

This section presents the object-detection performance of YOLOv8n and Faster R-CNN on the frozen MedPackID test partition. The evaluation covers class-wise detection accuracy, overall mAP and computational efficiency, as well as error patterns and confidence-threshold behavior. All results were obtained using the same test images, ground-truth annotations, class definitions, and object-matching criteria to maintain a consistent comparison between the two detectors.

3.1.1 Detection Accuracy and Computational Efficiency

Table 5 presents the class-wise performance of Faster R-CNN and YOLOv8n on the frozen test partition. Precision, recall, and F1-score represent performance at the predefined operational confidence threshold, whereas AP@0.5 summarizes the precision-recall relationship across ranked confidence scores. These metrics are widely used for evaluating classification and object-detection performance (Sokolova & Lapalme, 2009; Lin *et al.,* 2014).

Table 5. Class-wise object-detection performance on the frozen MedPackID test partition.

Model	Class	Precision	Recall	F1-Score	AP@0.5
Faster R-CNN	Packaging	0.985	0.957	0.971	0.945
Faster R-CNN	Pill	0.945	0.905	0.925	0.872
YOLOv8n	Packaging	0.985	0.978	0.982	0.977
YOLOv8n	Pill	0.968	0.947	0.957	0.958

The overall detector performance is summarized in Table 6. The mAP@0.5 values were calculated as the arithmetic mean of packaging and pill AP@0.5, consistent with standard object-detection evaluation practice (Lin *et al.*, 2014). Thus, Faster R-CNN obtained (0.945 + 0.872)/2 = 0.909, while YOLOv8n obtained (0.977 + 0.958)/2 = 0.968 after rounding. With batch size one on the same workstation, YOLOv8n required 12.5 ms per image (80.0 FPS), whereas Faster R-CNN required 46.8 ms per image (21.4 FPS).

Latency was measured using a batch size of one on the same workstation. Frames per second were calculated as 1,000 divided by the mean latency in milliseconds. In this comparison, YOLOv8n achieved a higher mAP@0.5 of 0.968, compared with 0.909 for Faster R-CNN. The difference of 0.059 indicates that YOLOv8n provided more consistent localization across the packaging and pill classes under the common test protocol. YOLOv8n also achieved a higher mAP@0.5:0.95 of 0.707, compared with 0.645 for Faster R-CNN.

Table 6. Overall performance and computational efficiency of the evaluated object detectors.

Model	mAP@0.5	mAP@0.5:0.95	Latency (ms/images)	FPS
Faster R-CNN	0.909	0.645	46.8	21.4
YOLOv8n	0.968	0.707	12.5	80.0

Both models showed lower performance at mAP@0.5:0.95 than at mAP@0.5. This reduction indicates that object localization became more difficult under stricter IoU requirements, consistent with the COCO evaluation protocol, which averages detection performance across multiple IoU thresholds from 0.50 to 0.95 (Lin *et al.*, 2014). Packaging objects achieved higher AP@0.5 than pill objects for both detectors. For Faster R-CNN, packaging AP@0.5 was 0.945, while pill AP@0.5 was 0.872. For YOLOv8n, the corresponding values were 0.977 and 0.958. The difference was more pronounced for Faster R-CNN, suggesting that its region-proposal and bounding-box refinement stages were more sensitive to the visual properties of pill objects.

Packaging objects generally occupied larger image regions and contained rectangular edges, printed structures, and relatively stable shapes. In contrast, pills were often smaller and showed greater variation in shape, scale, color, surface appearance, orientation, and background contrast. These characteristics are consistent with challenges reported in pill-recognition literature, where small object size, visual similarity, and acquisition variability complicate reliable recognition (My *et al.*, 2025; Usuyama *et al.*, 2020).

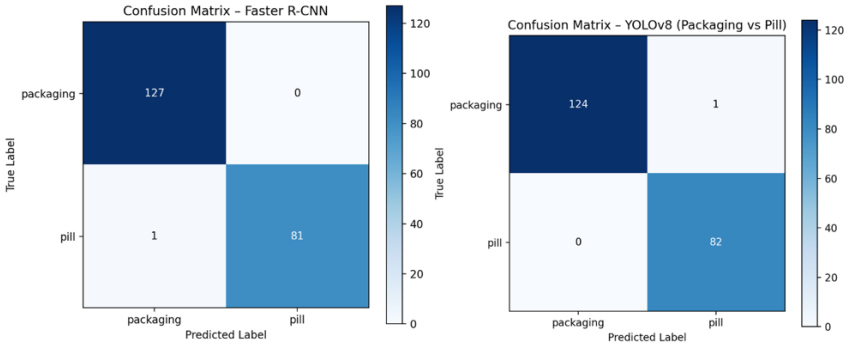
The threshold-specific results also showed that YOLOv8n obtained higher recall for both classes. Packaging recall increased from 0.957 for Faster R-CNN to 0.978 for YOLOv8n, while pill recall increased from 0.905 to 0.947. This pattern indicates that YOLOv8n missed fewer ground-truth objects at the selected operating threshold.

The precision difference was smaller. Both detectors achieved packaging precision of approximately 0.985, while pill precision was 0.945 for Faster R-CNN and 0.968 for YOLOv8n. These results indicate that YOLOv8n missed fewer ground-truth objects at the selected operating threshold while maintaining comparable precision.

YOLOv8n was also computationally more efficient in this comparison. Its mean latency of 12.5 ms per image corresponded to approximately 80.0 FPS, whereas Faster R-CNN required 46.8 ms per image and achieved approximately 21.4 FPS. This difference is consistent with the unified one-stage design of YOLO-style detectors (Redmon *et al.*, 2016) and the additional region-proposal and refinement operations used by Faster R-CNN (Ren *et al.*, 2015). However, the observed differences should not be attributed solely to architectural design because pretrained weights, image transformations, augmentation, batch size, optimization settings, confidence filtering, and non-maximum suppression configurations can also affect performance.

3.1.2 Error Patterns and Confidence-Threshold Behavior

Figure 3 presents classification outcomes among matched detections. The matrices describe whether localized objects were assigned to the correct pharmaceutical category. They do not represent missed objects or predictions without corresponding ground-truth objects and should therefore be interpreted together with precision and recall, which quantify false-positive and false-negative behavior (Sokolova & Lapalme, 2009).



(a) (b)

Figure 3. Confusion matrix of models: (a) Faster R-CNN model, (b) YOLOv8n

Matched-detection confusion matrices contained few off-diagonal entries. Unmatched predictions and unmatched ground-truth objects were reported separately through the false-positive and false-negative counts reflected in precision and recall.

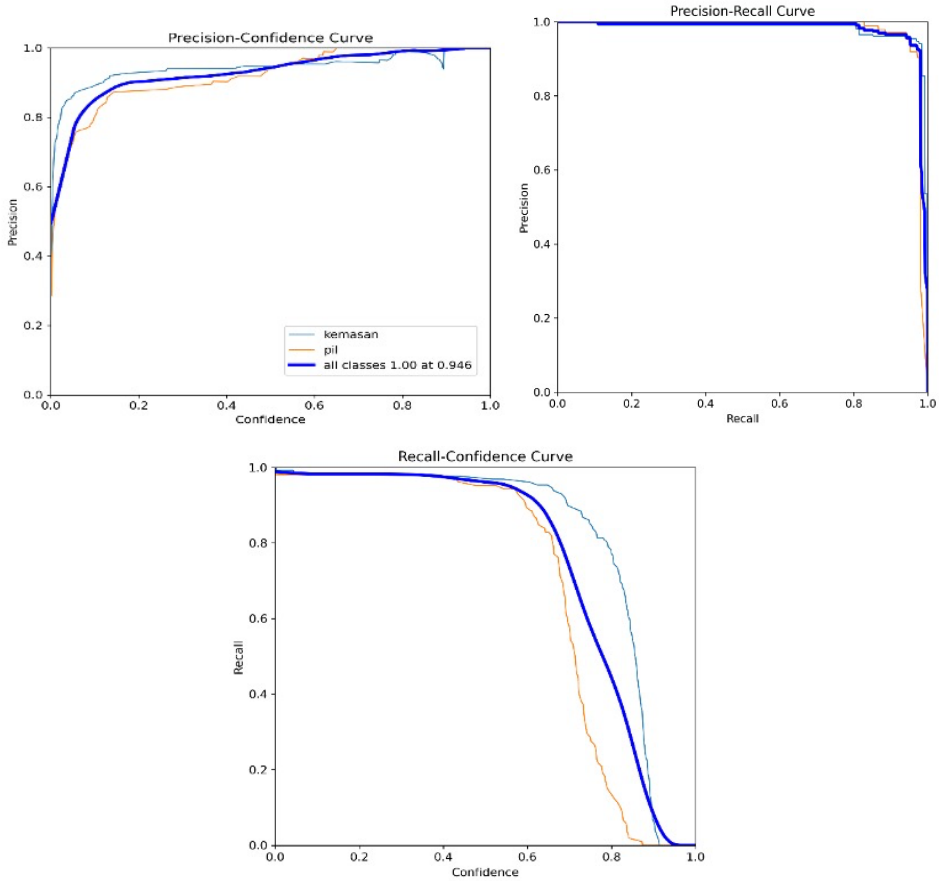


Figure 4. Precision, Recall and F-1 Confidence Curve for YOLOv8n

Figure 4 shows the threshold-dependent behavior of YOLOv8n. Increasing the confidence threshold reduced the number of low-confidence predictions and generally increased precision, while recall decreased because some correct detections were also excluded. This precision-recall trade-off is a standard characteristic of confidence-threshold-based object detection (Sokolova & Lapalme, 2009; Lin *et al.,* 2014). Average Precision was calculated from ranked confidence outputs rather than from the single operational threshold of 0.25.

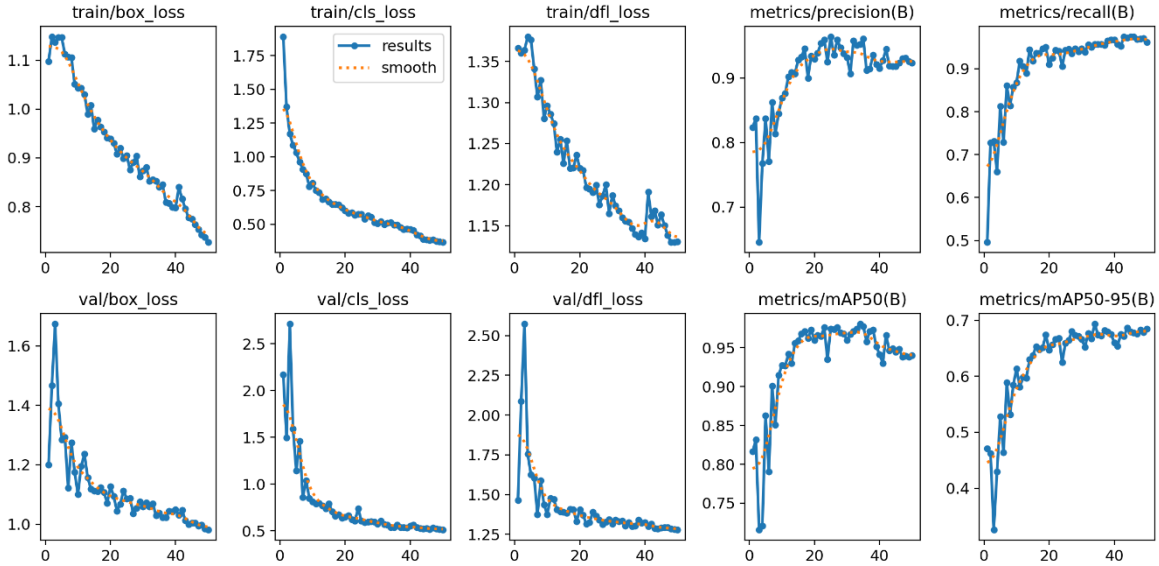


Figure 5. YOLOv8n training and validation performance curves on the MedPackID dataset.

Figure 5 shows the training and validation behavior of YOLOv8n across the reported training run. Bounding-box loss, classification loss, and distribution focal loss decreased during training, while validation precision, recall, mAP@0.5, and mAP@0.5:0.95 increased before stabilizing. These patterns are consistent with convergence behavior commonly monitored during YOLO training (Ultralytics, 2023). However, convergence within one training run does not establish robustness to unseen medicine identities, devices, sites, or environmental conditions.

3.2 Packaging OCR and Semantic Matching

Packaging recognition was evaluated at two levels. Raw OCR transcription quality was first measured by comparing EasyOCR outputs with manually verified ground-truth text. The OCR outputs were then normalized and ranked against a medicine dictionary using RapidFuzz WRatio. OCR evaluation commonly uses Character Error Rate and Word Error Rate based on edit-distance operations, while exact-match criteria provide a stricter measure of complete transcription correctness (Drobac & Lindén, 2020; Singh & Middleton, 2025).

The OCR evaluation used a manually verified packaging-text audit set comprising (N = 133) packaging images and a fixed medicine-candidate dictionary containing 243 medicine identities. CER was calculated over all ground-truth characters, WER over all ground-truth words, exact-match accuracy at the OCR-sample level, and Top-K retrieval accuracy at the query level. Ranked Top-K evaluation follows a common recognition protocol in which the correct class is considered successfully retrieved when it appears among the highest-ranked candidates (Russakovsky *et al.,* 2015).

Table 7. Raw OCR and packaging-based medicine-retrieval performance

Evaluation level	Metric	Value (%)
Raw OCR	CER	5.71
Raw OCR	WER	23.08
Raw OCR	Exact Match Accuracy	70.23
Semantic retrieval	Top-1 Accuracy	90.98
Semantic retrieval	Top-3 Accuracy	94.29

The CER of 5.71% indicates that most individual characters were transcribed correctly. The higher WER of 23.08% shows that relatively small character errors affected complete word tokens. This difference is expected because a single substitution, insertion, or deletion can cause an entire word to be counted as incorrect under edit-distance-based OCR evaluation (Drobac & Lindén, 2020) .

Exact-match accuracy reached 70.23%, substantially below the Top-1 semantic retrieval accuracy of 90.98%. The 20.75 percentage-point difference indicates that text normalization and fuzzy matching recovered medicine identities from many OCR outputs that were not exact transcriptions.

This recovery may occur when OCR produces minor spelling deviations, punctuation differences, omitted spaces, or dosage-format variations while preserving sufficient lexical similarity to the correct medicine entry. Semantic matching therefore reduced the effect of minor transcription errors. Top-3 accuracy reached 94.29%, which was 3.31 percentage points higher than Top-1 accuracy. The relatively small gap indicates that most successfully retrieved identities were already ranked first. However, the remaining errors show that lexical similarity alone cannot always distinguish medicines with similar brand names, active ingredients, or dosage variants.

The fuzzy-matching results must also be interpreted in relation to the size and composition of the candidate dictionary. A restricted candidate list can make the retrieval task easier, whereas a larger dictionary containing closely related medicine names and dosage variants produces a more challenging and realistic evaluation.

Figure 6 shows the processing sequence from packaging localization to medicine-candidate ranking. It shows how normalization and fuzzy matching can preserve the correct medicine candidate despite an imperfect OCR transcription. The example is qualitative and should not be interpreted as independent quantitative evidence of system accuracy.

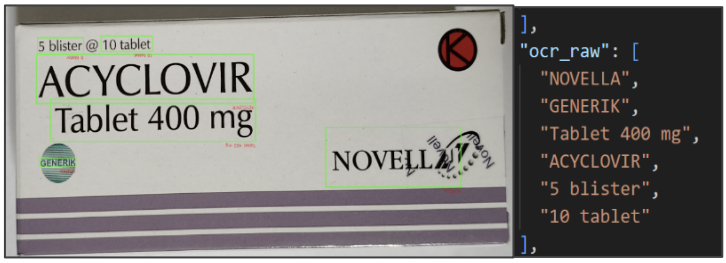


Figure 6. Visual of YOLO-based localization and EasyOCR-based text extraction.

OCR errors were associated with small text, blur, glare, non-horizontal text, partial occlusion, low contrast, and interference from packaging graphics. Such conditions are well-established challenges in scene-text detection and recognition (Baek *et al.*, 2019; Shi *et al.*, 2017). The use of the English EasyOCR model may also limit recognition of local abbreviations and Indonesian pharmaceutical terminology.

3.3 Visual Pill-Retrieval Results

The visual pill-retrieval experiment used 83 query images and 458 reference images representing 62 pill identities. Query images were separated from their corresponding reference images. Each pill image was encoded as a 512-dimensional L2- normalized feature vector using an ImageNet-pretrained ResNet-18 backbone, and query-reference correspondence was measured using cosine similarity.

Table 8. Performance of the visual pill-retrieval baseline.

Metric	Value
Number of query images	83
Number of reference images	458
Number of pill identities	62
Top-1 identification accuracy	85.4%
Top-3 identification accuracy	90.9%
Mean cosine similarity, correct predictions	0.830
Mean cosine similarity, incorrect predictions	0.710
Mean Top-1-to-Top-2 similarity margin	0.110

The Top-1 accuracy of 85.4% indicates that the correct pill identity was ranked first for most queries. Top-3 accuracy reached 90.9%, meaning that the correct identity appeared among the three highest-ranked candidates for most query images. Ranked retrieval using Top-K accuracy is consistent with established image-recognition and pill-identification evaluation protocols (Russakovsky *et al.,* 2015; Usuyama *et al.,* 2020).

The 5.5 percentage-point difference between Top-1 and Top-3 accuracy indicates that four to five additional queries contained the correct identity among the leading candidates but not at the first rank. This gap was larger than the 3.31 percentagepoint gap observed in packaging retrieval, indicating greater candidate ambiguity in the visual pill pathway. Correct predictions produced a mean cosine similarity of 0.830, compared with 0.710 for incorrect predictions. The difference of 0.120 indicates that correct matches were, on average, more similar to their references. However, mean values do not show whether the distributions overlap. The similarity scores therefore cannot independently define a reliable acceptance threshold.

The mean Top-1-to-Top-2 similarity margin was 0.110. A larger margin indicates greater separation between the first- and second-ranked candidates, whereas a small margin indicates ambiguity between visually similar identities. Because the evaluated implementation always returned the highest-ranked reference, the experiment represents closed-set nearest-neighbor retrieval rather than open-set or confidencecalibrated medicine recognition.

The lower Top-1 accuracy of pill retrieval compared with packaging retrieval shows that medicine identity was more readily recovered from explicit text than from generic visual features. Plain pills often share similar colors, circular or oval shapes, smooth surfaces, and weak or absent imprint information. Generic ImageNet features may emphasize these broad visual patterns without adequately representing medicinespecific differences.

These findings support visual embeddings as a useful baseline but not as a complete solution for medicine identification. Domain-specific fine-tuning, imprint-aware feature extraction, metric learning, multiple reference views, and explicit open-set rejection are required to improve fine-grained pill recognition.

4. Discussion

MedPackID evaluates three linked but distinct tasks: pharmaceutical object detection, packaging-text retrieval, and visual pill retrieval. These tasks should not be collapsed into a single accuracy value

because they use different inputs, candidate spaces, and error mechanisms. Detection is the upstream stage. A missed or poorly localized object can directly reduce the quality of OCR or visual retrieval even when the downstream method itself is unchanged.

Under the evaluated configurations, YOLOv8n produced higher mAP and lower latency than Faster R-CNN on the same frozen MedPackID test partition. This result is consistent with the efficiency motivation of one-stage YOLO detection (Redmon *et al.,* 2016) and with prior pharmaceutical work showing that YOLO-family detectors can provide favorable speed-accuracy trade-offs (Tan *et al.,* 2021). Faster R-CNN remains an important comparison because its region-proposal and refinement design is well established (Ren *et al.,* 2015) and its FPN backbone supports multi-scale feature representation (Lin *et al.,* 2017). The present study does not claim that YOLOv8n is universally superior to Faster R-CNN, because the two implementations also differed in pretrained weights, augmentation, image transformation, batch size, and non-maximum-suppression settings. The results therefore represent configuration-specific evidence on MedPackID rather than an architecture-only comparison.

Packaging localization was more accurate than pill localization for both detectors. This pattern is plausible in the context of pharmaceutical imagery because pills are typically smaller, more weakly textured, and more sensitive to partial occlusion and small bounding-box deviations. The result also aligns with the persistent challenges identified in pill-recognition research, including visually similar appearances, acquisition variability, and limited discriminative surface cues (My *et al.,* 2025). Accordingly, the lower pill AP should be interpreted as a dataset characteristic that warrants object-size and appearance-stratified analysis in future work.

For packaging recognition, raw OCR exact-match accuracy was substantially lower than Top-1 semantic retrieval accuracy. The difference indicates that normalization and fuzzy matching recovered correct candidate identities even when EasyOCR produced minor transcription errors. This finding is consistent with the broader scene-text literature, where recognition quality can degrade under irregular layouts and difficult visual conditions (Baek, 2019; Shi, 2017), and with medicine-packet recognition studies that use printed text as a primary semantic cue (Yue *et al.,* 2021). However, the MedPackID packaging result depends on a fixed candidate dictionary of 243 medicine identities. It should therefore be interpreted as constrained retrieval performance rather than unrestricted text-based medicine identification.

The visual pill pathway achieved lower Top-1 accuracy than the packaging pathway and showed a modest Top-1-to-Top-2 similarity margin. This is consistent with the fine-grained nature of pill identification reported by the NLM Pill Image Recognition Challenge (Yaniv *et al.,* 2016) and ePillID (Usuyama *et al.,* 2020), where consumer-reference matching remains difficult when pills share similar color, shape, and surface appearance. In MedPackID, generic ImageNet-pretrained ResNet-18 embeddings provide a transparent baseline, but they do not explicitly model pill imprints, metric-learning objectives, or unseen-identity rejection. The observed 85.4% Top-1 accuracy should therefore be treated as a closed-set baseline rather than as evidence of robust open-set recognition.

The main contribution of MedPackID is consequently benchmark design and conditional multimodal coverage rather than a new detector architecture. It combines packaging and pill bounding boxes in the same Indonesian consumer-device dataset and links detection to separate text-based and visual-retrieval pathways. This complements pill-only public benchmarks (Usuyama *et al.*, 2020; Yaniv *et al*,, 2016) and provides a regional testbed for studying how localization errors, OCR noise, and fine-grained visual ambiguity interact within one workflow. Because the current partitions are image-level rather than identity-disjoint, the reported results characterize within-dataset performance and do not establish generalization to entirely unseen medicine identities, devices, or sites.

5. Limitations and Future Development

MedPackID was collected at one site using one smartphone model, so the reported results establish within-dataset baselines rather than cross-device or cross-site generalization. Each detector configuration was represented by one reported training run, and the detector implementations differed in preprocessing, augmentation, batch size, and non-maximum-suppression settings. Performance differences therefore reflect the evaluated configurations and should not be attributed solely to one-stage versus two-stage architecture.

The predefined partitions were created at the image level rather than the medicine-identity level. Of 243 identities, 141 occurred in more than one partition, including 85 identities shared between training and test. No exact cross-partition duplicates were found, and sensitivity analysis after excluding three confirmed near-duplicate image pairs produced comparable estimates. Even so, the present results do not establish generalization to unseen medicine identities.

The packaging pathway remains sensitive to small text, blur, glare, perspective distortion, and local pharmaceutical terminology, while the pill pathway is limited to closed-set nearest-neighbor retrieval over 62 identities without calibrated rejection. Future development should therefore prioritize external-device and external-site validation, repeated training runs with variability estimates, identity-disjoint partitioning, multilingual or domain-adapted OCR, imprint-aware or metric-learning pill representations, calibrated open-set rejection, and stratified error analysis by object size, blur, illumination, occlusion, and medicine identity.

Table 9. Summary of study limitations and planned future development.

Limitation	Current evidence	Implication	Planned development
Single-site, single-device acquisition	One clinic; OPPO Reno12	External validity is not established	Evaluate additional devices and acquisition sites
Image-level rather than identity-disjoint split	141/243 identities occur in multiple partitions; 85 shared between train and test	Results do not measure unseen-identity generalization	Create identity-disjoint benchmark partitions
Single reported run per detector configuration	No repeated-run variability estimate	Training variability is unknown	Repeat training with variability estimates and confidence intervals
Packaging OCR constraints	English EasyOCR; fixed 243-identity candidate dictionary	Performance depends on text quality and candidate-space composition	Use multilingual/domain-adapted OCR and broader candidate sets
Closed-set pill retrieval	62 identities; no rejection threshold	Unknown identities cannot be rejected reliably	Add metric learning, imprint-aware features, and calibrated open-set rejection

6. Conclusions

This study introduced MedPackID, a public Indonesian dataset containing 1,444 consumer-device images and 1,540 manually reviewed bounding boxes for medicine packaging and pills. Using the same frozen test partition, YOLOv8n achieved mAP@0.5 of 0.968 and mAP@0.5:0.95 of 0.707, while Faster R-CNN achieved 0.909 and 0.645. Packaging-based retrieval achieved 90.98% Top-1 accuracy after OCR normalization and fuzzy matching, and the ResNet-18 closed-set pill baseline achieved 85.4% Top-1 accuracy. These results establish reproducible within-dataset reference performance for a regional benchmark that links object detection to conditional text-based and visual-retrieval pathways. They do

not establish clinical safety, open-set recognition, or generalization to unseen medicine identities, devices, or sites. Future work should address these limitations through identity-disjoint, multi-site, repeated-run, and open-set evaluation.

7. Acknowledgment and Funding Information

The authors also sincerely thank the management of Medica ITS Clinic for granting permission to collect the data.

8. Conflict of Interest Declaration

The authors declare that there is no conflict of interest regarding the publication of this article. The authors also confirm that this manuscript is original and free from plagiarism.

9. Author Contribution Statement

This journal uses Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Qory Hidayati	✓	✓	✓		✓	✓		✓	✓		✓		✓	
Hendra Kusuma				✓	✓	✓		✓		✓	✓	✓		
Muhammad Attamimi				✓						✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

10. Generative AI Statement

During the preparation of this work, the authors used generative AI tools to support language refinement, paragraph organization, draft development, manuscript structuring, and code explanation. These tools were used only as writing and technical assistance, not as a substitute for human critical thinking, interpretation, or scholarly judgment. All AI-assisted outputs were reviewed, edited, verified, and approved by the authors. The authors take full responsibility for the accuracy, integrity, and content of the published article.

11. Data availability

MedPackID Version 1 is publicly available through Mendeley Data under the CC BY 4.0 license at DOI: 10.17632/d23kndp3sv.1. The repository contains the image collection, YOLO and Pascal VOC annotations, dataset configuration files, and supporting metadata.

12. References

Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9365–9374). https://doi.org/10.1109/CVPR.2019.00959

Drobac, S., & Lindén, K. (2020). Optical character recognition with neural networks and post-correction with finite state methods. *International Journal on Document Analysis and Recognition*, *23*(4), 279–295. https://doi.org/10.1007/s10032-020-00359-9

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Hidayati, Q., Kusuma, H., & Attamimi, M. (2026). *MedPackID: A dataset of medicine packaging and pills in Indonesia*. Mendeley Data, Version 1. . URL http://dx.doi. org/10.17632/d23kndp3sv.1. Version 1

JaidedAI. (2020). *EasyOCR: Ready-to-use optical character recognition* [Computer software]. GitHub. https://github.com/JaidedAI/EasyOCR

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2117–2125). https://doi.org/10.1109/CVPR.2017.106

Lin, T. Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014). Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014* (Lecture Notes in Computer Science, Vol. 8693, pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

My, L. N. T., Le, V. T., Vo, T., & Hoang, V. T. (2025). A comprehensive review of pill image recognition. *Computers, Materials & Continua*, *82*(3), 3693–3740. https://doi.org/10.32604/cmc.2025.060793

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. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*, 8024–8035.

RapidFuzz Project. (2020). *RapidFuzz: Rapid fuzzy string matching in Python and C++* [Computer software]. GitHub. https://github.com/rapidfuzz/RapidFuzz

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*, 91–99.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Shi, B., Bai, X., & Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2298–2304. https://doi.org/10.1109/TPAMI.2016.2646371

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Torchvision Contributors. (2026). *Torchvision: Datasets, model architectures, and image transformations for computer vision* [Computer software]. GitHub. https://github.com/pytorch/vision

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1521–1528). https://doi.org/10.1109/CVPR.2011.5995347

Tzutalin. (2024). *LabelImg: Graphical image annotation tool* [Computer software]. GitHub. https://github.com/HumanSignal/labelImg

Ultralytics. (2023). *Ultralytics YOLOv8* [Computer software]. GitHub. https://github.com/ultralytics/ultralytics

Usuyama, N., Larios Delgado, N., Hall, A. K., & Lundin, J. (2020). ePillID dataset: A low-shot fine-grained benchmark for pill identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 3971–3977). https://doi.org/10.1109/CVPRW50498.2020.00463

World Health Organization. (2024). *Medication without harm: Policy brief*. World Health Organization. https://www.who.int/publications/i/item/9789240062764

Yaniv, Z., Faruque, J., Howe, S., Dunn, K., Sharlip, D., Bond, A., Perillan, P., Bodenreider, O., Ackerman, M. J., & Yoo, T. S. (2016). The National Library of Medicine Pill Image Recognition Challenge: An initial report. In *Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop* (pp. 1–9). https://doi.org/10.1109/AIPR.2016.8010584

Yue, X., Lyu, B., Li, H., Meng, L., & Furumoto, K. (2021). Real-time medicine packet recognition system in dispensing medicines for the elderly. *Measurement: Sensors*, *18*, 100072. https://doi.org/10.1016/j.measen.2021.100072

Tan, L., Huangfu, T., Wu, L., & Chen, W. (2021). Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identification. *BMC Medical Informatics and Decision Making*, *21*, 324. https://doi.org/10.1186/s12911-021-01691-8

Singh, L. G., & Middleton, S. E. (2025). Tabular context-aware optical character recognition and tabular data reconstruction for historical records. *International Journal on Document Analysis and Recognition (IJDAR)*, *28*(3), 357–376. https://doi.org/10.1007/s10032-025-00543-9

About Author(s)

Qory Hidayati, S.T, M.T, is currently pursuing her Ph.D. in Electrical Engineering at Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia and lecturer in the Department of Electrical Engineering at Balikpapan State Polytechnic, Balikpapan, Indonesia, where she focuses on Electronic, Control System, Artificial Intelligence, Fuzzy Logic, and Machine Learning. Scopus ID: 57215042421

Dr. Ir. Hendra Kusuma, M.Eng.Sc., is an Associate Professor at the Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia , specializing in Biometric, Image Processing, Machine Learning, Electronic and Data Acquisition. Scopus ID: 55847330100

Muhammad Attamimi, B.Eng., M.Eng., Ph.D., is a lecturer and researcher in the Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, specializing in Visual Recognition, Intelligent Systems, Artificial Intelligence, Intelligent Robotics. Scopus ID: 36439337100