Enhanced Cellular Detection in Cervical Cytopathology: A Systematic Study of YOLOv11 Training Paradigms

Sandra Marcos-Recio , Andrés Barrero-Bueno , Lautaro Rossi-Labianca , Ana Belén Gil-González * , Andrés Cardona-Mendoza , Sandra Janneth Perdomo-Lara

Posted Date: 13 April 2026

doi: 10.20944/preprints202604.0827.v1

Keywords: cervical cytology; artificial intelligence; instance detection; YOLOv11; deep learning; single class; multi class; data augmentation; WSI

 doi:10.20944/preprints202604.0827.v1

Article

Enhanced Cellular Detection in Cervical Cytopathology: A Systematic Study of YOLOv11 Training Paradigms

Sandra Marcos-Recio [1], Andrés Barrero-Bueno [2], Lautaro Rossi-Labianca [1], Ana Belén Gil-González [1,*], Andrés Cardona-Mendoza [1,3] and Sandra Janneth Perdomo-Lara [3]

[1] University of Salamanca, BISITE Research Group, Salamanca, Spain
[2] AIR Institute, Deep Research Lab, Salamanca, Spain
[3] El Bosque University, Bogotá, Colombia, Virtualization and Artificial Intelligence Advanced Solutions Laboratory (SavIA-Lab), Cellular and Molecular Immunology Group (INMUBO)
* Correspondence: abg@usal.es

Abstract

Automated cellular detection using deep learning is a key strategy for optimising cervical cancer screening by reducing the healthcare workload and inter-observer variability. However, analyzing Whole Slide Image (WSI) patches presents challenges like annotation scarcity, morphological complexity, and class imbalance. This study conducts a systematic evaluation of YOLOv11 (*n*, *s*, and *m* variants) to assess the impact of target variable granularity and training paradigms on performance. Four strategies were analyzed: independent and multi-class models, each evaluated at both specific cell label and diagnostic macro-group levels. To ensure clinical robustness, patient-level data partitioning was implemented to prevent information leakage. Performance was measured using precision, *recall*, and mAP (0.5 and 0.5:0.95). The results reveal critical trade-offs between fine-grained discrimination and model generalization when varying architectural complexity and labeling strategies. Findings indicate that diagnostic aggregation improves stability, while single-class training optimizes specialized detection. These results provide methodological guidelines for designing AI-assisted screening systems and establish a foundation for integrating YOLOv11 detectors into Multiple Instance Learning (MIL) frameworks at the WSI scale.

Keywords: cervical cytology; artificial intelligence; instance detection; YOLOv11; deep learning; single class; multi class; data augmentation; WSI

1. Introduction

Cervical cancer continues to be a significant cause of morbidity and mortality among women globally, with a particularly pronounced impact in regions with limited access to systematic screening programs and specialized personnel [1]. Cervical cytology, in either conventional or liquid-based modalities, is a consolidated tool for the early detection of precursor lesions and, by extension, the reduction of mortality associated with cervical cancer [2]. However, cytological interpretation remains a predominantly manual, time-intensive process dependent on the experience of the cytotechnologist or cytopathologist, which introduces inter-observer variability and limits the operational scalability of screening programs [3].

The progressive digitization of cytological slides through WSI technologies, coupled with the recent maturity of deep learning methods, has driven the development of automated medical image analysis systems aimed at supporting specialists in the detection, localization, and prioritization of findings [4]. Recent studies have shown that artificial intelligence can achieve accuracy levels comparable to expert personnel in identifying cervical lesions [5], reinforcing its potential as an assisted screening tool. In digital pathology, and more recently, in cervical cytology, these approaches

 doi:10.20944/preprints202604.0827.v1

address workflows based on high-resolution images, where cell pattern identification must be robust against staining variations, artifacts, blurring, and intra- and inter-patient heterogeneity [6]. This automated analysis paradigm is especially relevant as a support mechanism for population-based screening, allowing for standardized criteria and reduced workload in environments with a shortage of qualified personnel [7]. These capabilities make detection models promising tools that facilitate the integration of artificial intelligence into clinical screening workflows.

Among these, single-stage detectors, such as the YOLO family originally proposed by Redmon et al. [8], stand out for their balance between accuracy and computational efficiency [9]. Their ability to localize multiple small objects per image with reduced inference times makes them particularly suitable for cytological analysis, where numerous cells coexist, often partially overlapping, with high morphological variability and a complex visual context [10]. Consequently, these detectors are positioned as practical candidates for localizing and characterizing abnormal cells in WSI-derived images, operating at the patch level as the unit of analysis, and serving as a component of evidence for clinical screening flows.

However, despite the reported progress, there is still a lack of consensus on methodological decisions that significantly affect the performance of detectors in realistic medical scenarios, particularly when working with limited and imbalanced datasets, which is a common situation in cervical cytology [6]. Some studies have explored architectural variants or training improvements, but with heterogeneity in data partitioning, target variable definition, and metrics, which complicates direct comparisons between approaches [11]. Among the methodological decisions that seem to have the greatest impact are the following:

(i) Single-class versus multi-class formulation; in this work, single-class formulation was defined as training a detector with a single target label using an independent model for each specific cell label or diagnostic macro-group, while multi-class formulation was defined as training a single detector capable of simultaneously discriminating all specific cell labels or diagnostic macro-groups.

(ii) The granularity of the target variable, differentiating between a fine level based on specific cell labels and an aggregated level based on diagnostic macro-groups.

(iii) The use of geometric data augmentation applied only during training as a mechanism to improve generalization without biasing the evaluation.

Additionally, comparative evidence on how these choices interact with model size and capacity (variants with different trade-offs between precision and efficiency) remains limited, despite the availability of scalable architectures within the same family of models.

In this study, these variables were systematically studied through a controlled experimental evaluation of YOLOv11 for cell detection in digital cervical cytology. YOLOv11 was selected for its suitability for detecting small and numerous objects, as well as for offering scalable variants within the same architectural family, allowing for the comparison of models with different capacities under homogeneous conditions. Specifically, the *n* (*nano*), *s* (*small*), and *m* (*medium*) variants, which represent configurations of increasing complexity, were analyzed. Training strategies that combined single-class/multi-class formulations with two levels of target variable granularity—fine (specific cell labels) and aggregated (diagnostic macro-groups)—were compared, considering the effect of data augmentation applied exclusively to the training set. The evaluation was designed with patient-level splitting to minimize information leakage between training, validation, and testing, following digital pathology best practices [12] and approximating the measured performance to a real clinical use scenario.

The contribution of this study is to provide a reproducible comparative framework to understand the impact of (a) problem formulation (single-class vs. multi-class), (b) target variable granularity, (c) data augmentation, and (d) model size in a biomedical context with data constraints and class imbalance. In addition to informing the practical choice of training configurations for YOLOv11 detectors in cervical cytology, this analysis establishes a useful methodological foundation for integrat-

 doi:10.20944/preprints202604.0827.v1

ing WSI-level information and exploring global aggregation schemes, including approaches such as Multiple Instance Learning (MIL).

Based on the issues presented and the need to establish robust methodological criteria for using state-of-the-art detection models in digital cytology, this paper is structured as follows: Section 2 reviews the state of the art, addressing the challenges of AI in cervical cytology, the influence of the Bethesda taxonomy on model design, and the positioning of the YOLO family in this field. Section 3 details the experimental methodology, including the data source, factorial design of the experiments (single-class vs. multi-class strategies and target variable granularity), and strict patient-level partitioning protocol to ensure clinical validity. Subsequently, Section 4 presents the quantitative results, analyzing the effects of data augmentation, model capacity, and the interaction between the evaluated factors using precision, recall, and mAP metrics. Finally, Section 5 discusses the implications of these findings. Section 6 summarizes the study's conclusions, main limitations, and future lines of work.

2. Related Works

Cervical cytology is a particularly relevant scenario for the application of computer vision and deep learning techniques because of the massive nature of screening programs and the existence of morphological patterns with well-defined diagnostic significance [4]. However, the transition from controlled experimental environments to digital pathology scenarios based on Whole Slide Images (WSI) has revealed methodological limitations that affect the external validity and reproducibility of the proposed systems. Consequently, the analysis of the state of the art should not only focus on architectural evolution but also on experimental design decisions, labeling granularity, and mechanisms for integrating evidence at the full-slide scale [6,12,13]. From this perspective, this section first reviews the main methodological challenges and evolution of analysis approaches in cervical cytology, and then situates these advances within the broader framework of WSI analysis using Multiple Instance Learning.

2.1. AI in Cervical Cytology: Methodological Challenges and Evolution of Approaches

Cervical cytology represents a particularly favorable domain for artificial intelligence because of the repetitive nature of screening, the possibility of digitizing entire slides, and the availability of relatively standardized morphological criteria in clinical classification systems such as Bethesda [2,14]. However, recent systematic reviews have consistently pointed out that the reported model performance depends largely on methodological factors that are rarely controlled homogeneously, including inter-laboratory variability in sample preparation and staining, severe imbalance between diagnostic variables, and the high cost of exhaustive cell-level annotation [5–7]. In this context, system robustness depends not only on the architecture used but also on the experimental design, including data partitioning, target variable definition, and selection of clinically informative metrics [11,15].

Despite these limitations, various studies have indicated that integrating automated systems into large-scale screening workflows is feasible, provided that these systems are validated with sufficient clinical rigor [5,7,12]. In particular, these approaches can reduce the burden associated with repetitive tasks and support the initial prioritization of samples in assisted screening environments [5,7]. Along these lines, specific reviews on AI applied to Papanicolaou cytology, such as the one presented at PACBB'24, have shown that the recent evolution of the field has been marked not only by architectural improvement but also by the need for more homogeneous and transferable evaluation protocols for the clinical environment [16].

From a clinical perspective, the Bethesda terminology introduces a hierarchical structure that directly influences the model design [2,14]. Cellular labels such as ASC-US, LSIL, or HSIL have specific diagnostic meanings but can also be grouped into more aggregated levels closer to clinical decision-making [2,14]. Therefore, labelling granularity is not a purely formal issue but a methodological decision that affects problem separability, learning stability in the presence of imbalance, and system output interpretability [6,17].

Early deep learning work in digital cytology was primarily oriented toward classifying previously cropped cells in relatively controlled scenarios, demonstrating that convolutional neural networks

 doi:10.20944/preprints202604.0827.v1

could learn discriminative representations to differentiate between normal and abnormal cell phenotypes. [6,7]. However, as interest has shifted toward more realistic contexts, segmentation and detection tasks in dense scenes with cell overlap, contrast variations, artefacts, and morphological heterogeneity have become increasingly relevant. In this framework, segmentation strategies and instance detectors have taken centre stage, allowing for the localisation of multiple cellular structures in a single image and better approximating real digital cytology analysis conditions [6,10].

Among instance detectors, single-stage architectures, particularly the YOLO family, have gained prominence owing to their balance between accuracy and computational efficiency [8,9]. This property is particularly useful in cervical cytology, where multiple small objects must be localised per image under the practical constraints of inference time and scalability. Consequently, various adaptations of YOLO-based models have shown promising results for the detection of abnormal cells in complex scenes [9,10]. However, even within this line, the literature continues to show high methodological heterogeneity in aspects such as data partitioning, target variable definition, augmentation protocols, threshold selection, and evaluation metrics [6,11,15]. This diversity hinders direct comparisons between studies and limits the extraction of operational conclusions on key issues, such as the convenience of maintaining multi-class formulations, simplifying the problem through single-class strategies, or grouping the labels into more stable diagnostic variables.

2.2. From Local Patch Analysis to WSI Diagnostic Integration via Multiple Instance Learning

Although a substantial portion of the digital cytology literature has focused on local tasks, such as classification, segmentation, or cell-level detection, the actual diagnostic act occurs at the full-slide level. Therefore, WSI analysis constitutes a framework closer to clinical practice, in which relevant information is distributed across multiple heterogeneous regions and cannot always be summarised using independent patch-level decisions. This transition from local analysis to global interpretation has driven the development of hierarchical strategies capable of combining cellular finding detection with diagnostic aggregation mechanisms [6,12].

In this context, Multiple Instance Learning (MIL) has been consolidated as one of the most relevant methodological frameworks for WSI analysis when exhaustive cell-level annotations are unavailable. Under this paradigm, each slide is modelled as a set of instances or patches, and the model learns to identify the regions that contribute most significantly to global decision-making. These approaches have proven useful in both histopathology and cervical cytology, especially in weak supervision scenarios, and can be integrated with self-supervision strategies or previous local feature extractors [6,12,13].

From this perspective, cellular detectors should be understood not only as localisation tools but also as potential extractors of intermediate evidence within hierarchical pipelines oriented toward the final WSI-level diagnostic decision-making. Reviewing the state of the art, it is observed that although object detection has reached a notable degree of technical maturity with architectures such as YOLO, its application in digital cervical cytology continues to face relevant methodological challenges. The literature shows considerable fragmentation in aspects such as labelling granularity, patient-wise partitioning strategies, and the choice between single-class and multi-class formulations, making it difficult to translate the reported results into reproducible clinical scenarios. In this context, the present study addresses this methodological gap precisely: rather than proposing a new architecture, it performs a systematic evaluation aimed at isolating the effects of key training decisions on detection performance. Using YOLOv11 and a rigorous validation scheme that minimises information leakage, the aim is not only to report performance metrics but also to establish a reproducible methodological basis for designing robust cellular extractors with potential future integration into full-slide Multiple Instance Learning systems [6,12,13].

 doi:10.20944/preprints202604.0827.v1

3. Materials and Methods

3.1. Data Source and Patch Generation

This study utilised a proprietary Latin American digital cervical cytology dataset consisting of patches obtained from Whole Slide Images (WSI). A representative subset of the dataset, including labelled patches and corresponding annotations, was publicly documented in an institutional repository [18]. Patch extraction was previously performed using a software prototype and a processing pipeline specifically designed for digital cytological analysis, the methodological description of which was presented in a previous work [19]. The dataset was structured at the patient level and included cell-level Regions of Interest (ROI) annotations, along with the corresponding cell type classification.

Images were anonymised prior to their use in the research, ensuring the removal of patient-identifiable information and compliance with applicable ethical and data protection regulations.

The methodological objective was not to optimise WSI tiling but to evaluate the detector training strategies [12,13]. Therefore, this study assumes that patches have been previously extracted from the WSI scan via an external pipeline (parameters such as patch size, overlap, and effective magnification can be recorded when available; in the experimental reproduction, the determining factors are the patient-wise partition and augmentation protocols). The dataset comprises labelled 640×640 pixel image patches and YOLO-format annotations derived from conventional Papanicolaou test WSIs, as documented in the associated repository record [18].

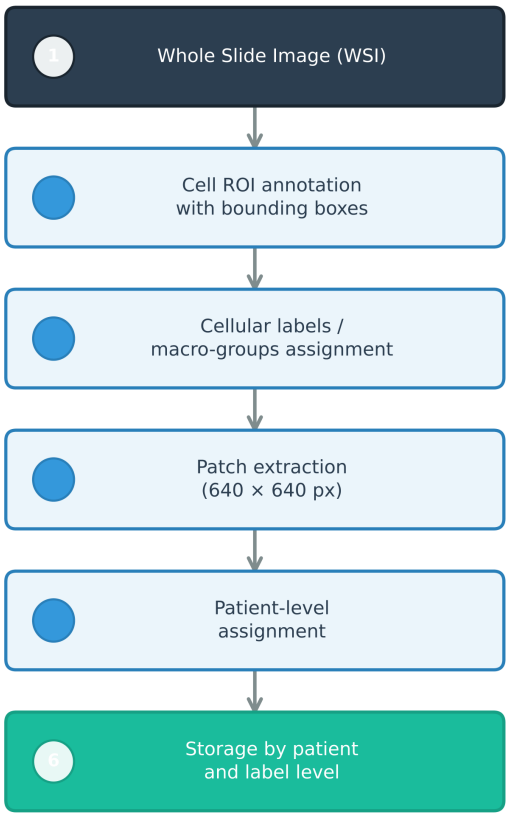

Figure 1. Workflow of dataset preparation from digital cervical cytology WSIs.

Each patch was associated with a patient and a reference label (*cell label* or *macro-group*) based on the original dataset. Annotations followed the YOLO format for detection using bounding boxes, with entries of the form `class_id x y w h`, where `x, y, w, h` are normalised in the interval $[0,1]$ with respect to the patch width and height. This information was used to construct strict patient-wise partitions and derivative macrogroup sets.

 doi:10.20944/preprints202604.0827.v1

Figure 2 shows representative examples of patches from the dataset along with their reference bounding box annotations, illustrating the types of cellular structures that constitute the target output of the detector used.

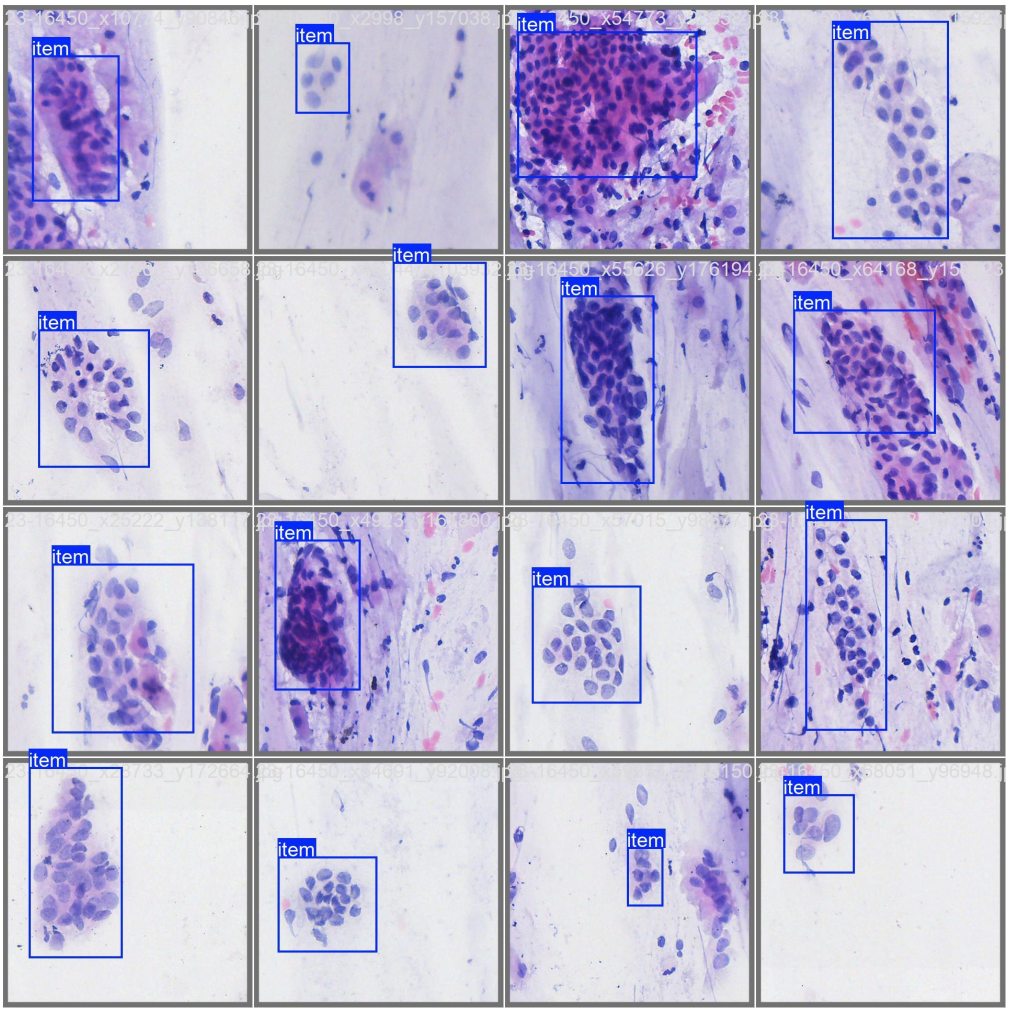

Figure 2. Representative examples of digital cervical cytology patches with their reference bounding box annotations.

3.2. Label Taxonomy and Granularity Levels

In this manuscript, the term *cell label* refers to the fine-level detector target variable (17 cell labels), while *macro-group* denotes a diagnostic aggregation of cell labels (four macro-groups) [5,6].

Two levels of label granularity were considered.

- Cell Labels (17): `CND`, `TRI`, `ACG_NOS`, `ADC`, `SCC-NQ_UNC`, `SCC-QRT_UNC`, `HSIL_UNC`, `LSIL_UNC`, `ASC-US_UNC`, `EC_GL`, `EM_GL`, `INT_UNC`, `LEUC_UNC`, `MET_UNC`, `PBS_UNC`, `RCT_UNC`, `SUP_UNC`.
- Macro-groups (4): `AUXILIARY`, `CANCER`, `PRECANCER`, `NILM`.

As a prior normalisation step, the labels `CRCT` and `DC` were merged into a single `RCT_UNC` to eliminate terminological ambiguity and unify the space before partitioning and training [14]. The complete mapping between the cell labels and diagnostic macro-groups used in the experiments is presented in Table 1.

To generate sets by macro-groups, YOLO annotations were transformed by rewriting the `class_id` of each box from the cell label identifier to the integer identifier of its macro-group (`AUXILIARY`, `CANCER`, `PRECANCER`, and `NILM`), while preserving `x,y,w,h`.

 doi:10.20944/preprints202604.0827.v1

Table 1. Mapping of cell labels to diagnostic macro-groups used in the experiments.

Cell Label	Macro-group
CND	AUXILIARY
TRI	AUXILIARY
ACG_NOS	CANCER
ADC	CANCER
SCC-NQ_UNC	CANCER
SCC-QRT_UNC	CANCER
HSIL_UNC	PRECANCER
LSIL_UNC	PRECANCER
ASC-US_UNC	PRECANCER
EC_GL	NILM
EM_GL	NILM
INT_UNC	NILM
LEUC_UNC	NILM
MET_UNC	NILM
PBS_UNC	NILM
RCT_UNC (CRCT+DC)	NILM
SUP_UNC	NILM

3.3. Experimental Design

The study was evaluated in a factorial manner: (i) training strategy (single-class vs. multi-class), (ii) target variable granularity (specific cell labels vs. diagnostic macro-groups), and (iii) model size (YOLOv11-*n*, YOLOv11-*s*, YOLOv11-*m*) [11]. Additionally, geometric data augmentation was incorporated as a factor (no augmentation vs. with augmentation), applied exclusively to the training set (Section 3.6) [4].

Each combination of the training strategy and label granularity was evaluated for the three model sizes with and without data augmentation, resulting in a full factorial design. Figure 3 shows a schematic representation of the variables evaluated, including the training strategy (single vs. multi-class), target variable granularity (specific cell labels vs. diagnostic macro-groups), YOLOv11 architecture capacity and data augmentation configuration. This design allows for a systematic comparison of the effect of each factor on the detector performance, isolating the impact of training decisions from architectural variations.

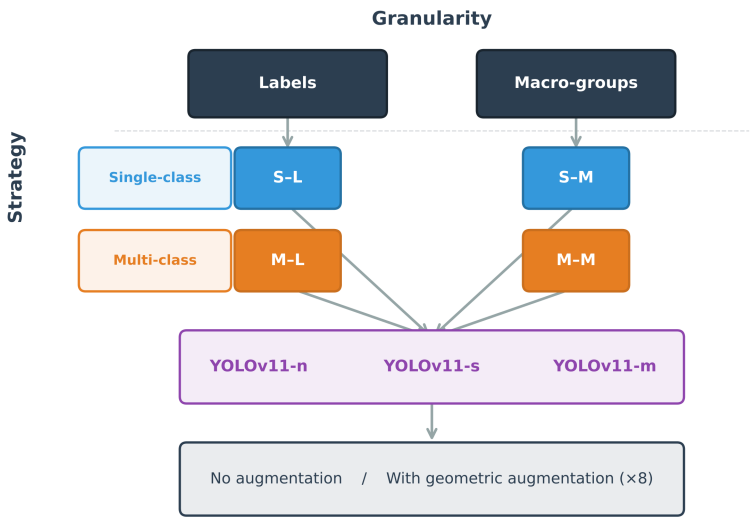

Figure 3. Factorial design.

The four operational scenarios derived from the combination of the training strategy and granularity level are summarised in Table 2.

 doi:10.20944/preprints202604.0827.v1

Table 2. Experimental factors and training scenarios.

ID	Type	Operational description
S-L	Single-class/labels	One YOLOv11 model per cell label; each model detects only its target label (`single_cls=True`, `nc=1`).
S-M	Single-class/macro-groups	One YOLOv11 model per macro-group; each model detects only its target macro-group (`single_cls=True`, `nc=1`).
M-L	Multi-class/labels	A single YOLOv11 model trained to simultaneously discriminate all cell labels (`nc=17`).
M-M	Multi-class/macro-groups	A single YOLOv11 model trained to simultaneously discriminate all macro-groups (`nc=4`).

In the single-class scenarios (S–L and S–M), a set of specialised models is trained for each target label. In the multi-class scenarios (M–L and M–M), a single detector is trained to simultaneously discriminate between all labels at the corresponding granularity level.

3.4. Training Configuration

The experiments were implemented using the YOLO framework (Ultralytics), initialised from pretrained YOLOv11 weights in its *n*, *s*, and *m* variants (e.g. `yolo11n.pt`, `yolo11s.pt`, `yolo11m.pt`) [20]. To maximise the comparability between scenarios, all training sessions (single-class and multi-class) used the same fixed configuration, varying only (i) the `data.yaml` file associated with the scenario (cell label vs. macro group; with or without augmentation) and (ii) the `single_cls` parameter [9,10].

Specifically, the following hyperparameters were fixed.

- Maximum epochs: 100 (`epochs=100`).
- Input size: 640 (`imgsz=640`).
- Early stopping: 20 (`patience=20`).
- Batch size: 16 (`batch=16`).
- Data loading workers: 4 (`workers=4`).
- Dropout regularization: 0.1 (`dropout=0.1`).
- Device: GPU (`device=cuda`).
- Non-deterministic training: `deterministic=False`.
- Internal YOLO augmentations disabled (`mosaic`, `mixup`, `copy_paste`, `perspective`, `shear`, `hsv`, `flip`, `scale`, `translate`, `degrees` = 0).

Execution Environment:

Ultralytics 8.3.198 (YOLOv11 compatibility), Python 3.10.16, PyTorch 2.7.0+cu118, NVIDIA Quadro RTX 4000 GPU (8 GB).

Parameters that were not explicitly specified, such as the optimiser or learning rate policy, were set to their default values. This ensured that the comparison reflected only the effect of the defined experimental factors (Section 3.3) rather than ad-hoc optimisation adjustments. Training utilised the default YOLOv11 loss function, which combined localisation (IoU), classification, and object confidence losses, as implemented by Ultralytics framework. This choice ensured that comparisons between single-class/multi-class scenarios and cell labels/macro-groups exclusively reflected the effect of the defined experimental factors without introducing additional biases from loss function modifications or optimisation tweaks.

Reproducibility:

Experiments are executed with `seed=0`; however, deterministic execution is disabled (`deterministic=False`), so small variations associated with GPU non-determinism and training randomness may be observed. This study reports a single execution for each configuration while keeping the rest of the pipeline fixed.

 doi:10.20944/preprints202604.0827.v1

Comparisons were conducted using an identical protocol, varying only the factors defined in Section 3.3.

Operational clarification is provided as follows:

- Single-class: an independent model is trained per target variable (cell label or macro-group) by enabling `single_cls=True` and using `nc=1` in the `data.yaml` for that model.
- Multi-class: a single model is trained for all labels of the considered granularity level, with `single_cls=False` and `nc=17` (cell labels) or `nc=4` (macro-groups).

3.5. Patient-Wise Partitioning

The partitioning was performed at the patient level to prevent information leakage: the same patient could not appear in more than one set (*train*, *valid*, or *test*) [3,7]. For each valid cell label, the number of available patches per patient was counted, and labels with fewer than 30 total images or fewer than three patients were excluded. The patients were sorted in descending order based on the number of images. The training set was built by assigning patients until approximately 70% of the total images for that cell label were reached. The remainder was divided into *valid* and *test* sets, aiming for a distribution close to 15%/15% through sequential assignments. This procedure ensured approximate proportions of 70/15/15 and maintained strict patient exclusion criteria. Figure 4 summarises the employed partitioning protocols. The data were distributed into training, validation, and test sets, ensuring that samples from the same patient exclusively belonged to one of these subsets. This procedure preserved the independence between partitions and maintained the image distribution necessary for model training and evaluation.

Once the cell label splits were constructed, the macro-group sets were derived by aggregating the labels that compose each macro-group (Table 1), preserving the previous *train*/*valid*/*test* assignment, and relabeling the class identifiers in the annotations.

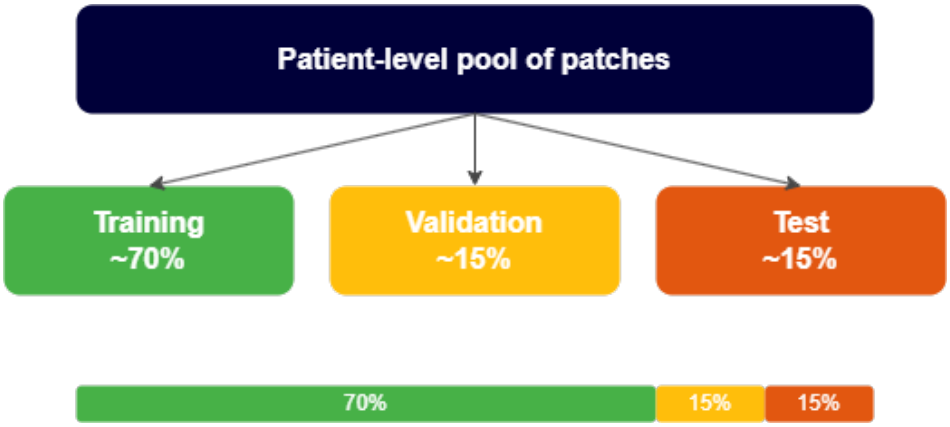

Figure 4. Independent patient-level partitioning protocol.

3.6. Data Augmentation

To increase the diversity of the training set, we applied *offline* geometric transformations to each patch. Each image was multiplied by an expansion factor of $\times 8$ using

- Horizontal flip.
- Vertical flip.
- Rotations of 90°, 180°, and 270°.
- Combinations of flipping and rotation.

All transformations preserved the YOLO boxes by adjusting them to patches. The multiplicative factor on the training set was $\times 8$ (1 original + 2 flips + 3 rotations + 2 combinations of flipping and rotation). Figure 5 summarises the data augmentation protocol employed. Each original patch underwent eight transformations derived from geometric reflections (*flip*) and rotation operations. This procedure preserved the consistency of the bounding boxes and promoted greater model robustness

 doi:10.20944/preprints202604.0827.v1

against spatial variations in the orientation of cytological samples. After applying each operation, the normalised coordinates of the boxes were verified, and those falling outside $[0, 1]$ or with non-positive dimensions were removed.

To ensure reproducibility, these transformations are deterministic; given an original patch, the exact same transformed images with the same boxes are always generated. This is achieved because the applied geometric operations (flips and rotations) do not depend on random values or external states, ensuring consistent results between runs.

The validation and test sets remained unaltered to avoid optimistic evaluations [6]. All other internal framework transformations, if activated by the default, were kept constant across all configurations.

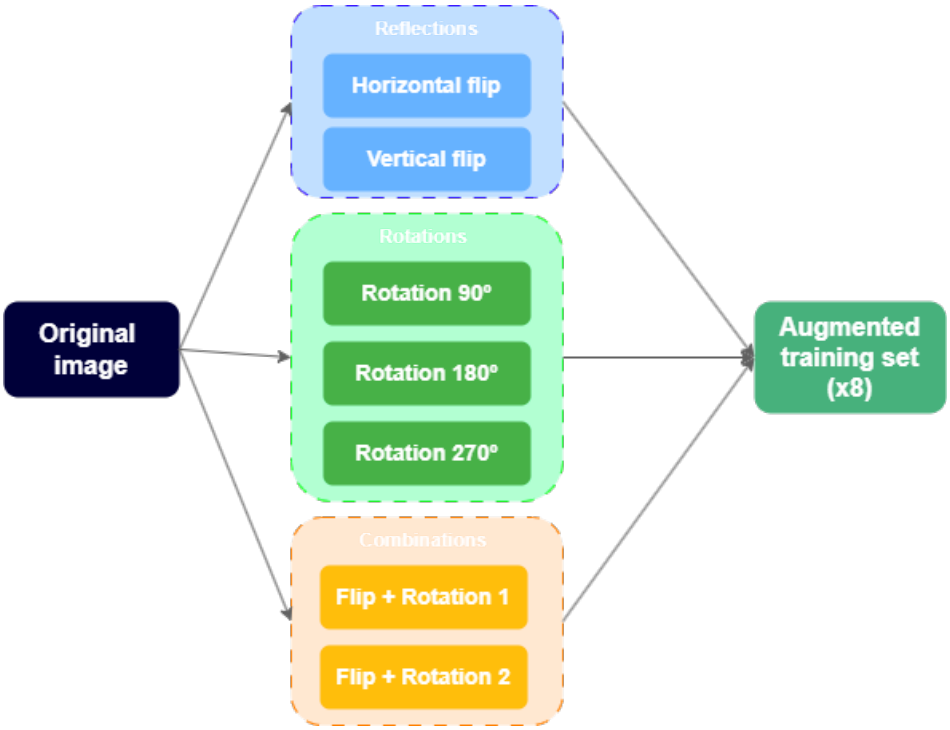

Figure 5. Geometric data augmentation protocol.

3.7. Evaluation Metrics

The performance was quantified using standard metrics employed in object detection: precision (P), recall (R), mAP@0.5, and mAP@0.5:0.95 [8,21]. Denoting true positives, false positives, and false negatives as TP, FP, and FN, respectively, we define

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{1}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{2}$$

mAP@0.5 corresponds to the mean Average Precision (AP) at an overlap threshold of IoU $= 0.5$. mAP@0.5:0.95 averages the AP over multiple IoU thresholds in the range $[0.5, 0.95]$ with an increment of 0.05 (COCO-style criterion) [21], thus providing a stricter measure of localisation quality.

4. Results

4.1. Effect of Data Augmentation

In this subsection, the results obtained without data augmentation and with $\times 8$ geometric augmentation are compared by averaging the performance metrics (P, R, mAP@0.5, and mAP@0.5:0.95) across the three evaluated model sizes (YOLOv11-n, YOLOv11-s, and YOLOv11-m). The reported values correspond to simple arithmetic means between model sizes, without weighting by the number

 doi:10.20944/preprints202604.0827.v1

of samples, to isolate the effect of data augmentation, independently of the model capacity. The analysis focused on a multi-class scenario, where the effect of augmentation could be evaluated in an aggregate manner. The average results obtained in this scenario, with and without geometric augmentation, are summarised in Table 3.

Table 3. Effect of geometric augmentation on average performance in the multi-class scenario.

Granularity Level	Augmentation	P	R	mAP@0.5	mAP@0.5:0.95
Labels	Without	0.201	0.187	0.102	0.052
Labels	With	0.130	0.212	0.111	0.058
Macro-groups	Without	0.276	0.283	0.238	0.124
Macro-groups	With	0.272	0.293	0.241	0.126

At the cell label level, geometric augmentation is primarily associated with increases in *recall*, whereas the impact on mAP@0.5 and mAP@0.5:0.95 is moderate, and the average precision shows a slight decrease. Labels with very few samples maintained values close to zero in both scenarios. At the macro-group level, the effect of augmentation was more stable, with moderate increases in *recall* and slight improvements in mAP, maintaining precision values similar to those of the scenario without augmentation.

Figure 6 shows the evaluation of the effect of synthetic training set expansion on the true positive rate (*recall*). It presents the multiscale average of the evaluated models by comparing the instance recovery capacity at the cell label and macro-group levels. The results suggest that geometric augmentation produces a more consistent improvement in *recall* at the aggregate level than at the fine level.

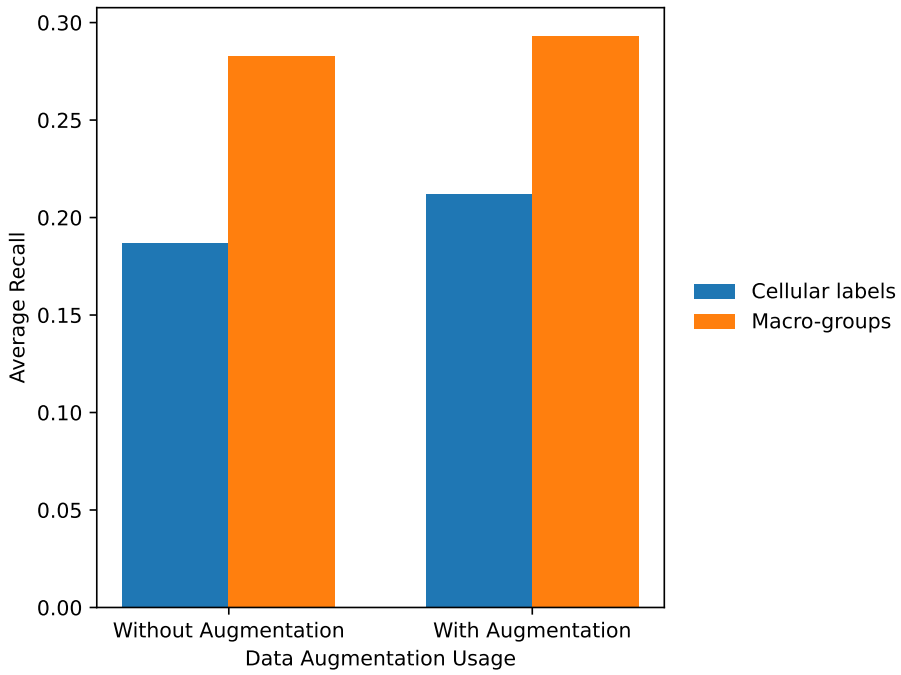

Figure 6. Analysis of the average *recall* under different data augmentation configurations.

4.2. Comparison between Granularities: Labels vs. Macro-groups

This subsection compares the results obtained when training the detectors using specific cell labels (17 labels) versus aggregated diagnostic macro-groups (4 macro-groups). The performance metrics were averaged over the three model sizes and across all evaluated configurations (single and multi-class, with and without data augmentation). The average results obtained for both granularity levels are summarised in Table 4.

 doi:10.20944/preprints202604.0827.v1

Table 4. Comparison of average performance between cell labels and macro-groups.

Granularity	P	R	mAP@0.5	mAP@0.5:0.95
Labels	0.166	0.199	0.107	0.055
Macro-groups	0.274	0.288	0.239	0.125

At the global level, the use of macro groups presented higher values for mAP@0.5 and mAP@0.5:0.95, accompanied by lower inter-configuration variability, as evidenced by a reduction in the interquartile range. In contrast, training at the cell-label level exhibits greater performance dispersion, particularly for labels with lower representations. Figure 7 shows the variability of the mean performance in terms of mAP@0.5 according to the labelling granularity, comparing the cell labels and diagnostic macro-groups. The values represent the aggregation of the results for all YOLOv11 variants and configurations, with and without data augmentation, allowing for an analysis of model stability under different levels of diagnostic abstraction.

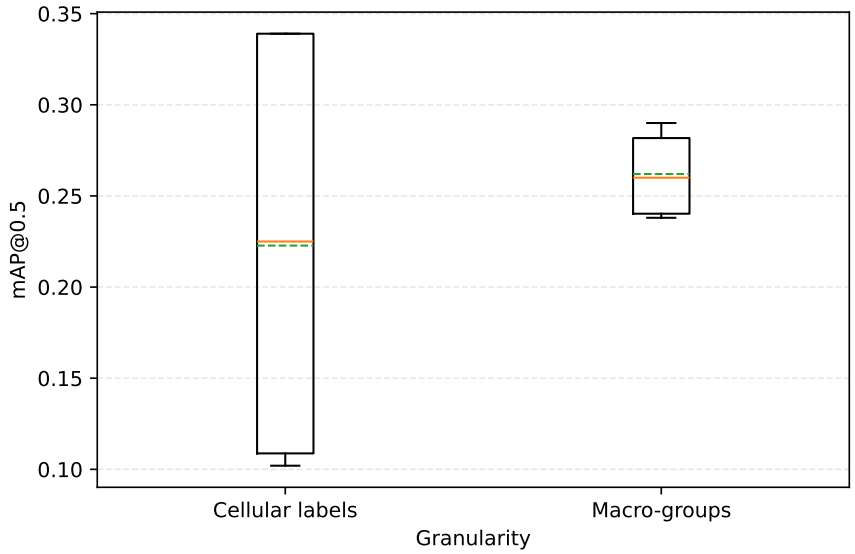

Figure 7. Distribution of mean performance (mAP@0.5) according to granularity.

4.3. Performance of Single-class vs. Multi-class Strategies

This subsection compares the results obtained through single-class training strategies (an independent model per label or macro-group) and multi-class strategies (a single model trained to simultaneously discriminate all labels or macro-groups of the corresponding granularity level), averaging the metrics over the three model sizes and across scenarios with and without data-augmentation. The average results obtained for both training strategies are presented in Table 5.

Table 5. Comparison of average performance between single-class and multi-class training strategies.

Strategy	Granularity	P	R	mAP@0.5	mAP@0.5:0.95
Single-class	Labels	0.392	0.416	0.339	0.183
Multi-class	Labels	0.166	0.199	0.107	0.055
Single-class	Macro-groups	0.287	0.387	0.283	0.129
Multi-class	Macro-groups	0.274	0.288	0.239	0.125

At the cellular label level, the single-class strategy exhibited superior performance compared to the multi-class strategy across all considered metrics, with notable increases in precision, *recall*, mAP@0.5, and mAP@0.5:0.95. Conversely, at the macro-group level, the differences between both

 doi:10.20944/preprints202604.0827.v1

strategies are attenuated: the single-class approach maintains slightly higher values, especially in *recall* and mAP@0.5, but the magnitude of the advantage is considerably smaller than that observed at the fine-grained level. This suggests that diagnostic aggregation reduces the complexity of the problem and narrows the performance gap between the two training paradigms.

Figure 8 illustrates the relationship between mAP@0.5 and the average *recall* for each combination of the training strategy and granularity level. The data points represent the consolidated performance of all YOLOv11 variants, including configurations with and without data augmentation, allowing for a visualisation of the *trade-off* between the localisation capability and model sensitivity according to the learning paradigm used.

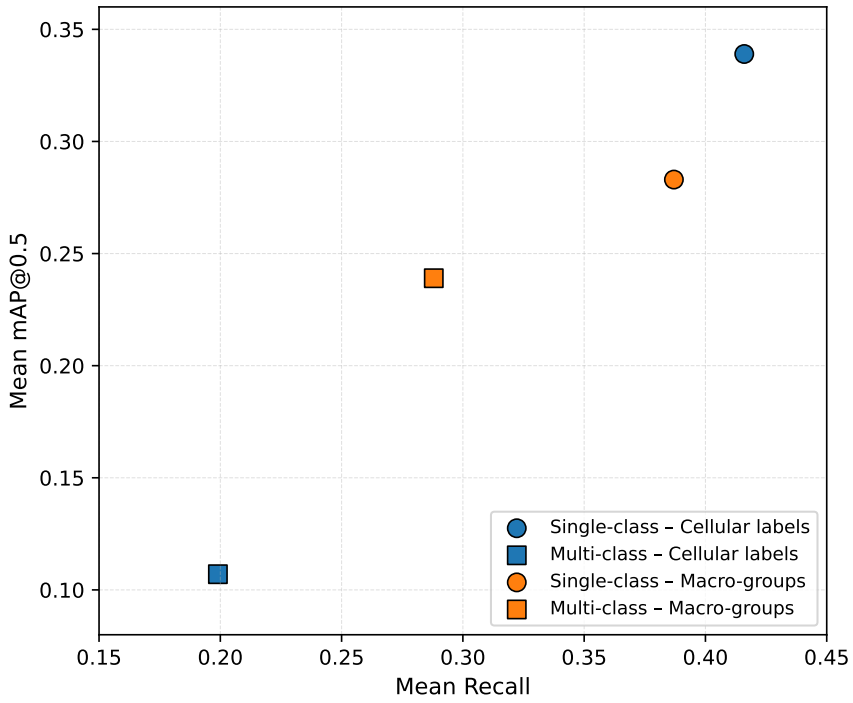

Figure 8. Correlation analysis between mean average precision (mAP@0.5) and average *recall*.

4.4. Combined Effect of Factors

To analyse the joint effect of the primary experimental factors, the results were summarised by averaging the performance metrics across the three model sizes (n, s, and m) using simple arithmetic means for each combination of the training strategy (single-class or multi-class), target variable granularity (cellular labels or macro-groups), and data augmentation usage. The average results for all evaluated combinations are summarised in Table 6.

Table 6. Factorial summary of average performance for the evaluated configurations.

Strategy	Granularity	Augmentation	P	R	mAP@0.5	mAP@0.5:0.95
Single-class	Labels	None	0.392	0.416	0.339	0.183
Single-class	Labels	With	0.380	0.430	0.355	0.196
Single-class	Macro-groups	None	0.256	0.333	0.279	0.119
Single-class	Macro-groups	With	0.290	0.389	0.290	0.134
Multi-class	Labels	None	0.201	0.187	0.102	0.052
Multi-class	Labels	With	0.130	0.212	0.111	0.058
Multi-class	Macro-groups	None	0.276	0.283	0.238	0.124
Multi-class	Macro-groups	With	0.272	0.293	0.241	0.126

Single-class configurations at the label level achieved the highest average mAP values, although greater performance heterogeneity was observed in previous fine-grained analyses. Macro-group-

 doi:10.20944/preprints202604.0827.v1

based configurations exhibited more stable and consistent behaviour, particularly in the multi-class scenario. Data augmentation is primarily associated with increases in *recall*, whereas its impact on mAP depends on the combination of strategy and granularity.

Figure 9 presents a matrix representation of the interaction between the three experimental axes considered: training paradigm (single-class vs. multi-class), target variable granularity (cellular labels vs. macro-groups), and application of data augmentation. The values correspond to the average performance of the three YOLOv11 variants, allowing for the identification of the most favorable configurations in terms of mAP@0.5 under different levels of diagnostic abstraction.

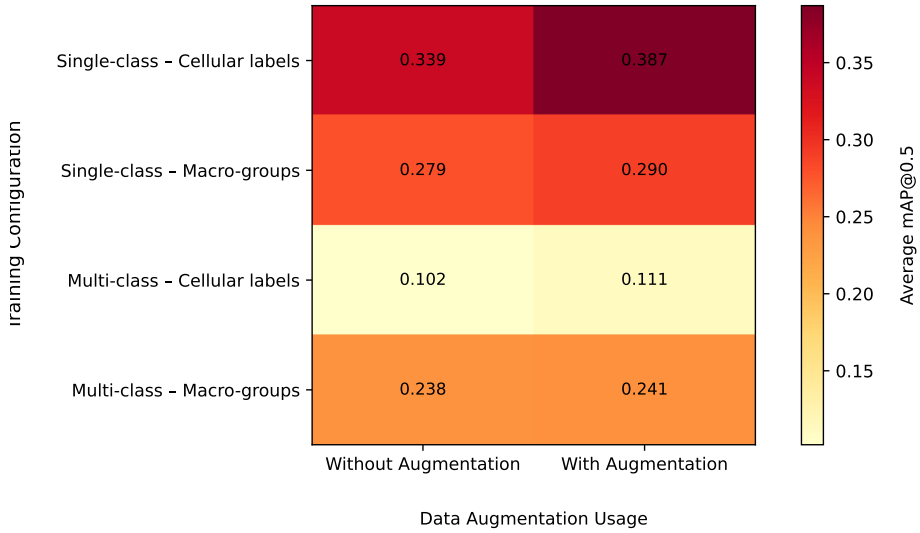

Figure 9. Sensitivity analysis using mAP@0.5 heatmaps.

The observed differences between the configurations should be interpreted descriptively, as no formal statistical significance tests were performed between the experimental scenarios, nor was there an analysis of variability between independent runs. The reported values correspond to simple arithmetic means calculated over the model sizes and the evaluated configurations.

5. Discussion

5.1. Integrated Interpretation of Results

The results obtained allow for a systematic characterization of the detection model behavior under different experimental configurations, simultaneously considering the effects of data augmentation, target variable granularity, and training strategy. The joint analysis of these factors reveals non-trivial interactions that condition both learning stability and the model generalization capacity.

First, geometric data augmentation had a consistent effect on *recall*, particularly in multi-class scenarios. This behavior suggests that the artificial expansion of the training space helps the model become more robust to spatial transformations and morphological variations. From a functional perspective, the increase in *recall* indicates an improvement in sensitivity [4,22], which is especially relevant in biomedical contexts, where missing detections can have critical implications. However, the impact on mAP was more moderate and depended on the granularity of the target variable. This result indicates that geometric augmentation, while increasing detection coverage, does not fully compensate for the structural scarcity of samples in minority labels or morphological ambiguity between similar labels. The slight reduction in precision observed in certain scenarios suggests an increase in false detections, which may be induced by the greater variability introduced during training. Consequently,

 doi:10.20944/preprints202604.0827.v1

geometric augmentation improves system sensitivity but does not inherently resolve the problems associated with extreme imbalances [17].

Second, the comparison between training at the cellular label level and the diagnostic macro-group level revealed a pattern consistent with imbalanced classification problems. Semantic aggregation reduces the fragmentation of the output space, increases the sample density per group, and favors the statistical stability of the optimization process. The substantial increase in mAP when working with macro-groups suggests that a significant portion of the error in the fine-grained multi-class scenario is due to the confusion between labels with similar morphologies. By reducing the semantic resolution of the problem, the structural ambiguity of the decision space is minimized, and the effective separability between outputs is improved [11]. This result indicates that part of the limitation observed at the fine label level is not exclusively attributable to the model architecture but rather to the intrinsic structure of the dataset.

Third, the comparison between single-class and multi-class strategies highlights a *trade-off* between specialization and joint learning. Single-class models achieved higher mAP values at the fine level, indicating that specialization allows for the optimization of specific decisions without competitive interference between outputs. This behavior is consistent with scenarios of heavy imbalance, where joint optimization may favor dominant labels at the expense of minority labels. However, at more aggregated levels of abstraction, the differences between both strategies diminish, suggesting that multi-class learning can benefit from the reuse of shared representations when the problem presents less semantic fragmentation [23]. In this sense, independent training seems to maximize local adjustment per label, whereas joint training favors more general representations, albeit at the cost of greater competition during the optimization.

Furthermore, the factorial analysis confirmed that the effect of data augmentation was not homogeneous but depended on its interaction with the training strategy and the target variable granularity. In single-class configurations at the label level, augmentation produced noticeable improvements in mAP, whereas its impact was more limited in highly fragmented multi-class scenarios. This result suggests that the complexity of the decision space in multi-class problems with high semantic resolution may require not only a quantitative increase in data but also additional strategies for explicit balancing, adaptive regularization, or modification of loss functions that are sensitive to imbalance [24]. The mere geometric expansion of the training set does not appear to be sufficient when the label structure exhibits strong morphological overlap.

5.2. Methodological Implications

From an applied perspective, the results allow for the extraction of several relevant methodological implications for future research. First, semantic aggregation is an effective strategy when sample availability per label is limited and significant morphological overlap exists, as it reduces the effective complexity of the problem and promotes more stable detection. Second, single-class training can be advantageous in scenarios of heavy imbalance, particularly when the primary objective is to maximize the performance of specific labels. Third, geometric augmentation consistently improves system sensitivity but does not replace the need to increase the actual diversity of samples or the convenience of applying specific balancing strategies when the label distribution is extremely skewed.

Altogether, these findings show that the optimization of YOLOv11 detectors in cervical cytology does not depend solely on the architectural capacity of the model but also on the interaction between the annotation scheme, training strategy, and data preparation protocols. Therefore, the methodological decisions adopted in this phase are decisive for building detection systems that can subsequently be integrated into more complex diagnostic workflows at the WSI scale.

6. Conclusions and Future Work

This study provides a systematic evaluation of YOLOv11 training strategies in the context of digital cervical cytology, revealing how the target variable granularity and learning paradigm (single-class vs. multi-class) directly affect the detection capability. The results demonstrate that aggregation into

 doi:10.20944/preprints202604.0827.v1

diagnostic macro-groups favors learning stability and improves overall performance, whereas single-class training offers clear advantages when the goal is to optimize fine-grained detection. Likewise, geometric data augmentation primarily contributes to increasing system sensitivity, although its effect depends on the experimental configuration and does not, on its own, eliminate the limitations stemming from extreme imbalance. Altogether, these findings emphasize that methodological optimization of detectors constitutes an essential step in advancing from isolated patch analysis to robust WSI-based screening systems.

6.1. Study Limitations

Despite the methodological contributions of this work, several limitations must be considered. First, the results were based on simple arithmetic means across model variants without weighting by sample volume or formal statistical significance testing. This absence of explicit contrasts limits the ability to establish categorical differences between certain configurations and aligns with a recurring challenge noted in the literature on machine learning reproducibility, where experimental variability and the lack of adequate statistical analysis can affect result interpretability [15]. Furthermore, each experiment was executed under a fixed protocol without estimating the variability associated with multiple random training initializations.

Second, advanced class-balancing strategies, such as adaptive loss function weighting and label-specific synthetic oversampling techniques, were not explored. Finally, the use of a dataset from a single institutional environment suggests caution regarding inter-center generalization, which is a persistent issue in digital pathology and AI-assisted cytology research.

6.2. Future Research Lines

In future studies, it would be pertinent to incorporate formal statistical analyses and evaluate the inter-run stability through multi-seed initialization schemes. From a technical perspective, it would be relevant to explore imbalance-aware loss functions, hybrid strategies combining multi-class training with label-specific specialized refinement, and more advanced data-balancing protocols.

Finally, external validation on independent cohorts and the integration of these detectors into Multiple Instance Learning (MIL) architectures will constitute decisive steps toward estimating the system's generalization capacity and fostering the transition from local cellular detection to automated whole-slide level diagnosis.

Author Contributions: Conceptualization, L.R.L., A.C.M., S.J.P.L. and A.B.G.G.; methodology, L.R.L., S.M.R., A.C.M. and A.B.G.G.; software, S.M.R., A.B.B., A.B.G.G. and L.R.L.; validation, S.M.R., A.B.B., L.R.L., A.B.G.G., A.C.M. and S.J.P.L.; formal analysis, S.M.R. and A.C.M.; investigation, S.M.R., A.B.B., L.R.L., A.B.G.G., A.C.M. and S.J.P.L.; resources, Colombian League Against Cancer; data curation, A.C.M. and S.M.R.; writing—original draft preparation, A.B.G.G.; writing—review and editing, S.M.R. and A.B.G.G.; visualization, S.M.R. and L.R.L.; supervision, A.B.G.G.; funding acquisition, A.B.G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the International Chair Project on Trustworthy Artificial Intelligence and Demographic Challenge within the National Strategy for Artificial Intelligence (ENIA), in the framework of the European Recovery, Transformation, and Resilience Plan (Reference: TSI-100933-2023-0001), funded by the Secretary of State for Digitalization and Artificial Intelligence and the European Union (Next Generation EU).

Data Availability Statement: The data supporting the findings of this study were fully anonymized. To promote transparency, a curated 10% subset of the dataset is publicly accessible via the University of Salamanca's institutional repository, GREDOS [18].While the complete anonymized dataset remains restricted at this stage to safeguard the ongoing investigation, access may be granted by the corresponding author upon reasonable request, ensuring the responsible and ethical reuse of the clinical data.

Acknowledgments: The authors gratefully acknowledge the support of the SavIA Lab research group and the Agencia Distrital para la Educación Superior, la Ciencia y la Tecnología - ATENEA (Colombia), the Colombian

 doi:10.20944/preprints202604.0827.v1

League Against Cancer, and the Bogotá District Health Secretariat for their support and collaboration in this research.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript.

AI	Artificial Intelligence
AP	Average Precision
ASC-US	Atypical Squamous Cells of Undetermined Significance
COCO	Common Objects in Context
FN	False Negatives
FP	False Positives
GPU	Graphics Processing Unit
HSIL	High-grade Squamous Intraepithelial Lesion
IoU	Intersection over Union
LSIL	Low-grade Squamous Intraepithelial Lesion
mAP	mean Average Precision
MIL	Multiple Instance Learning
NILM	Negative for Intraepithelial Lesion or Malignancy
P	Precision
R	Recall
ROI	Region of Interest
TP	True Positives
WSI	Whole Slide Images
YOLO	You Only Look Once

References

1. Bray, F.; et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **2018**, *68*, 394–424. https://doi.org/10.3322/caac.21492.
2. Saslow, D.; et al. American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. *CA: A Cancer Journal for Clinicians* **2012**, *62*, 147–172. https://doi.org/10.3322/caac.21139.
3. Becker, R.C.; et al. Variability in cytology interpretation and its impact on cervical cancer screening programs. *Journal of Cytology* **2018**, *35*, 83–89. https://doi.org/10.4103/joc.JOC_38_18.
4. Litjens, G.; et al. A survey on deep learning in medical image analysis. *Medical Image Analysis* **2017**, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005.
5. Liu, L.; Liu, J.; Su, Q.; et al. Performance of artificial intelligence for diagnosing cervical intraepithelial neoplasia and cervical cancer: a systematic review and meta-analysis. *eClinicalMedicine* **2024**, *80*, 102992. https://doi.org/10.1016/j.eclinm.2024.102992.
6. Jiang, P.; Li, X.; Shen, H.; et al. A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. *Artificial Intelligence Review* **2023**, *56*, 2687–2758. https://doi.org/10.1007/s10462-023-10588-z.
7. Hays, P. Artificial intelligence in cytopathological applications for cancer: a review of accuracy and analytic validity. *European Journal of Medical Research* **2024**, *29*. https://doi.org/10.1186/s40001-024-02138-2.
8. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
9. Cai, Z.; Zhou, K.; Liao, Z. Object detection in medical images: a review of YOLO architectures and applications. *Biomedical Signal Processing and Control* **2022**, *75*, 103583. https://doi.org/10.1016/j.bspc.2022.103583.

 doi:10.20944/preprints202604.0827.v1

10. Zhao, X.; et al. Deep learning based nucleus detection in cervical cytology images using YOLO. *Computer Methods and Programs in Biomedicine* **2021**, *198*, 105795. https://doi.org/10.1016/j.cmpb.2020.105795.
11. Zhang, Z.; et al. Challenges in deep learning for medical image analysis: a review. *Journal of Imaging* **2021**, *7*, 198. https://doi.org/10.3390/jimaging7090198.
12. Campanella, G.; Hanna, M.G.; Geneslaw, L.; et al. Clinical-grade computational pathology using weakly supervised deep learning. *Nature Medicine* **2019**, *25*, 1301–1309. https://doi.org/10.1038/s41591-019-0508-1.
13. Lu, M.Y.; et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering* **2021**, *5*, 555–570. https://doi.org/10.1038/s41551-021-00725-5.
14. Koss, L.G. *The Bethesda System for reporting cervical cytology: definitions, criteria, and explanatory notes*; Springer, 2021.
15. Pineau, J.; et al. Improving reproducibility in machine learning research. *Communications of the ACM* **2021**, *64*, 56–63.
16. Cardona-Mendoza, A.F.; García-González, A.B.; Díez-Baños, D.; Pérez-Leyva, S.; Rodríguez-Caballero, A. Artificial Intelligence Models for the Automating Papanicolaou Test Reading: A Systematic Review Toward Understanding Clinical Application. In *Practical Applications of Computational Biology and Bioinformatics. PACBB 2024*; Lecture Notes in Networks and Systems, Springer, 2024. https://doi.org/10.1007/978-3-031-87873-2_13.
17. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *Journal of Big Data* **2019**, *6*, 27. https://doi.org/10.1186/s40537-019-0192-5.
18. Cardona-Mendoza, A.; Gil-González, A.B.; et al. Labeled Patches Dataset for Semi-supervised YOLO Training on Cervical Cytology WSI. GREDOS Repository, University of Salamanca, 2025. Data set, https://doi.org/10.71636/g7s3-1n94.
19. Cardona-Mendoza, A.F.; García-González, A.B.; Hortua, H.J.; Labianca, R.L.; Perdomo-Lara, S. Prototype of a Comprehensive System for Automated Generation and Expert Validation of Labeled Patches on Papanicolaou Test WSI Images for Semi-supervised Training of YOLO Models in Automated Cervical Cytology Diagnosis. In *Practical Applications of Computational Biology and Bioinformatics. PACBB 2025*; Springer, 2026; Vol. 1720, *Lecture Notes in Networks and Systems*, pp. 1–11. https://doi.org/10.1007/978-3-032-10634-6_1.
20. Ultralytics. YOLOv11 repository. https://github.com/ultralytics/yolov11, 2024.
21. Lin, T.Y.; et al. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.
22. Perez, L.; Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621* **2017**.
23. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* **2017**.
24. Lin, T.Y.; et al. Focal Loss for Dense Object Detection. In Proceedings of the ICCV, 2017, pp. 2980–2988.