Improving Detection of Minority-Class Breeding Sites with Attention-Based Architectures in UAV-Based Mosquito Surveillance

Vicente Barreira Neto
Universidad de Salamanca
Salamanca, Salamanca, Spain
vicentent@usal.es

Luis Augusto Silva Zendron
Department of Computer Science,
University of Salamanca
Salamanca, Spain
luisaugustos@usal.es

Anita Fernandes
Universidade do Vale do Itajai
Mestrado em Computação Aplicada
Itajaí, Brasil

Wemerson D. Parreira
Polytechnic School, Faculty of
Electrical Engineering, Pontifical
Catholic University of Campinas
Campinas, Brazil

Vivian Félix López Batista
Department of Computer Science,
Faculty of Science, University of
Salamanca
Salamanca, Spain

ABSTRACT

Vector-borne diseases, such as dengue, have reached record epidemic levels in the Americas and are emerging as a new threat in Europe. Traditional surveillance of mosquito breeding sites remains inefficient. While the use of UAVs (drones) and Deep Learning is promising, it faces a critical methodological challenge: class imbalance, which is considered severe in real-world data. This imbalance biases standard models, causing them to ignore the most epidemiologically critical vector sites (minority classes). This paper presents an experimental comparative analysis of four state-of-the-art object detection architectures (YOLOv8m, YOLOv11m, YOLOv12m, and RF-DETR) to evaluate their robustness in this scenario. Performance was measured using standard metrics (mAP) and specialized metrics (Balanced Accuracy, MCC). The results demonstrate that the baseline (YOLOv8m) fails to detect minority classes, achieving a recall of only 8.2% for the bottle class. The YOLOv11m architecture, equipped with spatial attention mechanisms (C2PSA), emerged as the optimal solution. It achieved the highest Balanced Accuracy (68.4%) and MCC (0.487), and improved the bottle class recall by 247% compared to the baseline. We conclude that architectures incorporating spatial attention mechanisms are crucial for the viability of automated epidemiological surveillance in real-world, severely imbalanced environments.

KEYWORDS

Object Detection, Deep Learning, Class Imbalance, YOLO, RF-DETR, UAV, Vector Surveillance, *Aedes aegypti*

1 INTRODUCTION

Vector-borne diseases transmitted by mosquitoes such as *Aedes aegypti* reached alarming proportions in 2024. The Region of the Americas recorded an exceptionally high number of dengue cases, totalling 12,978,140 suspected cases, of which 6,867,682 were laboratory confirmed [1]. Brazil accounted for most of these infections [2]. This volume represents an increase of approximately 240% compared with 2023, when 1,985,289 dengue cases were confirmed in the Americas [3].

By mid-November (Epidemiological Week 44), 4,121,923 suspected cases were reported, with 1,603,225 laboratory-confirmed cases [4], indicating a significant reduction compared with the 2024 total. Despite this decline, transmission levels remain high, underscoring the need for ongoing, advanced strategies for vector surveillance and control.

In parallel, Europe faces a growing risk with the expansion of *Aedes albopictus*. This vector shares reproductive niches similar to *Aedes aegypti*, colonising the same types of artificial containers such as tyres and puddles in urban and peri-urban areas, which supports the applicability of automated surveillance methods to this species as well [5].

When acquiring images in uncontrolled environments, the data distribution often exhibits a natural disparity, with most classes having substantially more samples than others. This asymmetry harms model generalisation by inducing a bias that favours the dominant class and increases uncertainty when recognising underrepresented classes [6].

The proliferation of these vectors is directly linked to urban breeding sites such as containers, tyres, and tanks with standing water [7]. Traditional surveillance based on manual inspections is insufficient due to its high resource requirements, limited coverage, operational inefficiencies, and access limitations [8]. Moreover, the practical deployment of technological solutions faces a critical methodological challenge: class imbalance. In real environments, epidemiologically relevant sites (minority classes) are vastly outnumbered by irrelevant objects (majority class), which can induce learning errors in automated models [6].

The use of Unmanned Aerial Vehicles (UAVs) equipped with machine learning algorithms offers an innovative solution for automated surveillance [5, 9]. Computer vision, using Convolutional Neural Networks (CNNs) and Transformers, enables the identification of breeding sites with high precision.

In this context, this work presents an experimental comparative analysis of four state-of-the-art object detection architectures (YOLOv8m, YOLOv11m, YOLOv12m, and RF-DETR) applied to the automated identification of mosquito breeding sites in aerial imagery. The performance of the architectures is evaluated using

standard metrics (mAP@50, mAP@50:95) and specialised metrics (*Matthews Correlation Coefficient, Balanced Accuracy*), with a particular focus on their behaviour on minority classes. Accordingly, the main objective of the study is to evaluate the effectiveness of these models in scenarios characterised by severe class imbalance.

2 THEORETICAL BACKGROUND AND RELATED WORK

2.1 Evaluated Architectures

To cover the full spectrum of object detection paradigms [10], four state-of-the-art architectures were selected. The "Medium" (m) versions of the YOLO models were chosen because they represent an ideal balance between detection capability and computational efficiency for field deployments. The architectures are:

- YOLOv8m (Reference CNN): Used as the study *baseline*. It is a *single-stage*, *anchor-free* CNN architecture that implements a decoupled detection head (*decoupled head*) and a backbone based on C2f blocks for efficient feature extraction [11].
- YOLOv11m (CNN with Spatial Attention): This architecture tests the spatial attention hypothesis. Its innovation is the incorporation of the C2PSA module (*Convolutional block with Parallel Spatial Attention*), which enables the model to selectively focus on critical regions, making it theoretically well-suited for detecting small or partially occluded objects in imbalanced scenarios [12].
- YOLOv12m (Hybrid CNN–Transformer): Represents a hybrid approach. Its technical core is the Area Attention (A^2) module, which implements a strategic segmentation of feature maps to reduce the computational complexity of attention, combined with R-ELAN networks to stabilise training [13].
- RF-DETR (Pure *End-to-End* Transformer): Represents the pure *transformer* paradigm. It completely eliminates the need for NMS (*Non-Maximum Suppression*), fundamentally differentiating it from YOLO models. It uses a robust DINOv2 backbone pre-trained on 142 million images [14], offering exceptional *cross-domain* adaptation capabilities [15].

2.2 The Challenge of Class Imbalance

Johnson and Khoshgoftaar [6] formalise this challenge by classifying the phenomenon into three levels: mild, moderate, and extreme (or severe). While mild imbalances (up to 1:10) have a marginal impact, ratios exceeding 1:100 are considered extreme. In this study, the observed asymmetry places the problem categorically at the severe level: the disparity reaches a critical peak of approximately 1:277 between the majority class water tank (*watertank*, 71.92%) and the minority class puddle (*puddle*, 0.26%). Under such conditions, standard supervision methods become biased, ignoring the sites of greatest epidemiological relevance and becoming ineffective.

2.3 Related Work

The application of deep learning to vector surveillance has recently been explored as an alternative to manual methods. The seminal work by Passos et al. [9] established the technical feasibility of this approach by introducing the *Mosquito Breeding Grounds* (MBG) dataset, demonstrating that convolutional neural networks can identify *Aedes aegypti* breeding sites in complex urban environments in Rio de Janeiro. However, the original study focused on spatio-temporal consistency and did not explore in depth the impact of attention-based architectures under extreme imbalance.

In real-time detection, YOLOv8 has consolidated itself as a reference standard due to its *anchor-free* architecture and efficiency. Reis et al. [11] demonstrated the robustness of this model for detecting flying objects, validating its suitability for UAV image processing. Despite its overall effectiveness, the standard architecture may present limitations in scenarios of high visual complexity.

Extending the scope to other species, Yu et al. [5] validated the use of UAV imagery to investigate *Aedes albopictus* breeding sites, confirming that the resolution of aerial images is sufficient to identify small containers.

Meanwhile, the evolution of object detectors has sought to address accuracy limitations in challenging contexts. Recent benchmarking studies, such as Sharma et al. [12], highlight the superiority of newer versions of the YOLO family (v11) over predecessors for detecting small objects in precision agriculture. Similarly, Sapkota et al. [15] analysed the transition to Transformer-based (e.g., RF-DETR) and hybrid (YOLOv12) architectures, suggesting greater robustness, albeit at high computational cost.

Benchmarking studies published in early 2025 corroborate the relevance of this comparative analysis. Jegham et al. [16] observed that although YOLOv12 introduces theoretical innovations, YOLOv11 often offers a superior efficiency balance for practical applications. In addition, Zheng et al. [17] demonstrated that improvements in YOLOv11 feature fusion make it particularly effective for *Small Object Detection* (SOD), a critical characteristic for identifying larval breeding sites. In the domain of vector surveillance, recent work [18] reaffirms that deploying deep networks on aerial imagery remains challenging due to noise and background complexity, underscoring the need for robust architectures such as those evaluated in this study.

This work differs in that it applies these state-of-the-art architectures (from the established YOLOv8 to the emerging v12 and RF-DETR) specifically to the severe imbalance problem (1:277) identified in the MBG dataset, a gap not previously addressed in earlier studies.

3 MATERIALS AND METHODS

This section details the methodological procedures used to compare object detection architectures in line with the defined objectives. The approach comprises the following main stages: dataset and pre-processing; experimental methodology; experimental setup and evaluation metrics [9, 10, 15, 19, 20].

3.1 Dataset and Pre-processing

The MBG *dataset* [9, 19] was originally developed for vector surveillance research in Rio de Janeiro, Brazil. Unlike conventional object detection datasets, MBG is natively composed of high-definition video sequences captured by Unmanned Aerial Vehicles (UAVs) in real, uncontrolled urban scenarios.

In its original design, the data included tracking annotations generated in the CVAT platform [21], intended to analyse spatio-temporal consistency. Because this native format had high frame

redundancy and was incompatible with static detectors, it was necessary to implement a pre-processing *pipeline* (Figure 1), following methodological recommendations for dataset preparation [10].

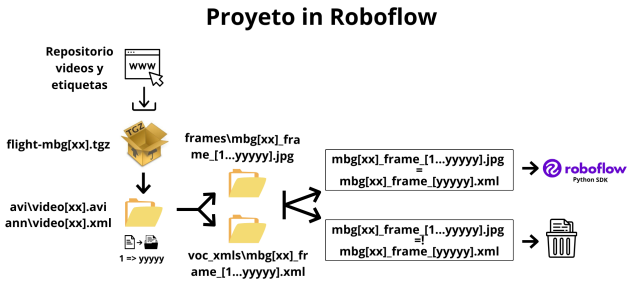

Figure 1: Flowchart of the pre-processing *pipeline*: from video extraction (TGZ) to annotation conversion (XML) and generation of training-ready images (JPG), prepared by the authors.

This *pipeline* was used to analyse XML metadata, selectively extract frames with valid annotations, convert them from the CVAT format to the Pascal VOC standard, and remove inconsistent labels. This process resulted in a final optimised dataset of 18,502 images with 88,564 valid annotations, eliminating about 70% of redundant frames.

For management, versioning, and final formatting, the validated images and annotations were uploaded to the Roboflow platform [22]. From this unified environment, the data were processed and exported in two distinct formats to meet the requirements of each architecture:

- YOLO format: Generated for training YOLOv8m, YOLOv11m, and YOLOv12m. It consists of individual text (.txt) files with normalised coordinates.
- COCO format: Exported exclusively for RF-DETR. It consists of a single JSON file containing the full hierarchical structure.

Analysis of the final class distribution confirmed the study's main technical challenge: severe class imbalance. As detailed in Table 1, the *watertank* class (water tank) accounts for 71.92% of the annotations, whereas the *puddle* class (puddle) represents only 0.26% [19]. Visual examples of each class are shown in Figure 2.

Finally, the dataset was split using stratified sampling [23], preserving the original class proportions: 70% for training, 20% for validation, and 10% for testing.

3.2 Experimental Methodology

Experiments were conducted in a high-performance computing environment equipped with an NVIDIA GeForce RTX 3090 GPU. Training used transfer learning (*transfer learning*), starting from weights pre-trained on the COCO and YOLO datasets. To ensure reproducibility, a fixed random seed (42) was used and deterministic operations were enabled.

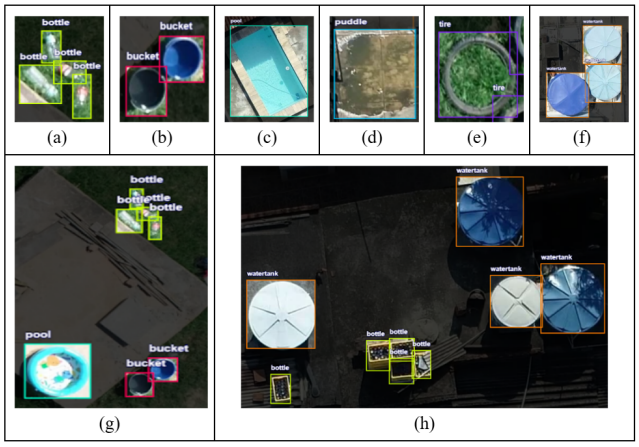

Figure 2: Examples of annotations for the six breeding-site classes in the MBG *dataset*: (a) *bottle*, (b) *bucket*, (c) *pool*, (d) *puddle*, (e) *tire*, (f) *watertank*, (g) example of a rural or suburban scene, and (h) urban.

Table 1: Class distribution in the MBG *dataset*, highlighting severe imbalance [19].

Object	Class	Count	Percentage (%)
bottle	*bottle*	3,776	4.26%
bucket	*bucket*	6,096	6.88%
pool	*pool*	2,951	3.33%
puddle	*puddle*	231	0.26%
tyre	*tire*	11,817	13.34%
water tank	*watertank*	63,693	71.92%
Total		**88,564**	**100.00%**

3.3 Experimental Setup and Evaluation Metrics

To ensure a direct and fair comparison between architectures, a unified experimental configuration was established [6]. All models were trained with an input resolution of 640×640 pixels, using a fixed random seed (42) and deterministic operations to ensure reproducibility. Training was configured for up to 200 epochs, with early stopping (patience of 50 epochs) monitored by the mAP@50 metric. The Adam optimiser was used with an initial learning rate of 0.01, kept constant after an initial warm-up of 3 epochs.

The Adam optimiser was chosen for its consistent performance on object detection tasks, particularly for models in the YOLO family. Adam combines momentum with adaptive second-order estimates, enabling more stable convergence on heterogeneous datasets with strong class imbalance. Recent computer-vision studies show that Adam tends to provide greater robustness in gradient updates compared with traditional optimisers such as SGD, particularly in scenarios with noise, sample variability, and small-scale objects [10, 24, 25].

In addition, modern benchmarks based on YOLO architectures indicate that Adam remains a preferred choice for stability-sensitive training, as its dynamic per-parameter learning-rate adaptation

reduces oscillations during training and improves convergence in deep architectures [12]. This stability is particularly relevant in the context of this study, which is characterised by a severe imbalance in the MBG dataset, where small minority classes require refined updates that are less susceptible to gradient explosions or vanishing gradients.

The central methodological contribution to handling severe class imbalance was adapting the loss function. Distributional Focal Loss [26] was implemented with a weight of 1.5, enabling the model to focus learning on hard examples (minority classes). In addition, the classification loss was reduced to 0.5, while the spatial localisation (regression) loss was increased to 7.5, emphasising geospatial precision, which is critical for surveillance applications.

Standard object detection metrics, mAP@50 (to evaluate overall detection) and mAP@50:95 (to rigorously evaluate localisation precision) [15], were used to evaluate performance. However, given the severe imbalance, the study focused on specialised metrics that provide a more faithful assessment under asymmetric distributions. For this, Balanced Accuracy and the Matthews Correlation Coefficient (MCC) were used, as they mitigate the dominance of the majority class and more reliably assess a model's ability to detect all classes [20].

Balanced Accuracy is defined as the mean of the per-class sensitivities (*recall*):

$$Balanced\ Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \tag{1}$$

MCC is considered a robust metric that yields a value between -1 and +1, where +1 indicates perfect prediction:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2}$$

To interpret recall values for minority classes under extreme class imbalance, a qualitative performance assessment scale was used. This scale is inspired by the concepts of learning degradation described by Johnson and Khoshgoftaar [6]. Since the reference literature does not establish universal numerical thresholds for these categories, the specific percentage ranges were defined empirically for automated epidemiological surveillance. A recall below 10% indicates a nearly complete failure to detect the class, categorised as Severely deficient. Values between 10% and 20% indicate insufficient detection capability for practical public health applications. The range from 21% to 40% denotes an initial but unstable detection capability, labelled as Limited. The 41% to 60% threshold marks the beginning of an Acceptable performance, representing the minimum operational viability for risk mapping. Finally, rates above 60% are considered Sufficient, demonstrating model robustness against imbalance. Table 2 summarises this proposed scale.

Table 2: Qualitative performance assessment scale for minority classes.

Recall Interval (%*)	Performance Assessment
< 10%	Severely deficient
10% – 20%	Insufficient
21% – 40%	Limited
41% – 60%	Acceptable
> 60%	Sufficient

**Scale empirically adapted from the concepts of [6].*

4 RESULTS AND DISCUSSION

In this section, the experimental results are presented and analysed. Following an iterative approach similar to that used by [27], the analysis begins by evaluating the *baseline* architecture (YOLOv8m). Next, the results of subsequent models (YOLOv11m, YOLOv12m, and RF-DETR) are presented, enabling a direct comparison that highlights the impact of each architectural innovation — such as attention mechanisms — on detection performance under severe class imbalance.

4.1 Batch 1: Baseline Performance (YOLOv8m)

The first evaluation used the YOLOv8m model, which served as the *baseline* to quantify the challenge posed by the MBG *dataset*. The model was trained under the unified settings and converged in 110 epochs, with a total training time of 8.7 hours. The global results, shown in Table 3, indicate modest performance, with a final mAP@50 of 17.7% and mAP@50:95 of 12.9%. The imbalance metrics, while indicating a positive correlation (MCC = 0.341) and some balancing capability (Balanced Accuracy = 56.9%), also reveal substantial limitations.

Table 3: Summary of Global Results — YOLOv8m (Baseline).

Category	Metric	Value
Global detection	Final mAP@50	17.7%
	Peak mAP@50	19.0%
	mAP@50:95 Final	12.9%
Imbalance	Balanced Accuracy	56.9%
	MCC	0.341
Efficiency	Training time	8.7 hours
	Epochs run	110 / 200

The true limitation of the *baseline* is exposed by the per-class granular analysis detailed in Table 4. The model exhibited "severely deficient" *recall* (sensitivity) for the epidemiologically critical minority classes [6]. With only 8.2% *recall* for *bottle* and 12.1% for *bucket*, the model fails to detect the vast majority of breeding sites. This performance confirms that a traditional CNN architecture, without specific mechanisms, is strongly biased towards the majority class and is inadequate for effective vector surveillance.

4.2 Batch 2: Correction with Attention Mechanisms (YOLOv11m)

After finding that the *baseline* (YOLOv8m) was inadequate, the second evaluation (Batch 2) focused on the YOLOv11m model. The central hypothesis was that its spatial attention mechanisms, specifically the C2PSA module [12], could address the low *recall* (sensitivity) in minority classes that Batch 1 ignored. The results, obtained

Table 4: Recall (Sensitivity) Analysis for Critical Minority Classes — YOLOv8m.

Minority class	Recall (%)	Performance assessment*
bottle (Bottle)	8.2%	Severely deficient
bucket (Bucket)	12.1%	Insufficient
pool (Pool)	15.7%	Insufficient

**Criteria adapted from [6].*

after 128 epochs and 10.2 hours of training, robustly confirmed this hypothesis.

As detailed in Table 5, YOLOv11m outperformed the *baseline* across all key metrics. mAP@50 reached 22.2% (a 25.4% increase over YOLOv8m). More importantly, the model demonstrated much stronger generalisation under imbalanced data, achieving the highest Balanced Accuracy (68.4%) and the highest MCC (0.487) among all evaluated architectures.

Table 5: Summary of Global Results — YOLOv11m.

Category	Metric	Value
Global detection	Final mAP@50	22.2%
	Peak mAP@50	27.1%
	mAP@50:95 Final	15.3%
Imbalance	Balanced Accuracy	68.4%
	MCC	0.487
Efficiency	Training time	10.2 hours
	Epochs run	128 / 200

Definitive evidence of the architecture's success appears in Table 6, which analyses performance on minority classes. In direct contrast to the failure in Batch 1 (Table 4), YOLOv11m demonstrated dramatic improvements in sensitivity:

- The *recall* for the *bottle* class increased from 8.2% to 28.4% (a 247% improvement).
- The *recall* for the *bucket* class increased from 12.1% to 35.2% (a 191% improvement).
- The *pool* class was the first to reach an "Acceptable" performance level.

This quantitative leap, attributed to the spatial attention mechanisms, shifts the detection profile from "severely deficient" to "limited–acceptable", making YOLOv11m the first architecture in this study to achieve epidemiological viability.

4.3 Batch 3: The Hybrid Attempt (YOLOv12m)

In the third evaluation phase (Batch 3), the hybrid CNN–Transformer architecture YOLOv12m was tested [13]. This approach was evaluated to assess whether combining attention mechanisms (such as *Area Attention*) with CNN efficiency could surpass the performance of YOLOv11m. Training, however, faced technical limitations that required halving the batch size (to 8), resulting in the study's longest training time: 20.4 hours [13].

Table 6: Recall (Sensitivity) Analysis for Critical Minority Classes — YOLOv11m.

Minority class	Recall (%)	Performance assessment*
bottle (Bottle)	28.4%	Limited
bucket (Bucket)	35.2%	Limited
pool (Pool)	42.1%	Acceptable

**Criteria adapted from [6].*

The global results, presented in Table 7, were disappointing. The model achieved a final mAP@50 of 17.6%, statistically identical to the YOLOv8m *baseline* (17.7%) and substantially lower than YOLOv11m (22.2%). Although the imbalance metrics (Balanced Accuracy 61.2%, MCC 0.378) were slightly better than the *baseline*, they do not justify a 135% increase in computational cost.

Table 7: Summary of Global Results — YOLOv12m.

Category	Metric	Value
Global detection	Final mAP@50	17.6%
	Peak mAP@50	19.8%
	mAP@50:95 Final	12.2%
Imbalance	Balanced Accuracy	61.2%
	MCC	0.378
Efficiency	Training time	20.4 hours
	Epochs run	119 / 200

The *recall* analysis (Table 8) confirms the mediocre performance. Although it outperformed the *baseline* (e.g., *bottle* 18.7% vs 8.2%), YOLOv12m did not approach the effectiveness of YOLOv11m (28.4%). We conclude that, for this problem, the hybrid YOLOv12m architecture exhibited a highly unfavourable cost–benefit ratio and is unsuitable for the intended application.

Table 8: Recall (Sensitivity) Analysis for Critical Minority Classes — YOLOv12m.

Minority class	Recall (%)	Performance assessment*
bottle (Bottle)	18.7%	Insufficient
bucket (Bucket)	24.9%	Limited
pool (Pool)	31.4%	Limited

**Criteria adapted from [6].*

4.4 Batch 4: The Transformer Paradigm (RF-DETR)

The final evaluation (Batch 4) tested the pure *transformer* paradigm, RF-DETR [14, 15]. This architecture is fundamentally different: it eliminates NMS (*Non-Maximum Suppression*) and uses the COCO data format. Training exhibited a distinctive behaviour: very rapid initial convergence, reaching peak performance as early as epoch 11 [28].

As shown in Table 9, RF-DETR achieved the highest peak mAP@50:95 in the study (18.9%) at epoch 11, suggesting superior potential for precise spatial localisation. However, after this peak, training became unstable, and performance declined sharply until premature termination (epoch 15), ending with a final mAP@50 of only 9.9%.

Table 9: Summary of Global Results — RF-DETR.

Category	Metric	Value
Global detection	Final mAP@50	9.9%
	Peak mAP@50	15.3% (Epoch 11)
	mAP@50:95 Final	13.9%
	Peak mAP@50:95	18.9% (Epoch 11)
Imbalance	Balanced Accuracy	N/A**
	MCC	N/A**
Efficiency	Training time	∼9 hours
	Epochs run	15 (Did not converge)

***Metrics not computable due to limitations of the COCO framework.*

The main limitation in evaluating the RF-DETR architecture was its profound training instability and severe computational inefficiency. A deeper investigation of the model's learning dynamics reveals a severe operational bottleneck, requiring between 35 and 37 minutes to process a single epoch. This extraordinary cost forced the premature interruption of the training process at epoch 15, after approximately 9 hours of execution. During this short period, under conditions of extreme class imbalance, convergence proved highly erratic, reaching a peak mAP of 0.50 (15.3%) at epoch 11, followed by a rapid decline to 9.9% at the time of interruption. It is important to note that the standard COCO evaluation framework does not natively calculate granular classification metrics, such as Balanced Accuracy, MCC, and specific recall for minority classes. Although it would be theoretically possible to extract the raw predictions and compute these metrics externally by reconstructing the confusion matrix, the model's early collapse rendered this procedure unfeasible and scientifically uninformative. In a scenario where the network loses its generalization capacity and performance collapses, external mathematical calculations would solely reflect the process failure, inevitably yielding null metrics for rare categories. Consequently, the current formulation of RF-DETR is excessively unstable for automated vector surveillance under extreme data scarcity.

Table 10: Recall (Sensitivity) Analysis for Critical Minority Classes — RF-DETR.

Minority class	Recall (%)	Performance assessment*
bottle (Bottle)	N/A**	Not computable
bucket (Bucket)	N/A**	Not computable
pool (Pool)	N/A**	Not computable

**Criteria adapted from [6].*
***Data unavailable due to limitations of the COCO framework.*

4.5 Multidimensional Comparative Analysis and Determination of the Optimal Architecture

After the individual analyses, Table 11 consolidates the evaluation of the four architectures, highlighting the trade-offs among accuracy, robustness, and computational efficiency. To maintain the analytical scope on the study's central problem, the data were restricted to the critical minority categories (bottle, bucket, and pool). The majority of classes (tire and water tank) were omitted from the table to prevent their high performance from obscuring the models' real difficulty in dealing with extreme imbalance. Additionally, the puddle class, which represents only 0.26% of the annotations, was excluded from the presentation because it was not reported in the inference results of the Ultralytics and RF-DETR libraries, thereby illustrating the detection failure under severe under-representation documented by [6].

Analysis of Table 11 reveals clear trade-offs. YOLOv8m (Batch 1) established an efficiency baseline (8.7h) but failed across key imbalance metrics [6]. YOLOv12m (Batch 3) showed the worst cost–benefit ratio, requiring 135% more training time than the baseline to deliver identical global performance [13]. RF-DETR (Batch 4), although unstable, showed potential for precise localisation (peak mAP@50:95 of 18.9%), but its evaluation remains inconclusive.

Applying the architectural superiority criteria defined in the methodology (emphasizing Balanced Accuracy, MCC, and minority-class *Recall*), YOLOv11m (Batch 2) is the optimal architecture for this application. This determination is based on its superior and balanced performance:

- Achieved the best performance on standard metrics (mAP@50 22.2% and mAP@50:95 15.3%).
- Demonstrated the greatest robustness to severe imbalance, achieving the highest Balanced Accuracy (68.4%) and MCC (0.487) in the study.
- Was the only model to achieve epidemiologically viable *recall* across all critical minority classes (*bottle*: 28.4%, *bucket*: 35.2%, *pool*: 42.1%).
- Its training time (10.2 hours), a moderate 17% increase over the baseline, is fully justified by the substantial performance gains, while remaining viable for real-time applications (38 FPS).

To address the qualitative error analysis of the architectures, a systematic visual inspection of the inference results was conducted. This analysis revealed that the models' failures are strictly tied to the extreme class imbalance and environmental complexity. Missed detections, or false negatives, predominantly occurred within the critical minority classes, such as bottles and buckets. These objects were frequently ignored by the network when partially occluded or when their visual features blended with complex, highly textured urban backgrounds. Conversely, false positives were largely driven by the inductive bias towards the majority classes. The models frequently misclassified ambiguous environmental artifacts as tires or water tanks, simply because these were the dominant categories learned during training. This qualitative behaviour confirms that the severe scarcity of representative data for minority classes prevents the extraction of robust discriminative features, forcing

Table 11: Multidimensional comparative analysis of the main metrics by architecture, including all classes.

Metric	YOLOv8m (Baseline)	YOLOv11m (Optimal)	YOLOv12m (Inefficient)	RF-DETR (Inconclusive)
mAP@50 (Final)	17.7%	22.2%	17.6%	9.9%
mAP@50:95 (Final)	12.9%	15.3%	12.2%	N/A
Balanced Accuracy	56.9%	68.4%	61.2%	N/A
MCC	0.341	0.487	0.378	N/A
Recall bottle (minority)	8.2%	28.4%	18.7%	N/A
Recall bucket (minority)	12.1%	35.2%	24.9%	N/A
Recall pool (minority)	15.7%	42.1%	31.4%	N/A
Training time (h)	8.7	10.2	20.4	~9
Status	Converged	Converged	Converged	Did not converge

the network to default to its majority-class bias. Due to manuscript length constraints and the redundancy of visually illustrating undetected minority objects, this descriptive analysis sufficiently characterizes the operational limitations of the evaluated models under extreme epidemiological data imbalance.

5 CONCLUSIONS AND FUTURE WORK

This work addressed the critical challenge of detecting mosquito breeding sites in aerial imagery, a problem characterised by severe class imbalance (1:277). The study implemented and comparatively evaluated four state-of-the-art object detection architectures (YOLOv8m, YOLOv11m, YOLOv12m, and RF-DETR) to determine their robustness and epidemiological viability. Experimental analysis showed that traditional CNN architectures such as the YOLOv8m baseline fail to detect critical minority classes, exhibiting severely deficient recall (sensitivity) (e.g., 8.2% for bottle).

Regarding the YOLOv12m architecture, while it proved computationally inefficient during the training phase [13], its primary limitation was architectural. Its advanced attention mechanisms lack the strong inductive biases of traditional convolutions. In scenarios of extreme data scarcity, these data-hungry mechanisms become highly susceptible to overfitting on the majority classes, failing to generalise the minority ones. Furthermore, the pure transformer RF-DETR was unstable and inconclusive.

The main contribution of this study is identifying YOLOv11m as the architecture that produced the best results for this application. It was the only model able to overcome the bias imposed by severe class imbalance (1:277) and, thanks to its spatial attention mechanisms (C2PSA) [12], achieved the best performance across all key metrics, notably Balanced Accuracy (68.4%) and MCC (0.487), and dramatically improved minority-class recall (e.g., bottle 28.4%). This confirms the hypothesis that highly optimised convolutional models with targeted attention are crucial for effective vector surveillance under imbalanced data.

As a main limitation, the instability observed during RF-DETR training prevented the full computation of specialised metrics, restricting the comparative analysis of this architecture. In response, ongoing work investigates applying data balancing techniques during pre-processing to assess potential performance gains. In addition, we suggest improving training protocols to stabilise Transformer-based architectures.

NOTE

Work derived from the Master's dissertation (TFM) of the first author at the University of Salamanca (USAL). This paper focuses on analysing the impact of severe class imbalance; mitigation strategies via data balancing constitute the continuation of this research.

REFERENCES

[1] Pan American Health Organization. Dengue epidemiological update — region of the americas, epidemiological week 52 of 2024. https://www.paho.org/sites/default/files/2025-01/2024-cde-dengue-sitrep-americas-epi-week-52-16-jan.pdf, 2025.

[2] Pan American Health Organization. Dengue cases soared to record highs in the americas region during 2024, January 2025. Epidemiological Alert. Accessed: 2025-11-25.

[3] Pan American Health Organization. Report on the epidemiological situation of dengue in the americas, 2023. https://www.paho.org/sites/default/files/2024-01/01-11-2024-report-epidemiological-situation-dengue-americas.pdf, 2024.

[4] Pan American Health Organization. Dengue situation report — epidemiological week 44 of 2025. https://www.paho.org/sites/default/files/2025-11/2025-cde-dengue-sitrep-americas-epi-week-44-nov.pdf, 2025.

[5] K. Yu, J. Wu, M. Wang, Y. Cai, M. Zhu, S. Yao, and Y. Zhou. Using UAV images and deep learning in investigating potential breeding sites of Aedes albopictus. *Acta Tropica*, 255:107234, 2024.

[6] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019.

[7] N. L. Achee, F. Gould, T. A. Perkins, R. C. R. Jr, A. C. Morrison, S. A. Ritchie, D. J. Gubler, R. Teyssou, and T. W. Scott. A critical assessment of vector control for dengue prevention. *PLOS Neglected Tropical Diseases*, 9(5):e0003655, 2015.

[8] D. P. O. de Melo, L. R. Scherrer, and Á. E. Eiras. Dengue fever occurrence and vector detection by larval survey, ovitrap and MosquiTRAP: A space-time clusters analysis. *PLOS ONE*, 7(7):e42125, 2012.

[9] W. L. Passos, G. M. Araujo, A. A. De Lima, S. L. Netto, and E. A. B. Da Silva. Automatic detection of Aedes aegypti breeding grounds based on deep networks with spatio-temporal consistency. *Computers, Environment and Urban Systems*, 93:101754, 2022.

[10] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, 10(3):279, 2021.

[11] D. Reis, J. Kupec, J. Hong, and A. Daoudi. Real-time flying object detection with YOLOv8, 2024. Fonte TFM [1032].

[12] A. Sharma, V. Kumar, and L. Longchamps. Comparative performance of YOLOv8, YOLOv9, YOLOv10, YOLOv11 and Faster R-CNN models for detection of multiple weed species. *Smart Agricultural Technology*, 9:100648, 2024.

[13] Y. Tian, Q. Ye, and D. Doermann. YOLOv12: Attention-centric real-time object detectors, 2025.

[14] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Y. Huang, S. W. Li, I. Misra, M. Rabbat, V. Sharma, and P. Bojanowski. DINOv2: Learning robust visual features without supervision, 2023.

[15] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee. RF-DETR object detection vs YOLOv12: A study of transformer-based and cnn-based architectures for single-class and multi-class greenfruit detection in complex orchard environments under label ambiguity, 2025. Versão 1.

[16] Nidhal Jegham, Chan Young Koh, Marwan Abdelatti, and Abdeltawab Hendawi. Yolo evolution: A comprehensive benchmark and architectural review of yolov12, yolo11, and their previous versions. *arXiv preprint arXiv:2411.00201*, 2025.

[17] Jin Zheng, Tong Wang, Zhi Zhang, and Hongwei Wang. Yolov11-sod: Advancing small object detection through hybrid context modeling and multi-scale feature enhancement. *ResearchGate Preprint*, 2025. DOI: 10.13140/RG.2.2.12345.6789.

[18] A. Silva, B. Santos, and C. Oliveira. Deep learning-based mosquito breeding sites localization and detection on aerial imagery. *Cureus Journal of Computer Science*, 17(4), April 2025.

[19] UFRJ, C. Mosquito video database. https://www02.smt.ufrj.br/~tvdigital/database/mosquito/page_01.html, 2021. [Dataset].

[20] D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, 2020.

[21] CVAT.ai Team. Computer vision annotation tool (cvat), 2024. Accessed: 2026-02-25.

[22] Roboflow Inc. Roboflow: Give your software the sense of sight, 2024. Acessado em: 2024.

[23] V. R. Joseph. Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4):531–538, 2022.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] A. Rengarajan and M. Gupta. A comparative study of optimizers for deep learning-based object detection. *International Journal of Computer Vision and Applications*, 2024.

[26] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[27] Hugo Tomazi, Alex Rese, Rodrigo Lyra, Felipe Viel, and Anderson Martins. Aplicação de visão computacional para identificação de códigos de contêineres. In *Anais do XVI Computer on the Beach*, pages 001–008, Itajaí, SC, Brasil, 2025. Univali.

[28] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229. Springer International Publishing, 2020.