REVISTA OWL (*OWL Journal* )
www.revistaowl.com.br – ISSN: 2965-2634

# MOSAICOS SINTÉTICOS COM CONTEXTO CONTROLADO ALTERAM O PONTO DE OPERAÇÃO SEM MELHORAR A DISCRIMINAÇÃO EXTERNA NA CLASSIFICAÇÃO DE BACILOSCOPIAS PAUCIBACILARES

## *SYNTHETIC MOSAICS WITH CONTROLLED CONTEXT SHIFT THE OPERATING POINT WITHOUT IMPROVING EXTERNAL DISCRIMINATION IN PAUCIBACILLARY SMEAR CLASSIFICATION*

DOI: 10.5281/zenodo.23968001

*Fabio Cavalcante Binatti*[1]
*Thiago Felipe de Carvalho Lourenço*[2]
*Rayner do Nascimento Souza*[3]
*Monalise Daiane dos Santos Ferreira*[4]
*Rodrigo Rodrigues Antunes*[5]

1 Master's degree in Optimization and Processes from the Galileo Institute of Technology and Education of the Amazon (ITEGAM); PhD candidate in Electrical Engineering at the Federal University of Amazonas (UFAM). Researcher at the Sustainable and Technological Institute of the Amazon (ISATECH), Manaus, Amazonas, Brazil. Corresponding author. E-mail: fabio.binatti@isatech.org.br. ORCID: https://orcid.org/0009-0000-5891-0187.

2 Postgraduate specialization in Software Quality Engineering from the National Service for Commercial Training (Senac); Bachelor's degree in Electrical Production Engineering from FUCAPI University Center. Researcher at the Sustainable and Technological Institute of the Amazon (ISATECH), Manaus, Amazonas, Brazil. E-mail: thiago.lourenco@isatech.org.br. ORCID: https://orcid.org/0009-0008-0093-9209.

3 Master's degree in Intellectual Property from Amazonas State University (UEA); PhD candidate in Biotechnology at the Federal University of Amazonas (UFAM). Researcher at the Sustainable and Technological Institute of the Amazon (ISATECH), Manaus, Amazonas, Brazil. E-mail: rayner.souza@isatech.org.br. ORCID: https://orcid.org/0000-0002-8336-3311.

4 Postgraduate specialization in Microbiology and Immunology from the Metropolitan College of Manaus (FAMETRO); Bachelor's degree in Biomedicine from the same institution. Researcher at the Sustainable and Technological Institute of the Amazon (ISATECH), Manaus, Amazonas, Brazil. E-mail: monalise.ferreira@isatech.org.br. ORCID: https://orcid.org/0009-0000-8491-2338.

5 PhD in Applied Geosciences from the University of Brasília (UnB), with postdoctoral research at the Institute of Geosciences of the University of Brasília (2024 to 2026). Researcher at the Sustainable and Technological



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

REVISTA OWL (*OWL Journal* )
www.revistaowl.com.br – ISSN: 2965-2634

**RESUMO**

Redes neurais treinadas com imagens sintéticas são uma saída para a escassez de lâminas paucibacilares anotadas na baciloscopia automatizada da tuberculose, mas pouco se sabe sobre como a forma de construir essas imagens se reflete em lâminas reais. Este estudo comparou mosaicos montados em grade rígida com mosaicos de contexto controlado, nos quais imagens positivas e negativas compartilham o mesmo fundo para que a rede aprenda o bacilo, e não o cenário. Foram treinadas 75 redes (cinco condições, cinco arquiteturas e três sementes), com limiar de decisão fixado ainda nos dados sintéticos, e testadas em duas bases externas: Base externa 1 (Wellgen; 100 imagens positivas e 900 negativas, microscópio digital) e Makerere (1.217 positivas e 37 negativas, câmera de celular). Pela métrica usual, a área sob a curva ROC (AUROC), o contexto controlado não fez diferença (Δ +0,012 e −0,022; p de Holm = 1,0). No limiar definido durante o treino, porém, o comportamento das redes mudou marcadamente: na Base externa 1, a especificidade subiu de 0,419 para 0,778, enquanto a sensibilidade caiu de 0,997 para 0,747, padrão mantido após excluir as execuções que não convergiram. A sensibilidade também dependeu do tamanho aparente do bacilo, que varia com o equipamento de aquisição, e até um modelo treinado só com imagens reais perdeu o equilíbrio entre sensibilidade e especificidade ao mudar de base. Esses resultados mostram que a AUROC, isoladamente, pode esconder mudanças clinicamente relevantes, e que a validação de sistemas automatizados deve relatar o desempenho no limiar congelado e a escala de aquisição das imagens.

**Palavras-chave:** Tuberculose; Baciloscopia; Ziehl-Neelsen; Aprendizado profundo; Dados sintéticos; Validação externa.

**ABSTRACT**

Neural networks trained on synthetic images offer a way to address the scarcity of annotated paucibacillary smear images in automated tuberculosis microscopy, but little is known about how the construction of these images affects performance on real microscopy fields. This study compared mosaics assembled on a rigid grid with controlled-context mosaics, in which positive and negative examples are generated using the same background procedure to encourage the network to focus on bacilli rather than contextual cues. We trained 75 networks (five conditions, five architectures, and three seeds), with the decision threshold fixed using synthetic validation data, and tested them on two external datasets: External dataset 1 (Wellgen; 100 positive and 900 negative images, digital microscope) and Makerere (1,217 positive and 37 negative images, mobile-phone camera). Controlled context did not yield a statistically significant improvement in AUROC on either external dataset (Δ = +0.012 and −0.022, respectively; Holm-adjusted p = 1.0 for both comparisons). At the frozen threshold, however, network behavior changed markedly: in External dataset 1, specificity increased from 0.419 to 0.778, whereas sensitivity decreased from 0.997 to 0.747, a pattern that persisted after excluding runs that failed to converge. Sensitivity also depended on the apparent size of bacilli, which varies with the acquisition device, and even a model trained exclusively on real images lost its balance between sensitivity and specificity when transferred to another dataset. These findings demonstrate

Institute of the Amazon (ISATECH), Manaus, Amazonas, Brazil. E-mail: rodrigo.antunes@isatech.org.br. ORCID: https://orcid.org/0000-0001-9368-5077.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

that AUROC alone can mask clinically relevant changes and that the validation of automated systems should report performance at the frozen threshold and the image acquisition scale.

**Keywords:** Tuberculosis; Sputum smear microscopy; Ziehl-Neelsen; Deep learning; Synthetic data; External validation.

## 1. INTRODUCTION

Tuberculosis, caused by *Mycobacterium tuberculosis*, remains one of the leading causes of death from a single infectious agent. In 2023, global incidence was estimated at 134 cases per 100,000 population, with approximately 1.25 million deaths; in 2024, 8.3 million cases were notified, while about 2.4 million people remained undiagnosed (WORLD HEALTH ORGANIZATION, 2024, 2025). More than 80% of cases and deaths occur in low- and middle-income countries (WORLD HEALTH ORGANIZATION, 2026). In Brazil, the North Region recorded the highest regional incidence in 2025; in the state of Amazonas, the rate reached 93 cases per 100,000 population, and the state had the highest tuberculosis mortality in 2024 (BRASIL, 2026), in a region where long distances and a dispersed laboratory network make low-cost diagnostic methods that can operate with minimal infrastructure essential.

Sputum smear microscopy with Ziehl-Neelsen (ZN) staining remains a low-cost alternative in resource-limited settings. Its sensitivity, however, is affected by the examiner's experience, staining quality, fatigue, and the need to inspect numerous microscopic fields (CARVALHO BRITO et al., 2026; SERRÃO et al., 2024; AULIA et al., 2024; XU et al., 2025). In paucibacillary samples, few bacilli are scattered among many negative regions, and an automated system must simultaneously avoid missing the rare fields containing bacilli and avoid generating false alarms in the numerous negative fields, because even a low per-field false-positive rate can make the slide-level decision unworkable (SERRÃO et al., 2024; SANTIAGO-MOZOS et al., 2014). Although culture and molecular methods offer diagnostic advantages, they depend on time or specialized infrastructure (LUO et al., 2026; MOTA CARVALHO et al., 2023); in this setting, automated methods can support the standardization of slide reading.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

Advances in deep learning have driven the automated detection of bacilli, but the published evidence still favors testing in the same domain used for development. Reviews and experimental studies report sensitivities and specificities frequently above 90%, with promising results for classification and detection architectures (MOTA CARVALHO et al., 2023; TAMURA et al., 2024; YAN et al., 2025; CHEN et al., 2024; WANG; PAN; SHU, 2024; ZHENG; ZHANG; LIU, 2026); however, random splits of a single collection, restricted availability of datasets, and heterogeneous metrics limit comparability (MOTA CARVALHO et al., 2023; TAMURA et al., 2024). When external testing is performed, the picture changes. In the Nightingale challenge, a model maintained an area under the ROC curve (AUROC) of 0.995 on the external set but lost specificity and concentrated false negatives in images with few bacilli (BEDŐHÁZI et al., 2025).

In a study on automated detection of malaria parasites, models trained on high-quality images reached an accuracy of up to 97.98% on the source set but showed an accuracy close to 50% when evaluated on lower-quality images from Sudan, highlighting the difficulty of generalizing across acquisition conditions (ZHONG et al., 2023). In chest radiography, a commercial software with an AUROC of 0.85 did not reach the minimum sensitivity and specificity at the manufacturer's thresholds (TAVAZIVA et al., 2022). Similar patterns in computed tomography and microscopy (NIJIATI et al., 2023; VISUÑA et al., 2025), together with the influence of staining, resolution, and acquisition geometry in smear microscopy collections (GREESHMA; VISHNUKUMAR, 2026), suggest that generalization failure lies less in the ability to rank images than in the operating point chosen in the development domain.

One alternative to the scarcity of annotations is to expand training with synthetic data. Cut-and-paste and *domain randomization* approaches seek to diversify the appearance of the material and to narrow the gap between simulation and real acquisition (DWIBEDI; MISRA; HEBERT, 2017; TOBIN et al., 2017; TREMBLAY et al., 2018). In this study, the term “synthetic data” refers to mosaics built from patches of real microscopy images, rearranged and combined through compositing and pasting procedures. This approach differs from the



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

full generation of images by generative models and allows the visual characteristics of the background and of the bacilli to be controlled. In the literature examined in this study, no external evaluation of smear-microscopy classifiers trained exclusively on synthetic mosaics was identified among the studies reviewed. Shortcut learning offers a plausible explanation for part of the loss of transportability: the model may rely on cues that are associated with the classes during training but are irrelevant to the presence of bacilli (GEIRHOS et al., 2020; ZECH et al., 2018). In rigid-grid mosaics, differences in color, background texture, and seams between patches can act as spurious cues associated with the class. In Ziehl-Neelsen-stained smear microscopy, crystallized deposits of red stain can cause errors in automated segmentation, whereas artifacts and overstaining hinder the identification of bacilli (KHUTLANG et al., 2010; SHAH et al., 2017). Because these factors can modify the score distribution without changing its ranking, AUROC may appear stable while the threshold defined in the synthetic domain ceases to work on real images.

This study compared two strategies for generating synthetic mosaics for paucibacillary smear microscopy classifiers: a rigid-grid generator and a controlled-context generator, designed on the basis of a prior audit of the rigid-grid shortcuts, in which the backgrounds of both classes follow the same process and the presence of the bacillus is the main difference.

The pre-specified hypothesis was a gain in image-level AUROC on two external datasets of different origin, staining, and acquisition. Five neural network architectures and three random seeds were used as measurement instruments rather than as objects of comparison. As a secondary analysis, the dependence of sensitivity on image scale was evaluated; after the primary result had been observed, performance at the frozen threshold and a real-image reference were examined.

## 2. MATERIALS AND METHODS

### 2.1 Study design and outcomes

The networks in the main analyses were trained only on synthetic mosaics and evaluated, without any adjustment, on three sets of real whole-field images: an internal set

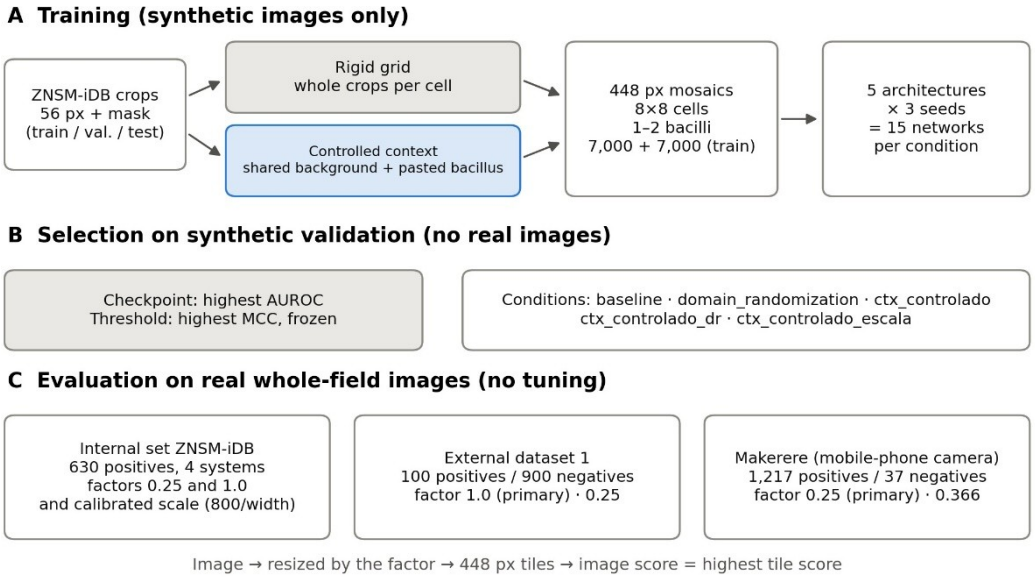


*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

and two external datasets (Figure 1). Each experimental condition was trained with five architectures and three seeds, with identical hyperparameters, and reporting follows the TRIPOD+AI guideline where applicable (COLLINS et al., 2024).

Outcomes were ranked before the comparison between conditions. The primary outcome was image-level AUROC on External dataset 1 and on Makerere, each at its pre-specified scale factor. Secondary outcomes were sensitivity on Makerere and sensitivity on the internal set at the calibrated scale, and AUROC or sensitivity at inappropriate scale factors were robustness outcomes. The following were analyzed after the primary outcome had been observed (post hoc): sensitivity, specificity, and Matthews correlation coefficient (MCC) at the frozen threshold; sensitivity by bacillary load; the real-image reference; and the sensitivity analysis excluding the non-converged runs.

Figure 1 – Study design




Source: Prepared by the authors (2026). (A) Training only on synthetic mosaics generated by rigid grid or by controlled context, with 15 networks per condition. (B) Checkpoint and threshold selected on synthetic validation and frozen. (C) Evaluation without adjustment on three sets of real whole-field images.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

### 2.2 Datasets

The mosaics were built from ZNSM-iDB (SHAH et al., 2017), a public dataset of ZN-stained sputum fields digitized with three microscopy systems using a 100× objective (SHAH et al., 2017; NAGARAJU et al., 2025). The field images were first split into training, validation, and test sets, and a portion was reserved as the internal set; only then were patches extracted, so that patches from the same image never appear in different partitions. Because the public dataset provides no slide identifier, slide-level independence could not be verified. Positive patches were centered on bacilli manually segmented in CVAT, in which only isolated, good-quality bacilli were annotated; negative patches were extracted from unannotated regions. The partitions contain 3,002 positive and 12,663 negative patches in training, 1,544 and 5,376 in validation, and 1,326 and 5,581 in testing. Table 1 summarizes the real evaluation sets, none of which was used for training, model selection, or threshold setting.

Table 1 – Real image sets used for evaluation

| Set | Role | Positives / negatives | Acquisition | Scale factor |
|---|---|---|---|---|
| Internal (ZNSM-iDB) | Transfer from mosaic to whole field | 630 / 0 | Microscope, ZN; 800 × 600 to 3984 × 2988 px | min(1; 800/width) |
| External dataset 1 (Wellgen) | Primary outcome | 100 / 900 | Digital microscope, 450×, 0.1725 µm/pixel; modified Kinyoun | 1.0 (native) |
| Makerere | Primary outcome | 1,217 / 37 | Mobile-phone camera attached to the microscope, ZN; Uganda | 0.25 |

Source: Prepared by the authors (2026). Note: the scale factor brings the apparent size of the bacillus close to that in the training mosaics (Section 2.5).

The internal set contains 630 positive images from ZNSM-iDB, set aside before patch extraction but originating from the same slides and systems; it measures the transfer from mosaic to whole field, not generalization across laboratories. Fourteen of these images are identical, by SHA256, to files labeled as without bacilli in the dataset itself; because it was



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

not possible to determine which label is correct, they were kept, and sensitivity was recalculated without them (Supplementary Material, Table S1).

External dataset 1 is the external set released by Wellgen Medical (LIN; WELLGEN MEDICAL, 2024) and used as the external test in the Nightingale challenge study, whose data availability statement points to the same identifier (BEDŐHÁZI et al., 2025). The slides come from an Asian medical center distinct from the development set of that challenge and were stained with the modified Kinyoun method; positives come from culture-positive cases, with bacilli identified by an algorithm and checked by two technologists (BEDŐHÁZI et al., 2025; LIN; WELLGEN MEDICAL, 2024). The criterion for negatives is not described for the external set.

Makerere comprises 1,265 ZN images acquired in Uganda with a mobile-phone camera (WANG; PAN; SHU, 2024; QUINN et al., 2016); according to the annotations, 1,218 have at least one annotated bacillus and 47 have none, which characterizes images without annotated bacilli rather than confirmed absence of bacilli. The vignetting filter (Section 2.5), with parameters fixed before the external evaluation, excluded 11 images with 26% to 100% dark pixels, four of them completely dark: one positive and ten negatives (Supplementary Material, Table S2). This left 37 evaluable negatives. The number of annotated bacilli defined load strata of 1 (n = 70), 2 to 3 (n = 230), 4 to 7 (n = 374), and 8 or more (n = 543).

### 2.3 Mosaic generators

All mosaics measure 448 × 448 pixels, on an 8 × 8 grid of 56-pixel cells; a positive mosaic contains one or two bacilli, a range chosen to represent paucibacillary slides, and a negative mosaic contains none. Each generator produced, with seed 42, 7,000 positive and 7,000 negative mosaics for training, in addition to 700 mosaics of each class for validation and 700 of each class for the synthetic test (Figure 2).

In the rigid-grid generator, each cell receives a whole patch, with its own rotation, reflection, and brightness and contrast adjustment, and the patches are juxtaposed without smoothing. Before the new generator was designed, simple classifiers (logistic regression and

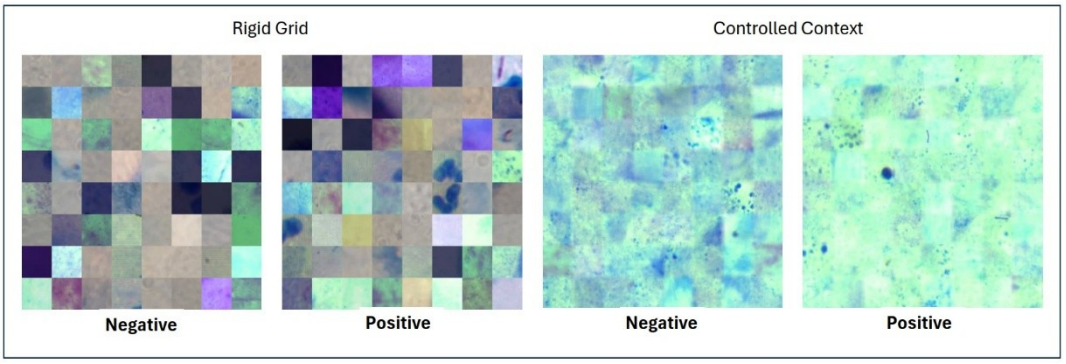


*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

gradient boosting) were trained on 11 color and texture features of the training patches and tested on the validation and test patches (Supplementary Material, Table S3). The color cluster of the patch predicted the class with an AUROC of 0.818 (95% CI 0.811 to 0.825). With the bacillus digitally erased in both classes, the background still distinguished positives from negatives with an AUROC of 0.724 (0.713 to 0.734), even within the same color cluster. The intensity jump at the seams was 16.4 times greater than in normal texture, and 112-pixel patches downscaled to 56 were more frequent among negatives (37%) than among positives (22%).

The controlled-context generator was designed to neutralize these cues: (i) it uses only 56-pixel patches; (ii) it clusters the patches by color with the k-means algorithm (k = 24, fixed a priori and fitted only to the training negatives); (iii) it draws a single color cluster per mosaic, with equal probability for both classes; (iv) it always assembles the 64 cells from negative patches and, in positives, pastes only the bacillus, cut out by the segmentation mask, at least 6 pixels from the cell border; (v) it applies two pastes to every mosaic, those that are not bacilli being sham background pastes, so that paste borders appear equally in both classes; (vi) it joins the cells with a 6-pixel soft overlap; and (vii) it applies only rotation and reflection to the cells. After generation, an audit checked whether the background surrounding the pasted bacillus still distinguished it from a sham paste, with the criterion of revising the generator if AUROC exceeded 0.6.

**Figure 2 –** Examples of training mosaics from the two generators




Source: Prepared by the authors (2026).



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

### 2.4 Experimental conditions and training

The five conditions combined a generator and a training data augmentation technique (Table 2). The main contrast compares two generation packages rather than an isolated component. All conditions had 14,000 training mosaics and 1,400 validation mosaics, produced by the respective generator.

**Table 2 –** Experimental conditions

| Condition | Mosaic | Training augmentation | Compared with |
|---|---|---|---|
| baseline | Rigid grid | Basic | Reference |
| domain_randomization | Rigid grid | Domain randomization | baseline |
| ctx_controlado | Controlled context | Basic | baseline (main comparison) |
| ctx_controlado_dr | Controlled context | Domain randomization | ctx_controlado |
| ctx_controlado_escala | Controlled context | Basic + random scaling | ctx_controlado |

Source: Prepared by the authors (2026). Note: each condition was trained with five architectures and three seeds (15 networks).

Basic augmentation applies reflections, rotations, slight color variation, and blur. Domain randomization (TOBIN et al., 2017) broadens color variation and adds geometric, sharpness, compression, and noise distortions, with color augmentation preferred over stain normalization (TELLEZ et al., 2019; SHEN et al., 2022). Random scaling resizes 75% of the training mosaics by factors between 0.5 and 2.0, which takes the bacillus from about 22 to 11 to 44 pixels. Detailed parameters are given in the Supplementary Material (M1).

The architectures used were ResNet-50 (HE et al., 2016), EfficientNet-B3 (TAN; LE, 2019), ConvNeXt-Tiny (LIU et al., 2022), ViT-B/16 (DOSOVITSKIY et al., 2021), and Swin-T (LIU et al., 2021), with ImageNet-pretrained weights from the timm library (WIGHTMAN, 2019), a single output, and a 448 × 448-pixel input. Training used the AdamW optimizer (LOSHCHILOV; HUTTER, 2019), with a learning rate and weight decay of $1 \times 10^{-4}$, batches of 8 mosaics, up to 20 epochs, and early stopping after 5 epochs without improvement, with seeds 42, 123, and 2024. The selected checkpoint was the one with the highest AUROC on synthetic validation, and the decision threshold was the one that maximized MCC (CHICCO; JURMAN, 2020) on that same validation set. The threshold was then frozen and applied unchanged to all real images. Per-architecture ensembles and an ensemble of the 15 networks



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

of each condition were pre-specified, with the ensemble score defined as the mean of the networks' probabilities.

**2.5 Whole-field inference and scale**

Each real image was resized by the scale factor and divided into 448 × 448-pixel tiles. Tiles with more than 15% dark pixels were discarded as vignetting, and the image score was the highest probability among the valid tiles, which is equivalent to considering an image positive if any region is suspicious. The number of tiles was nearly constant within each dataset: 30 in 997 of the 1,000 images of Dataset 1, exactly 1 in Makerere, and 1 to 4 in the internal set at the calibrated scale. Because the threshold is chosen on single-tile mosaics and applied to the maximum over several tiles, it is not adjusted to the number of tiles in the real image.

The scale factor of each dataset was set before evaluation to bring the apparent size of the bacillus close to that in the mosaics. In the training patches, the side of the bacillus bounding box had a median of 22 pixels (interquartile range 17 to 30). In Makerere, the longest side of the annotated boxes had a median of 80 pixels, and the factor 0.25 brings the bacillus to about 20 pixels; this measurement uses annotations from the evaluated set itself, not model predictions. Dataset 1 was evaluated at native scale, at which a bacillus of 2 to 4 µm spans about 12 to 23 pixels (BEDŐHÁZI et al., 2025). In the internal set, the calibrated scale downscaled each image by the factor 800 divided by its width, assuming the same field of view across the ZNSM-iDB systems. Inappropriate factors (0.25 in Dataset 1, 0.366 in Makerere, and 1.0 in the internal set) assessed robustness to scale.

**2.6 Statistical analysis**

AUROC, sensitivity, specificity, and MCC were computed at the image level. Except for the ensembles, results summarize 15 runs. Conditions were compared on the same images and the same architecture–seed pairs, and the effect was the mean of the paired differences. The 95% confidence intervals (95% CI) were obtained from a paired bootstrap with 2,000



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

resamples that, at each iteration, resampled the images within each class, the architectures, and, within each architecture, the seeds; the architecture was the resampling unit because seeds of the same architecture are not independent. With only five architectures, the intervals may have lower-than-nominal coverage, and the number of pairs favoring each condition is therefore also reported. The Holm correction (HOLM, 1979) was applied to the main comparison on the two primary outcomes; in the secondary and post hoc analyses, intervals and p-values are descriptive. Proportions from a single classifier received Wilson CIs.

Four runs did not converge during synthetic training (Section 3.1) and were kept in the main analysis, in accordance with the protocol. As a sensitivity analysis, defined after these runs were observed, the comparisons were repeated without the two affected architecture–seed pairs (13 pairs; Supplementary Material, Table S4).

**2.7 Real-image reference and interpretability**

To check whether the loss of the operating point is specific to synthetic data, the five architectures were also trained on whole ZNSM-iDB images, with image-level supervision and multiple-instance learning (DIETTERICH; LATHROP; LOZANO-PÉREZ, 1997), in two variants: a full variant, with 201 positive and 90 negative training images, and a restricted variant, with 17 positives and 64 negatives, which excluded any overlap with the internal set (Supplementary Material, M2). SHA256 and texture-signature audits found no overlap between the real training images and the external datasets; in the internal set, 254 images were identical by SHA256 to training images of the reference, and the texture signature indicated other probable same-scene matches, totaling 284 matches, all excluded in the restricted variant. In an exploratory analysis, Grad-CAM (SELVARAJU et al., 2017) and HiResCAM (DRAELOS; CARIN, 2020) activation maps were compared among baseline, ctx_controlado, and ctx_controlado_dr across the three seeds (M3).

The pipeline was implemented in Python with PyTorch 2.11 and timm 1.0.29, predominantly on an NVIDIA L4 GPU in Google Colab. One ResNet-50 run, with seed 42 in



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

the ctx_controlado_escala condition, used a Tesla T4 GPU. Deterministic algorithms were used, and all statistics can be recomputed from the preserved image-level predictions.

## 3. RESULTS

### 3.1 Generator audit and synthetic learning

In the controlled-context generator, the audit did not detect the rigid-grid cues: the background surrounding the pasted bacillus did not distinguish it from a sham paste (AUROC 0.511; approximate 95% CI 0.489 to 0.533), and the intensity jump at the seams fell from 16.4 times the normal texture to 1.82 times, making the seams much less conspicuous, although not invisible (Supplementary Material, Table S3). Among the 29,492 patches in the synthetic partitions, 23 (less than 0.1%) had a match in another partition, probably owing to duplicated images within the ZNSM-iDB itself, and no evaluation image matched the external datasets.

All conditions learned the synthetic task, with mean AUROC of 0.943 to 0.967 on the synthetic test set (Supplementary Material, Table S5). Four of the 75 runs, however, did not converge: ConvNeXt-Tiny with seed 2024 in the three controlled-context conditions and ViT-B/16 with seed 123 in domain_randomization. In these runs, the validation loss remained close to ln 2 (0.693), indicating nearly constant outputs; because the configuration was identical to that of the other runs, this pattern points to an optimization failure rather than a configuration error.

### 3.2 Primary outcome: AUROC on the external datasets

Controlled context did not improve per-image AUROC on either external dataset (Table 3; Figure 3). On Dataset 1, mean AUROC was 0.905 with ctx_controlado and 0.893 with baseline (Δ +0.012; 95% CI −0.071 to 0.127), with ctx_controlado outperforming baseline in 7 of the 15 pairs; on Makerere, it was 0.768 versus 0.790 (Δ −0.022; −0.091 to 0.059), with ctx_controlado outperforming baseline in 8 of the 15 pairs despite the slightly lower mean. After Holm correction, p = 1.0 for both comparisons. Excluding the pairs with



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

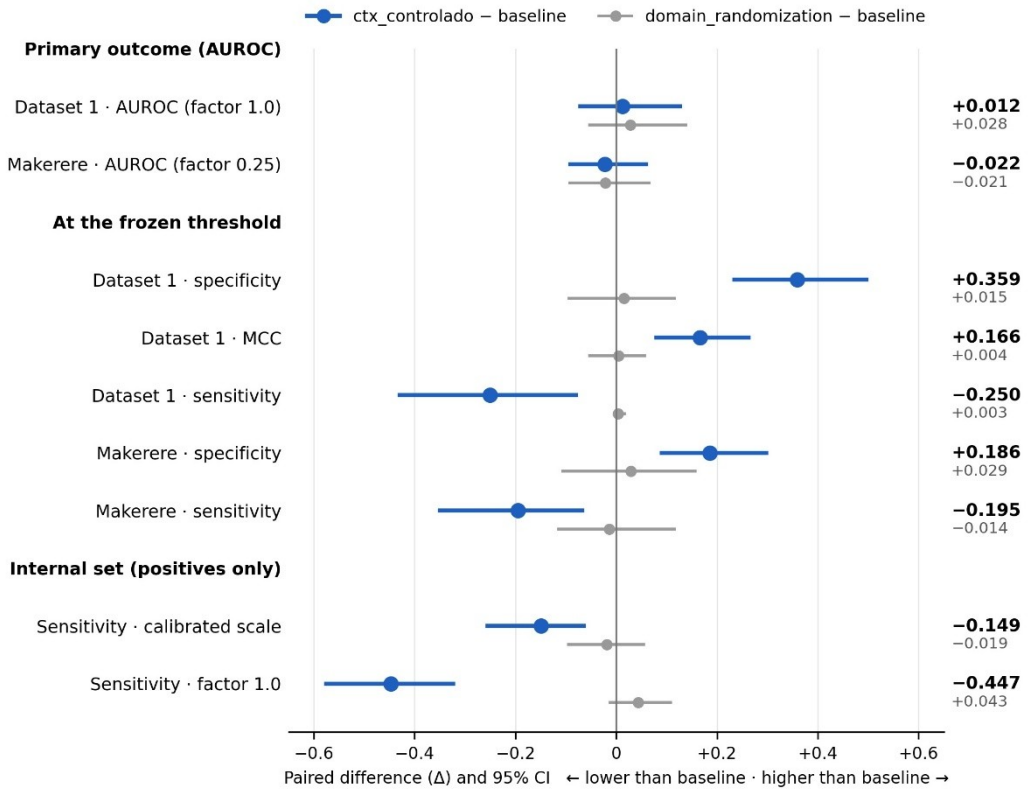
non-converged runs, the differences were +0.024 (−0.068 to 0.132) and −0.011 (−0.074 to 0.059). Domain randomization applied to the rigid grid also did not change AUROC.

**Table 3 –** Performance on the external datasets (mean of 15 runs)

| Condition | D1 AUROC | D1 Sens. | D1 Spec. | D1 MCC | Mak AUROC | Mak Sens. | Mak Spec. | Mak MCC |
|---|---|---|---|---|---|---|---|---|
| baseline | 0.893 | 0.997 | 0.419 | 0.266 | 0.790 | 0.736 | 0.638 | 0.193 |
| domain_randomization | 0.921 | 1.000 | 0.435 | 0.270 | 0.769 | 0.722 | 0.667 | 0.171 |
| ctx_controlado | 0.905 | 0.747 | 0.778 | 0.432 | 0.768 | 0.541 | 0.823 | 0.156 |
| ctx_controlado_dr | 0.908 | 0.919 | 0.658 | 0.401 | 0.813 | 0.729 | 0.721 | 0.194 |
| ctx_controlado_escala | 0.800 | 0.439 | 0.809 | 0.211 | 0.786 | 0.538 | 0.843 | 0.158 |

Source: Prepared by the authors (2026). Note: D1, External dataset 1 (100 positives and 900 negatives; factor 1.0); Mak, Makerere (1,217 positives and 37 negatives; factor 0.25). AUROC is the primary outcome; sensitivity (Sens.), specificity (Spec.), and MCC at the frozen threshold are post hoc analyses.

**Figure 3 –** Paired differences relative to baseline




Source: Prepared by the authors (2026). Note: mean of 15 architecture–seed pairs, with 95% CI from the paired bootstrap. Blue, ctx_controlado − baseline; gray, domain_randomization − baseline.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

### 3.3 Operating point at the frozen threshold

The difference between the generators emerged at the threshold rather than in the ranking of images (Figure 4). On Dataset 1, baseline classified almost all images as positive (sensitivity of 0.997 and specificity of 0.419). With ctx_controlado, specificity rose to 0.778 (Δ +0.359; 95% CI 0.234 to 0.496) and MCC to 0.432 (Δ +0.166; 0.080 to 0.263), but sensitivity fell to 0.747 (Δ −0.250; −0.429 to −0.080). Specificity was higher in 14 of the 15 pairs and in all five architectures, and sensitivity was higher in no pair (Supplementary Material, Table S6).

Excluding the pairs with non-converged runs, the pattern held, with specificity +0.387 (0.254 to 0.533) and sensitivity −0.284 (−0.463 to −0.099). On Makerere, specificity rose from 0.638 to 0.823 (Δ +0.186; 0.090 to 0.297) and sensitivity fell from 0.736 to 0.541 (Δ −0.195; −0.350 to −0.068), with no difference in MCC; with 37 negatives, this specificity estimate is imprecise.

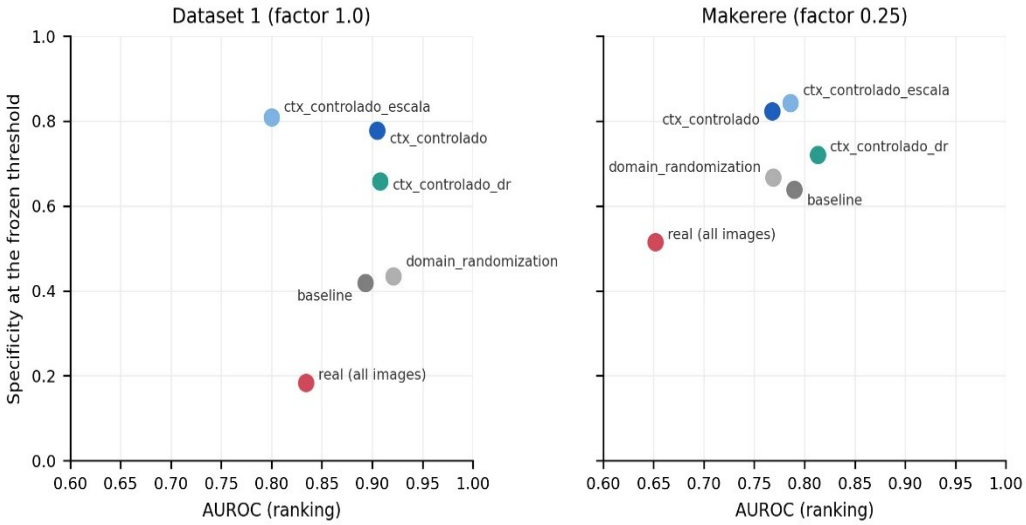
In the 15-network ensembles, the pattern is visible in the counts (Supplementary Material, Table S7). On Dataset 1, the ctx_controlado ensemble correctly classified 880 of the 900 negatives and detected 63 of the 100 positives, versus 588 and 100 for baseline.

On Makerere, it correctly classified 35 of the 37 negatives (0.946; Wilson CI 0.823 to 0.985) and 49.2% of the positives, versus 27 of the 37 (0.730; 0.570 to 0.846) and 75.5% for baseline. The shift is not explained by systematically different thresholds, as their medians were similar across conditions, ranging from 0.978 to 0.995 (Supplementary Material, Table S8).



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

**Figure 4 –** AUROC versus specificity at the frozen threshold, by condition




Source: Prepared by the authors (2026). Note: External dataset 1 at factor 1.0 and Makerere at factor 0.25; mean across runs. Conditions with similar AUROC differ substantially in specificity.

### 3.4 Image scale

On the internal set, the sensitivity of the controlled-context conditions depended strongly on scale (Table 4). At the original scale, ctx_controlado detected 38.3% of the positive images, versus 83.0% for baseline (Δ −0.447; −0.575 to −0.323); with the calibrated scale, the difference decreased to −0.149 (−0.255 to −0.065).

The loss was concentrated in the higher-resolution systems: for ctx_controlado, sensitivity fell from 0.657 on the 800 × 600 pixel images to 0.135 on the 3984 × 2988 pixel images, and calibration raised it to 0.75 to 0.82 in these systems (Figure 5). Excluding the 14 images with label inconsistency increased sensitivity by 0.008 to 0.018 in the synthetic conditions and by 0.004 to 0.007 in the restricted reference, without changing any comparison (Supplementary Material, Table S9).



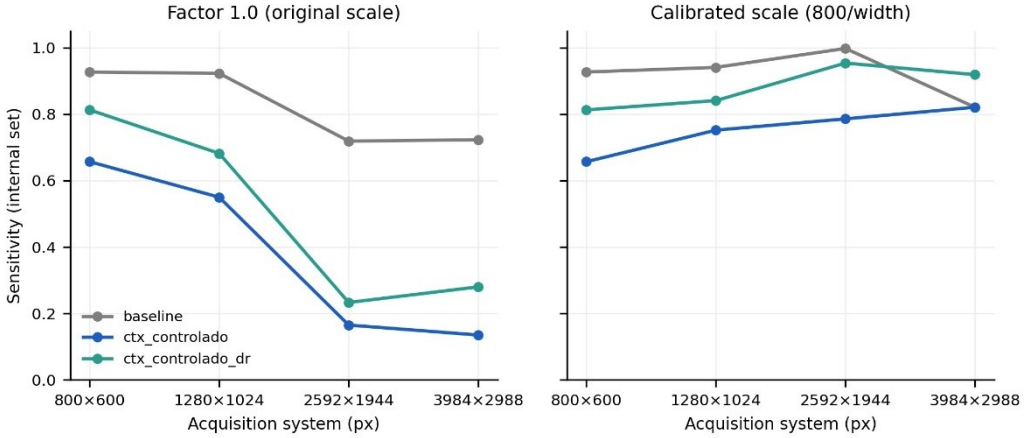
*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

**Table 4 –** Sensitivity on the internal set by image scale

| Condition | Factor 0.25 | Factor 1.0 (original) | Calibrated scale |
|---|---|---|---|
| baseline | 0.698 | 0.830 | 0.905 |
| domain_randomization | 0.645 | 0.874 | 0.887 |
| ctx_controlado | 0.638 | 0.383 | 0.757 |
| ctx_controlado_dr | 0.657 | 0.518 | 0.873 |
| ctx_controlado_escala | 0.601 | 0.491 | 0.691 |

Source: Prepared by the authors (2026). Note: positive images only (627, 623, and 595 evaluable at each scale); mean of 15 runs, frozen threshold.

**Figure 5 –** Sensitivity on the internal set by acquisition system




Source: Prepared by the authors (2026). Note: original scale (factor 1.0) and calibrated scale (800/width); mean of 15 runs.

Random scaling during training increased AUROC on Dataset 1 evaluated at the wrong scale (Δ +0.142; 0.027 to 0.261), but reduced MCC at the native scale (0.211 versus 0.432; Δ −0.220; −0.350 to −0.091), and its ensemble detected only 8 of the 100 positives. Inference combining multiple scales by the maximum score, with the same threshold, reduced specificity on Dataset 1 from 0.778 to 0.572 for ctx_controlado.

### 3.5 Exploratory analyses

When combined with controlled context, domain randomization recovered part of the lost sensitivity: on Dataset 1, from 0.747 to 0.919 (Δ +0.173; 0.032 to 0.331), and, on the internal set at the calibrated scale, from 0.757 to 0.873 (Δ +0.116; 0.019 to 0.246), both of

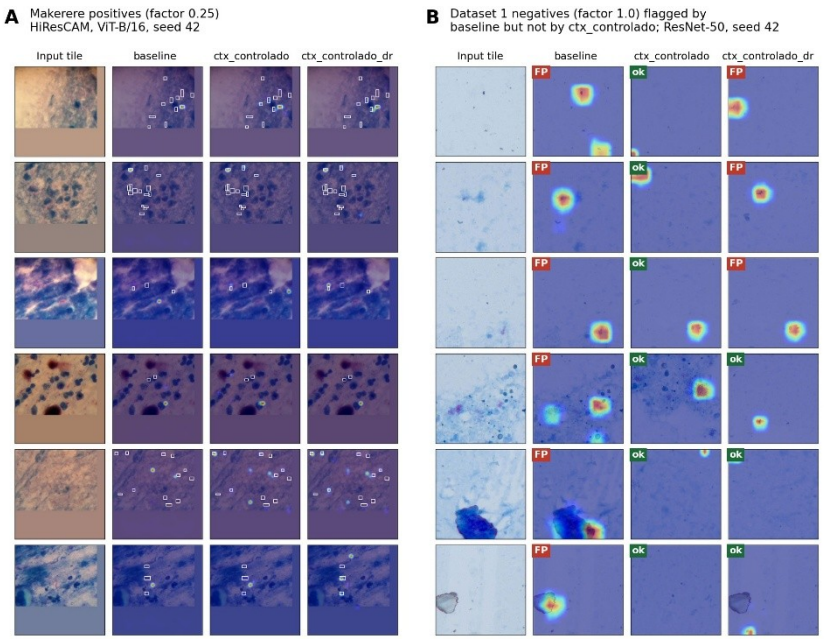


*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

which held after excluding the pairs with non-converged runs. The gain on Makerere (Δ +0.188; 0.003 to 0.374) did not hold in this analysis. On Makerere, sensitivity increased with the number of bacilli in all conditions; in images with a single bacillus, ensemble sensitivity was 0.514 (CI 0.400 to 0.628) for baseline, 0.229 (0.146 to 0.340) for ctx_controlado, and 0.571 (0.455 to 0.681) for ctx_controlado_dr (Supplementary Material, Table S10). AUROC varied more across architectures than across conditions, and the effect of controlled context on AUROC changed sign across architectures (Supplementary Material, Table S11).

In the activation maps, baseline produced more false positives than ctx_controlado in all three seeds (49.1% to 59.6% versus 5.8% to 28.6% of 277 hard negatives), but the localization of these errors in redder or more edge-rich regions was not consistently confirmed. On Makerere, the advantage of ctx_controlado in the spatial hit rate of the maps on the annotated boxes, observed in one seed, was not replicated in the others (Figure 6).

**Figure 6 –** Examples of activation maps from the exploratory analysis




Source: Prepared by the authors (2026).



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

### 3.6 Real-image reference

The real-image reference also lost its operating point outside its own domain (Table 5). The full variant achieved an AUROC of 0.905 on its own validation set and of 0.834 on Dataset 1, but a specificity of only 0.184, lower than that of ctx_controlado (Δ −0.594; −0.791 to −0.422) and of baseline. The restricted variant, with 17 training positives, had an AUROC of 0.695 on Dataset 1 and of 0.527 on Makerere.

**Table 5 –** Real-image reference compared with controlled-context mosaics

| Measure | Real (full) | Real (restricted) | ctx_controlado | ctx_controlado_dr |
|---|---|---|---|---|
| Training images (pos / neg) | 201 / 90 | 17 / 64 | 7,000 / 7,000 | 7,000 / 7,000 |
| AUROC on own validation | 0.905 | 0.799 | 0.965 | 0.951 |
| Dataset 1 · AUROC | 0.834 | 0.695 | 0.905 | 0.908 |
| Dataset 1 · specificity | 0.184 | 0.429 | 0.778 | 0.658 |
| Makerere · AUROC | 0.652 | 0.527 | 0.768 | 0.813 |
| Makerere · specificity | 0.515 | 0.880 | 0.823 | 0.721 |
| Internal · calibrated sensitivity | Not evaluated | 0.453 | 0.757 | 0.873 |

Source: Prepared by the authors (2026). Note: mean of 15 runs (14 for the restricted variant); for the mosaics, the own validation is the synthetic test set. The full variant was not evaluated on the internal set, which contains images from its training data.

## 4. DISCUSSION

This study evaluated two mosaic synthesis procedures under the same training protocol, with a pre-specified external outcome. The central result was negative: controlled context did not improve AUROC on either of the two external datasets. In post hoc analyses at the frozen threshold, however, it shifted the operating point toward higher specificity and lower sensitivity, consistently across pairs, across architectures, and after excluding the non-converged runs. Matching the image scale proved decisive for detection, and training with real images did not prevent performance loss under domain shift either. Discrimination, threshold behavior, and robustness to scale should therefore be interpreted separately, and AUROC alone does not indicate how a system will behave in the laboratory that adopts it.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

### 4.1 Why AUROC remained stable while the operating point shifted

The results are compatible with the hypothesis that contextual cues affect the score distribution. Before training, the color cluster distinguished positive from negative patches (AUROC 0.82) and, even after the bacillus was removed, the backgrounds of the two classes remained discriminable (AUROC 0.72). A network that uses these cues may raise the scores of images with a positive-looking background without substantially changing the ranking between positives and negatives. Baseline, with a sensitivity of 0.997 and a specificity of only 0.419 on Dataset 1, illustrates this behavior. Crystallized deposits of red stain have previously been identified as a source of errors in automated bacillus segmentation in Ziehl-Neelsen-stained images (KHUTLANG et al., 2010). In addition, studies of automated smear microscopy analysis have related limitations in specificity to characteristics of the preparations, including uneven thickness and dust contamination, indicating that non-bacillary artifacts may contribute to spurious positive classifications (WITARTO et al., 2024). This evidence supports the plausibility that visual features not directly related to the presence of bacilli interfere with classification. In this context, the controlled-context generator was designed to reduce these cues by making the two classes share the same background formation process and the same paste boundaries.

The interpretation, however, remains exploratory. The generators change several components simultaneously, use different synthetic distributions for checkpoint and threshold selection, and apply a threshold chosen on a single tile to the maximum score over up to 30 tiles per image, which may penalize conditions whose real negative images receive high scores. In the exploratory interpretability analysis, baseline showed a higher frequency of false positives, but the activation maps did not demonstrate, consistently across seeds, that these errors were concentrated in redder or more edge-rich regions. Therefore, although the results are compatible with an influence of contextual cues, the shift in the operating point cannot be attributed exclusively to the removal of these shortcuts.

The methodological implication is that AUROC alone does not adequately characterize model transportability. In smear microscopy, small per-patch false-alarm rates



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

may result in hundreds of false positives per slide (SERRÃO et al., 2024; SANTIAGO-MOZOS et al., 2014). Similar problems in transferring the decision threshold have also been observed in external validations of models applied to chest radiography and to automated tuberculosis detection in the Nightingale challenge (BEDŐHÁZI et al., 2025; TAVAZIVA et al., 2022). The present results add that the strategy used to construct the synthetic training data can itself modify the operating point without producing significant changes in AUROC. Future studies should therefore jointly report discrimination, sensitivity, specificity, and MCC at thresholds defined without access to the test set, accompanied by confidence intervals and by external evaluations with an adequate number of negative images (COLLINS et al., 2024; VAN CALSTER et al., 2019).

**4.2 Image scale as a determinant of sensitivity**

Acquisition resolution introduced another source of instability. For ctx_controlado, sensitivity at the native scale went from 0.657 on the 800 × 600 pixel system to 0.135 on the 3984 × 2988 pixel system and, after calibration, recovered to 0.75 to 0.82 in the higher-resolution systems. This dependence on apparent size is consistent with the difficulty of representing very small bacilli in resizing operations (LUO et al., 2026; CHEN et al., 2024; ZHENG; ZHANG; LIU, 2026; KHUTLANG et al., 2010; SANTOS et al., 2024).

The hypothesis that the rigid grid compensates for this effect through contextual cues is plausible but was not demonstrated by the activation maps. Scale augmentation increased robustness at a wrong scale at the expense of performance at the correct scale, and combining scales without a new threshold increased false positives. These results favor explicit scale normalization rather than relying solely on learned invariance. Whenever possible, the system should record the objective, camera, and physical resolution (µm/pixel), which was documented in only one of the three datasets; in mobile-phone images, in which the device and its coupling to the eyepiece vary, estimating the apparent size of the bacillus may be particularly useful. The calibration used here, however, depended on prior knowledge of the acquisition setup and does not yet constitute an automatic solution.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

Bacillary load reinforced this dependence: on Makerere, images with a single bacillus were the most difficult, a pattern also observed in the Nightingale external set (BEDŐHÁZI et al., 2025). This is precisely the scenario of greatest interest for automated screening and, at the same time, the most technically challenging one.

### 4.3 Domain randomization and the real-image reference

The effect of domain randomization varied with the generator. On the rigid grid, the additional perturbations had little effect on the results; with controlled context, sensitivity was recovered on Dataset 1 and on the internal set, and ensemble AUROC on Makerere was higher, but the sensitivity gain on Makerere did not withstand the exclusion of the non-converged runs. One hypothesis is that diversifying appearance helps more once background and artifacts no longer predict the class. Among the conditions explored, ctx_controlado_dr showed the most promising operational trade-off, which remains to be confirmed in future studies and does not constitute an indication for clinical use.

The reference trained on real fields also showed external fragility: with an AUROC of 0.905 on validation within its own domain, it had a specificity of 0.184 on Dataset 1, and the restricted version, with only 17 training positives, performed even worse. These findings do not establish the superiority of synthetic data, because the reference was added in a post hoc analysis and used image-level supervision and multiple-instance learning; they show that vulnerability to domain shift persists when training uses real images from a single laboratory. The variability across architectures, greater than the variability across several conditions, and the four non-converged runs further reinforce the risk of conclusions based on a single architecture.

### 4.4 Implications for practice

From a diagnostic standpoint, the ctx_controlado profile does not represent a net gain. In smear microscopy screening, a false negative may delay treatment and sustain transmission, whereas false positives increase the manual review workload and, if not



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

reviewed, lead to unnecessary treatment. The preferable profile depends on the intended use, such as screening with mandatory human review of flagged fields or autonomous reading, and on minimum sensitivity and specificity requirements that this study did not define. Moreover, the results are per field image: the positivity of a slide depends on many fields, and per-patient performance also depends on specimen quality and the number of slides. Three methodological recommendations emerge from the experiments: audit generators for shortcuts before training; record and, when feasible, standardize the acquisition scale; and fix the operating point without using the external dataset, reporting its metrics with confidence intervals. Local recalibration should be evaluated only on an independent set and for a defined clinical purpose (TAVAZIVA et al., 2022).

## 5. CONCLUSION

The controlled-context strategy did not yield a demonstrable gain in AUROC on two external datasets. At the threshold selected exclusively on synthetic validation, however, it increased specificity and MCC on External dataset 1 at the expense of sensitivity, which characterizes a change in the decision profile rather than a uniform improvement. The apparent size of the bacilli substantially affected sensitivity, and even models trained on real fields showed external instability. The contribution of this study is methodological: to assess the transportability of smear microscopy systems, AUROC, frozen-threshold metrics, and acquisition scale must be reported jointly, and diagnostic feasibility remains contingent on prospective slide-level and patient-level validation.

**Acknowledgments.** The authors thank the Instituto Sustentável e Tecnológico da Amazônia (ISATECH) for the infrastructure and support provided for the development of this research.

**Funding.** This work results from a research, development, and innovation project carried out by the Instituto Sustentável e Tecnológico da Amazônia (ISATECH) with resources from Tectoy S.A., invested under the Brazilian Informatics Law.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

**Conflict of interest.** The authors declare no conflict of interest. The funder had no role in the study design, data collection and analysis, interpretation of the results, writing of the manuscript, or the decision to submit it for publication.

**Ethical aspects.** This study used exclusively microscopy images from public databases (ZNSM-iDB, Wellgen Medical/Mendeley Data, and Makerere), made available in anonymized form by their original custodians, without access to clinical data or to any information that could identify individuals. There was no direct participation of human subjects. For this reason, in accordance with Article 1, sole paragraph, items II, III and V, of Resolution CNS No. 510/2016 (Brazilian National Health Council), the research is exempt from registration and review by the CEP/Conep system.

**Data and materials availability.** The datasets used are public and are described in Section 2.2. The supplementary tables (S1 to S11) and the detailed methodological descriptions (M1 to M3) accompany this article as supplementary material. The per-image predictions and the code are available from the corresponding author upon request.

## REFERENCES

AULIA, S. et al. A novel digitized microscopic images of ZN-stained sputum smear and its classification based on IUATLD grades. **IEEE Access**, v. 12, p. 51364-51380, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3386208.

BEDŐHÁZI, Z. et al. Streamlining tuberculosis detection with foundation model-based weakly supervised transformer. **Computers in Biology and Medicine**, v. 195, 110554, 2025. DOI: https://doi.org/10.1016/j.compbiomed.2025.110554.

BRASIL. Ministério da Saúde. **Boletim Epidemiológico: Tuberculose 2026**: número especial. Brasília: Ministério da Saúde, 2026. Available at: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/especiais/2026/boletim-epidemiologico-de-tuberculose-numero-especial-mar-2026.pdf. Accessed on: 13 Aug. 2026.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

CARVALHO BRITO, V. de et al. Multi-task deep learning for bacilli detection and counting in sputum smear microscopy. **Biomedical Signal Processing and Control**, v. 112, 108531, 2026. DOI: https://doi.org/10.1016/j.bspc.2025.108531.

CHEN, H. et al. Research on improved YOLOv8s model for detecting mycobacterium tuberculosis. **Heliyon**, v. 10, n. 18, e38088, 2024. DOI: https://doi.org/10.1016/j.heliyon.2024.e38088.

CHICCO, D.; JURMAN, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. **BMC Genomics**, v. 21, 6, 2020. DOI: https://doi.org/10.1186/s12864-019-6413-7.

COLLINS, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. **BMJ**, v. 385, e078378, 2024. DOI: https://doi.org/10.1136/bmj-2023-078378.

DIETTERICH, T. G.; LATHROP, R. H.; LOZANO-PÉREZ, T. Solving the multiple instance problem with axis-parallel rectangles. **Artificial Intelligence**, v. 89, n. 1-2, p. 31-71, 1997. DOI: https://doi.org/10.1016/S0004-3702(96)00034-3.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 9., 2021. **Proceedings** [...]. [S. l.]: ICLR, 2021.

DRAELOS, R. L.; CARIN, L. **Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks**. [S. l.]: arXiv, 2020. Preprint arXiv:2011.08891.

DWIBEDI, D.; MISRA, I.; HEBERT, M. Cut, paste and learn: surprisingly easy synthesis for instance detection. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2017, Venice. Proceedings [...]. Piscataway: IEEE, 2017. p. 1310-1319. DOI: https://doi.org/10.1109/ICCV.2017.146.

GEIRHOS, R. et al. Shortcut learning in deep neural networks. **Nature Machine Intelligence**, v. 2, n. 11, p. 665-673, 2020. DOI: https://doi.org/10.1038/s42256-020-00257-z.

GREESHMA, K.; VISHNUKUMAR, S. Automated tuberculosis diagnosis: a hybrid approach using attention-residual U-Net segmentation with ensemble classification. **Franklin Open**, v. 14, 100479, 2026. DOI: https://doi.org/10.1016/j.fraope.2025.100479.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

HE, K. et al. Deep residual learning for image recognition. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016, Las Vegas. **Proceedings** [...]. Piscataway: IEEE, 2016. p. 770-778.

HOLM, S. A simple sequentially rejective multiple test procedure. **Scandinavian Journal of Statistics**, v. 6, n. 2, p. 65-70, 1979.

KHUTLANG, R. et al. Classification of Mycobacterium tuberculosis in images of ZN-stained sputum smears. **IEEE Transactions on Information Technology in Biomedicine**, v. 14, n. 4, p. 949-957, 2010. DOI: https://doi.org/10.1109/TITB.2009.2028339.

LIN, E.; WELLGEN MEDICAL. **Detecting Mycobacterium tuberculosis Bacilli from TB Smear Images**. Version 1. [S. l.]: Mendeley Data, 2024. Dataset. DOI: https://doi.org/10.17632/yhybmfrmgs.1.

LIU, Z. et al. Swin Transformer: hierarchical vision transformer using shifted windows. In: IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2021. **Proceedings** [...]. Piscataway: IEEE, 2021. p. 10012-10022.

LIU, Z. et al. A ConvNet for the 2020s. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2022, New Orleans. **Proceedings** [...]. Piscataway: IEEE, 2022. p. 11976-11986.

LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 7., 2019, New Orleans. **Proceedings** [...]. [S. l.]: ICLR, 2019.

LUO, S. et al. A Tubercle Bacilli detection network combining frequency guidance and adaptive multi path fusion. **Engineering Applications of Artificial Intelligence**, v. 177, 114956, 2026. DOI: https://doi.org/10.1016/j.engappai.2026.114956.

MOTA CARVALHO, T. F. et al. A systematic review and repeatability study on the use of deep learning for classifying and detecting tuberculosis bacilli in microscopic images. **Progress in Biophysics and Molecular Biology**, v. 180-181, p. 1-18, 2023. DOI: https://doi.org/10.1016/j.pbiomolbio.2023.03.002.

NAGARAJU, Y. et al. BacilliFinder: revolutionizing tuberculosis detection with computer vision. Indian Journal of Tuberculosis, v. 72, suppl. 2, p. S7-S17, 2025. DOI: https://doi.org/10.1016/j.ijtb.2023.12.005.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

NIJIATI, M. et al. Deep learning on longitudinal CT scans: automated prediction of treatment outcomes in hospitalized tuberculosis patients. **iScience**, v. 26, 108326, 2023. DOI: https://doi.org/10.1016/j.isci.2023.108326.

QUINN, J. A. et al. Deep convolutional neural networks for microscopy-based point of care diagnostics. In: MACHINE LEARNING FOR HEALTHCARE CONFERENCE, 1., 2016. **Proceedings** [...]. [S. l.]: PMLR, 2016. v. 56, p. 271-281.

SANTIAGO-MOZOS, R. et al. An automated screening system for tuberculosis. **IEEE Journal of Biomedical and Health Informatics**, v. 18, n. 3, p. 855-862, 2014. DOI: https://doi.org/10.1109/JBHI.2013.2282874.

SANTOS, P. R. S. dos et al. KochDet: BiFPN-based deep architecture for tuberculosis diagnosis. **Biomedical Signal Processing and Control**, v. 91, 106056, 2024. DOI: https://doi.org/10.1016/j.bspc.2024.106056.

SELVARAJU, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2017, Venice. **Proceedings** [...]. Piscataway: IEEE, 2017. p. 618-626. DOI: https://doi.org/10.1109/ICCV.2017.74.

SERRÃO, M. K. M. et al. Automatic bright-field smear microscopy for diagnosis of pulmonary tuberculosis. **Computers in Biology and Medicine**, v. 172, 108167, 2024. DOI: https://doi.org/10.1016/j.compbiomed.2024.108167.

SHAH, M. I. et al. Ziehl–Neelsen sputum smear microscopy image database: a resource to facilitate automated bacilli detection for tuberculosis diagnosis. **Journal of Medical Imaging**, v. 4, n. 2, 027503, 2017. DOI: https://doi.org/10.1117/1.JMI.4.2.027503.

SHEN, Y. et al. RandStainNA: learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In: MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION, 25., 2022, Singapore. **Proceedings** [...]. Cham: Springer, 2022. p. 212-221. (Lecture Notes in Computer Science, v. 13432) DOI: https://doi.org/10.1007/978-3-031-16434-7_21.

TAMURA, G. et al. Machine-learning methods for detecting tuberculosis in Ziehl-Neelsen stained slides: a systematic literature review. **Intelligent Systems with Applications**, v. 22, 200365, 2024. DOI: https://doi.org/10.1016/j.iswa.2024.200365.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

TAN, M.; LE, Q. V. EfficientNet: rethinking model scaling for convolutional neural networks. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 36., 2019, Long Beach. **Proceedings** [...]. [S. l.]: PMLR, 2019. v. 97, p. 6105-6114.

TAVAZIVA, G. et al. Diagnostic accuracy of a commercially available, deep learning-based chest X-ray interpretation software for detecting culture-confirmed pulmonary tuberculosis. **International Journal of Infectious Diseases**, v. 122, p. 15-20, 2022. DOI: https://doi.org/10.1016/j.ijid.2022.05.037.

TELLEZ, D. et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. **Medical Image Analysis**, v. 58, 101544, 2019. DOI: https://doi.org/10.1016/j.media.2019.101544.

TOBIN, J. et al. Domain randomization for transferring deep neural networks from simulation to the real world. In: IEEE/RSJ INTERNATIONAL CONFERENCE ON INTELLIGENT ROBOTS AND SYSTEMS, 2017, Vancouver. **Proceedings** [...]. Piscataway: IEEE, 2017. p. 23-30. DOI: https://doi.org/10.1109/IROS.2017.8202133.

TREMBLAY, J. et al. Training deep networks with synthetic data: bridging the reality gap by domain randomization. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS, 2018, Salt Lake City. **Proceedings** [...]. Piscataway: IEEE, 2018. p. 969-977.

VAN CALSTER, B. et al. Calibration: the Achilles heel of predictive analytics. **BMC Medicine**, v. 17, 230, 2019. DOI: https://doi.org/10.1186/s12916-019-1466-7.

VISUÑA, L. et al. Automated monitoring of Mycobacterium population growth in time-lapse microscopy with deep learning pseudo-supervised algorithm. **Engineering Applications of Artificial Intelligence**, v. 162, 112503, 2025. DOI: https://doi.org/10.1016/j.engappai.2025.112503.

WANG, Y.; PAN, L.; SHU, X. YOLO-FMS: a lightweight and efficient model for medical microscopic smear detection. **IEEE Access**, v. 12, p. 125253-125265, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3455555.

WIGHTMAN, R. PyTorch Image Models. [S. l.]: GitHub, 2019. Software. Available at: https://github.com/huggingface/pytorch-image-models. DOI: https://doi.org/10.5281/zenodo.4414861.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

WITARTO, A. B. et al. AI-based analysis of Ziehl-Neelsen-stained sputum smears for Mycobacterium tuberculosis as a screening method for active tuberculosis. Life, v. 14, n. 11, article 1418, 2024. DOI: https://doi.org/10.3390/life14111418.

WORLD HEALTH ORGANIZATION. **Global tuberculosis report 2024**. Geneva: World Health Organization, 2024. Available at: https://www.who.int/publications/i/item/9789240101531. Accessed on: 23 Jul. 2026.

WORLD HEALTH ORGANIZATION. Global tuberculosis report 2025. Geneva: World Health Organization, 2025. Available at: https://www.who.int/publications/i/item/9789240116924. Accessed on: 23 Jul. 2026.

WORLD HEALTH ORGANIZATION. Tuberculosis. Geneva: World Health Organization, 2026. Available at: https://www.who.int/news-room/fact-sheets/detail/tuberculosis. Accessed on: 23 Jul. 2026.

XU, Y. et al. Towards multi-view sputum smear quality classification. **Biomedical Signal Processing and Control**, v. 102, 107217, 2025. DOI: https://doi.org/10.1016/j.bspc.2024.107217.

YAN, T. et al. ResGloTBNet: an interpretable deep residual network with global long-range dependency for tuberculosis screening of sputum smear microscopy images. **Medical Engineering & Physics**, v. 137, 104300, 2025. DOI: https://doi.org/10.1016/j.medengphy.2025.104300.

ZECH, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. **PLOS Medicine**, v. 15, n. 11, e1002683, 2018. DOI: https://doi.org/10.1371/journal.pmed.1002683.

ZHENG, R.; ZHANG, C.; LIU, Y. Enhanced feature fusion and detail-preserving network for small object detection in medical microscopic images. **Digital Signal Processing**, v. 174, 105938, 2026. DOI: https://doi.org/10.1016/j.dsp.2026.105938.

ZHONG, Y. et al. Efficient malaria parasite detection from diverse images of thick blood smears for cross-regional model accuracy. **IEEE Open Journal of Engineering in Medicine and Biology**, v. 4, p. 226-233, 2023. DOI: https://doi.org/10.1109/OJEMB.2023.3328435.



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

REVISTA OWL (*OWL Journal* )
www.revistaowl.com.br – ISSN: 2965-2634

**SUPPLEMENTARY MATERIAL**

Synthetic mosaics with controlled context shift the operating point without improving external discrimination in paucibacillary smear classification

**SUPPLEMENTARY METHODS**

**M1. Augmentation, weights, and checkpoint.** Basic augmentation: horizontal and vertical flips and 90° rotation (p = 0.5 each), brightness and contrast ±10%, saturation ±5%, hue ±0.02, and mild Gaussian blur (p = 0.10). Domain randomization: in addition to the basic augmentation, affine transformation (rotation up to 12°, translation up to 2.5%, scaling 0.96 to 1.04; p = 0.5), brightness and contrast ±20%, saturation ±15%, hue ±0.04, gamma 0.88 to 1.12 (p = 0.20), resolution degradation down to 0.78 (p = 0.15), blur (p = 0.12), sharpening (p = 0.12), JPEG compression with quality 72 to 96 (p = 0.10), and Gaussian noise (σ = 0.015; p = 0.15). Random scaling: a factor drawn on a logarithmic scale between 0.5 and 2.0 in 75% of the mosaics; when zooming in, the window of a positive mosaic is drawn among those that contain the entire bacillus; when zooming out, the canvas is filled by mirror padding in both classes. Pretrained weights (timm 1.0.29): resnet50.a1_in1k, efficientnet_b3.ra2_in1k, convnext_tiny.in12k_ft_in1k, vit_base_patch16_224.augreg2_in21k_ft_in1k, and swin_tiny_patch4_window7_224.ms_in1k; the ViT was pretrained with 0.5/0.5 normalization but, like the other architectures, received ImageNet normalization. Early stopping could only occur from the fifth epoch onward; pos_weight = n_neg/n_pos = 1. One run (ResNet-50, seed 42, ctx_controlado_escala) used a Tesla T4 GPU.

**M2. Real-image reference.** The split into training and validation sets was stratified by class and acquisition system, with a target ratio of 75% and 25%, respectively, keeping near-duplicate images in the same partition (texture correlation ≥ 0.90 in the restricted variant and ≥ 0.95 in the full variant). In the full variant, 291 images were used for training (201 positives and 90 negatives) and 97 for validation (68 positives and 29 negatives), corresponding to the 75% and 25% ratios. In the restricted variant, the split yielded 81 training images (17



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

positives and 64 negatives) and 35 validation images (7 positives and 28 negatives), equivalent to 69.8% and 30.2%, respectively.

Training used batches of two whole-field images and positive-class weighting defined by *pos_weight* = n_neg/n_pos (0.45 in the full variant and 3.76 in the restricted variant), with all other hyperparameters kept as in the training with synthetic mosaics. In the restricted variant, the ConvNeXt-Tiny run with seed 42 was interrupted by a file-writing error in cloud storage, unrelated to model convergence. This run was not repeated; therefore, comparisons involving the restricted variant considered 14 architecture–seed pairs.

**M3. Texture signature and interpretability.** The signature downsamples the grayscale image to 80 × 80 px, subtracts a smoothed version, crops the central 64 × 64 px, and normalizes the result; for patches, the image is downsampled to 64 × 64 px and the center crop is 56 × 56 px. For interpretability, 300 Makerere positives were sampled with seed 0, and the tile with the highest probability was analyzed. A hit was defined as the map peak falling inside an annotated box, with a tolerance of 16 px (half a map cell), and the lift as the fraction of activation inside the boxes divided by the fraction of the area they occupy (1 = chance). In Dataset 1, the negatives were those flagged as positive by at least one condition, supplemented by random sampling (seed 0); for each false positive, the peak was characterized by the percentile of the a* channel in LAB within a 9 × 9 px window and by the percentile of the Sobel gradient magnitude.

## SUPPLEMENTARY TABLES

**Table S1 –** Internal-set images labeled as positive and identical, by SHA256, to files in the Without_bacilli folders of the ZNSM-iDB.

| Internal file | Resolution | Identical file in the ZNSM-iDB | SHA256 (prefix) |
|---|---|---|---|
| 1432.jpg | 800 × 600 | Without_bacilli_Micros cope-1_set1/3.jpg | 9f0d0c7fc662 |
| 1786.jpg | 800 × 600 | Without_bacilli_Micros cope-1_set1/2.jpg | 556f22bce709 |
| 1828.jpg | 800 × 600 | Without_bacilli_Micros cope-1_set1/7.jpg | 98b5bcd57ba3 |
| 1835.jpg | 1280 × 1024 | Without_bacilli_Micros | 3aa1541258ac |



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

| | | | |
|---|---|---|---|
| | | cope-2_Set-1/20.jpg | |
| 1840.jpg | 1280 × 1024 | Without_bacilli_Micros cope-2_Set-1/21.jpg | 8633b1b6bda2 |
| 1845.jpg | 1280 × 1024 | Without_bacilli_Micros cope-2_Set-1/22.jpg | 5887e525a771 |
| 1854.jpg | 1280 × 1024 | Without_bacilli_Micros cope-2_Set-1/24.jpg | 7e9eb2c6ebd4 |
| 1857.jpg | 1280 × 1024 | Without_bacilli_Micros cope-2_Set-1/12.jpg | 551da1bc059f |
| 1885.jpg | 1280 × 1024 | Without_bacilli_Micros cope-2_Set-1/30.jpg | d588d81129a2 |
| 1890.jpg | 1280 × 1024 | Without_bacilli_Micros cope-2_Set-1/31.jpg | 5b2b0a348ff3 |
| 1915.jpg | 1280 × 1024 | Without_bacilli_Micros cope-2_Set-1/29.jpg | f88738c908bd |
| 2059.jpg | 800 × 600 | Without_bacilli_Micros cope-1_set1/47.jpg | f665984b1fcf |
| 2114.jpg | 1280 × 1024 | Without_bacilli_Micros cope-2_Set-1/14.jpg | 334e9fb772f4 |
| 2145.jpg | 1280 × 1024 | Without_bacilli_Micros cope-2_Set-1/15.jpg | 11b887159926 |

Source: Prepared by the authors (2026). Note: None of them appears in the Single_or_few or Ocluded folders. The correct label was not determined.

**Table S2 –** Flow of the Makerere images.

| Step | Positives | Negatives | Total |
|---|---|---|---|
| Files with Pascal VOC XML | 1,218 | 47 | 1,265 |
| Excluded by the vignetting filter (factor 0.25) | 1 | 10 | 11 |
| Main evaluation (factor 0.25) | 1,217 | 37 | 1,254 |
| Evaluation at factor 0.366 | 1,216 | 37 | 1,253 |

Source: Prepared by the authors (2026). Note: Positive, at least one box in the Pascal VOC file; negative, no box. Images excluded at factor 0.25 and fraction of dark pixels: positive tuberculosis-phone-0322 (0.27); negatives 0086 (0.66), 0316 (1.00), 0464 (0.29), 0539 (0.27), 0842 (0.26), 1057 (0.41), 1095 (1.00), 1112 (1.00), 1162 (1.00), and 1201 (0.27). The filter excludes tiles with more than 15% dark pixels.

**Table S3 –** Shortcut audit.

| Step | Data | Classifier | Result |
|---|---|---|---|
| Rigid grid: color cluster predicting the class | 15,665 training patches (3,002 pos / 12,663 neg); k-means, k = 24 | Fraction of positives in the cluster | AUROC 0.818 (0.811–0.825) |
| Rigid grid: background without the bacillus | Training 3,000 + 3,000; test 2,870 pos / 6,000 neg; 11 features | Gradient boosting | AUROC 0.724 (0.713–0.734) |
| | idem | Logistic regression | AUROC 0.643 (0.630–0.655) |
| | idem, bacillus present (control) | Gradient boosting | AUROC 0.959 (0.955–0.963) |
| Rigid grid: background with color- | Reweighting of | Gradient boosting | AUROC 0.735 → 0.727 |



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

| | | | |
|---|---|---|---|
| matched negatives | negatives by cluster | | |
| Rigid grid: seams | Negative mosaics | Ratio of the intensity jump at the seam / away from it | L *16.4; a* 14.1 |
| Rigid grid: patch size | Training sample | — | 112 px in 22% of pos and 37% of neg |
| Controlled context: background > 5 px from the mask | Cells: training 3,000, validation 2,800 | Gradient boosting | AUROC 0.511 (0.489–0.533)* |
| Controlled context: 3–5 px ring | idem | Gradient boosting | AUROC 0.484 (0.462–0.506)* |
| Controlled context: seams | Negative mosaics | Ratio of the intensity jump at the seam / away from it | 1.82 |

Source: Prepared by the authors (2026). Note: The 95% CIs were estimated by bootstrap, except for the results marked with an asterisk (*), for which the Hanley and McNeil approximation was used (HANLEY; McNEIL, 1982). Features: mean and standard deviation of *L, a*, and *b*, variance of the Laplacian, mean gradient, high frequency, edge minus center (L*), and reddish fraction. In the controlled context, the classifier distinguishes cells with a pasted bacillus from cells with a sham paste.

**Table S4 –** Sensitivity analysis: paired comparisons with all 15 architecture × seed pairs (main analysis) and without the two pairs with non-converged runs (ConvNeXt-Tiny with seed 2024 and ViT-B/16 with seed 123; 13 pairs).

| **Comparison** | **Outcome** | **Δ, 15 pairs (95% CI)** | **Δ, 13 pairs (95% CI)** | **Conclusion** |
|---|---|---|---|---|
| ctx_controlado − baseline | Dataset 1 AUROC f1.0 | 0.012 (−0.071 to 0.127) | 0.024 (−0.068 to 0.132) | CI status unchanged |
| ctx_controlado − baseline | Makerere AUROC f0.25 | −0.022 (−0.091 to 0.059) | −0.011 (−0.074 to 0.059) | CI status unchanged |
| ctx_controlado − baseline | Makerere sensitivity f0.25 | −0.195 (−0.350 to −0.068) | −0.171 (−0.275 to −0.089) | CI status unchanged |
| ctx_controlado − baseline | Internal calibrated sensitivity | −0.149 (−0.255 to −0.065) | −0.157 (−0.262 to −0.069) | CI status unchanged |
| ctx_controlado − baseline | Dataset 1 AUROC f0.25 (wrong scale) | −0.033 (−0.158 to 0.094) | −0.028 (−0.159 to 0.105) | CI status unchanged |
| ctx_controlado − baseline | Makerere AUROC f0.366 (wrong scale) | −0.049 (−0.112 to 0.009) | −0.044 (−0.101 to 0.011) | CI status unchanged |
| ctx_controlado − baseline | Internal sensitivity f1.0 (wrong scale) | −0.447 (−0.575 to −0.323) | −0.465 (−0.583 to −0.338) | CI status unchanged |
| ctx_controlado − baseline | Dataset 1 specificity f1.0 | 0.359 (0.234 to 0.496) | 0.387 (0.254 to 0.533) | CI status unchanged |
| ctx_controlado − baseline | Dataset 1 MCC f1.0 | 0.166 (0.080 to 0.263) | 0.180 (0.087 to 0.290) | CI status unchanged |
| ctx_controlado − baseline | Dataset 1 sensitivity f1.0 | −0.250 (−0.429 to −0.080) | −0.284 (−0.463 to −0.099) | CI status unchanged |
| ctx_controlado − baseline | Makerere specificity f0.25 | 0.186 (0.090 to 0.297) | 0.198 (0.111 to 0.317) | CI status unchanged |
| ctx_controlado − baseline | Makerere MCC f0.25 | −0.037 (−0.116 to 0.034) | −0.024 (−0.086 to 0.048) | CI status unchanged |



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

| | | | | |
|---|---|---|---|---|
| ctx_controlado_dr − ctx_controlado | Dataset 1 AUROC f1.0 | 0.003 (−0.079 to 0.072) | 0.029 (−0.015 to 0.087) | CI status unchanged |
| ctx_controlado_dr − ctx_controlado | Makerere AUROC f0.25 | 0.045 (−0.020 to 0.123) | 0.056 (−0.013 to 0.125) | CI status unchanged |
| ctx_controlado_dr − ctx_controlado | Makerere sensitivity f0.25 | 0.188 (0.003 to 0.374) | 0.158 (−0.013 to 0.294) | CI now includes zero |
| ctx_controlado_dr − ctx_controlado | Internal calibrated sensitivity | 0.116 (0.019 to 0.246) | 0.117 (0.009 to 0.246) | CI status unchanged |
| ctx_controlado_dr − ctx_controlado | Dataset 1 AUROC f0.25 (wrong scale) | −0.014 (−0.086 to 0.051) | 0.002 (−0.063 to 0.072) | CI status unchanged |
| ctx_controlado_dr − ctx_controlado | Makerere AUROC f0.366 (wrong scale) | 0.059 (−0.009 to 0.126) | 0.081 (0.022 to 0.130) | CI now excludes zero |
| ctx_controlado_dr − ctx_controlado | Internal sensitivity f1.0 (wrong scale) | 0.135 (0.005 to 0.278) | 0.090 (−0.038 to 0.198) | CI now includes zero |
| ctx_controlado_dr − ctx_controlado | Dataset 1 specificity f1.0 | −0.120 (−0.251 to −0.006) | −0.078 (−0.173 to 0.032) | CI now includes zero |
| ctx_controlado_dr − ctx_controlado | Dataset 1 MCC f1.0 | −0.030 (−0.126 to 0.056) | −0.002 (−0.088 to 0.090) | CI status unchanged |
| ctx_controlado_dr − ctx_controlado | Dataset 1 sensitivity f1.0 | 0.173 (0.032 to 0.331) | 0.195 (0.033 to 0.347) | CI status unchanged |
| ctx_controlado_dr − ctx_controlado | Makerere specificity f0.25 | −0.103 (−0.292 to 0.054) | −0.060 (−0.162 to 0.069) | CI status unchanged |
| ctx_controlado_dr − ctx_controlado | Makerere MCC f0.25 | 0.038 (−0.028 to 0.096) | 0.043 (−0.039 to 0.097) | CI status unchanged |
| ctx_controlado_escala − ctx_controlado | Dataset 1 AUROC f1.0 | −0.105 (−0.188 to −0.026) | −0.089 (−0.159 to 0.002) | CI now includes zero |
| ctx_controlado_escala − ctx_controlado | Makerere AUROC f0.25 | 0.019 (−0.036 to 0.076) | 0.028 (−0.028 to 0.076) | CI status unchanged |
| ctx_controlado_escala − ctx_controlado | Makerere sensitivity f0.25 | −0.003 (−0.097 to 0.114) | 0.005 (−0.101 to 0.147) | CI status unchanged |
| ctx_controlado_escala − ctx_controlado | Internal calibrated sensitivity | −0.066 (−0.232 to 0.074) | −0.015 (−0.097 to 0.102) | CI status unchanged |
| ctx_controlado_escala − ctx_controlado | Dataset 1 AUROC f0.25 (wrong scale) | 0.142 (0.027 to 0.261) | 0.178 (0.056 to 0.279) | CI status unchanged |
| ctx_controlado_escala − ctx_controlado | Makerere AUROC f0.366 (wrong scale) | 0.012 (−0.049 to 0.072) | 0.030 (−0.029 to 0.080) | CI status unchanged |
| ctx_controlado_escala − ctx_controlado | Internal sensitivity f1.0 (wrong scale) | 0.108 (−0.025 to 0.224) | 0.121 (0.019 to 0.240) | CI now excludes zero |
| ctx_controlado_escala − ctx_controlado | Dataset 1 specificity f1.0 | 0.031 (−0.081 to 0.163) | 0.022 (−0.089 to 0.137) | CI status unchanged |
| ctx_controlado_escala − ctx_controlado | Dataset 1 MCC f1.0 | −0.220 (−0.350 to −0.091) | −0.220 (−0.338 to −0.052) | CI status unchanged |
| ctx_controlado_escala − ctx_controlado | Dataset 1 sensitivity f1.0 | −0.308 (−0.531 to −0.055) | −0.283 (−0.490 to 0.035) | CI now includes zero |
| ctx_controlado_escala − ctx_controlado | Makerere specificity f0.25 | 0.020 (−0.074 to 0.110) | 0.008 (−0.105 to 0.122) | CI status unchanged |
| ctx_controlado_escala − | Makerere MCC | 0.001 (−0.042 to | −0.004 (−0.060 to | CI status |



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

| | | | | |
|---|---|---|---|---|
| ctx_controlado | f0.25 | 0.038) | 0.042) | unchanged |
| domain_randomization − baseline | Dataset 1 AUROC f1.0 | 0.028 (−0.053 to 0.138) | 0.052 (−0.002 to 0.146) | CI status unchanged |
| domain_randomization − baseline | Makerere AUROC f0.25 | −0.021 (−0.092 to 0.065) | −0.010 (−0.077 to 0.060) | CI status unchanged |
| domain_randomization − baseline | Makerere sensitivity f0.25 | −0.014 (−0.115 to 0.116) | −0.026 (−0.150 to 0.097) | CI status unchanged |
| domain_randomization − baseline | Internal calibrated sensitivity | −0.019 (−0.095 to 0.054) | 0.001 (−0.046 to 0.054) | CI status unchanged |
| domain_randomization − baseline | Dataset 1 AUROC f0.25 (wrong scale) | −0.063 (−0.154 to 0.025) | −0.038 (−0.133 to 0.036) | CI status unchanged |
| domain_randomization − baseline | Makerere AUROC f0.366 (wrong scale) | −0.015 (−0.071 to 0.031) | −0.002 (−0.045 to 0.031) | CI status unchanged |
| domain_randomization − baseline | Internal sensitivity f1.0 (wrong scale) | 0.043 (−0.012 to 0.108) | 0.037 (−0.037 to 0.102) | CI status unchanged |
| domain_randomization − baseline | Dataset 1 specificity f1.0 | 0.015 (−0.093 to 0.116) | 0.035 (−0.053 to 0.138) | CI status unchanged |
| domain_randomization − baseline | Dataset 1 MCC f1.0 | 0.004 (−0.053 to 0.056) | 0.013 (−0.034 to 0.067) | CI status unchanged |
| domain_randomization − baseline | Dataset 1 sensitivity f1.0 | 0.003 (0.000 to 0.016) | 0.004 (0.000 to 0.016) | CI status unchanged |
| domain_randomization − baseline | Makerere specificity f0.25 | 0.029 (−0.106 to 0.157) | 0.073 (−0.043 to 0.230) | CI status unchanged |
| domain_randomization − baseline | Makerere MCC f0.25 | −0.022 (−0.095 to 0.048) | −0.015 (−0.102 to 0.048) | CI status unchanged |

Source: Prepared by the authors (2026). Note: "CI status unchanged": the CI excludes zero in both analyses or includes it in both.

**Table S5 –** Synthetic test (700 + 700 mosaics from the respective generator).

| Condition | AUROC | MCC | Sensitivity | Specificity | ECE | Ensemble AUROC |
|---|---|---|---|---|---|---|
| baseline | 0.967 | 0.831 | 0.918 | 0.912 | 0.085 | 0.985 |
| domain_randomization | 0.943 | 0.795 | 0.897 | 0.897 | 0.078 | 0.988 |
| ctx_controlado | 0.965 | 0.893 | 0.921 | 0.968 | 0.038 | 0.999 |
| ctx_controlado_dr | 0.951 | 0.851 | 0.946 | 0.907 | 0.039 | 0.996 |
| ctx_controlado_escala | 0.955 | 0.864 | 0.934 | 0.929 | 0.037 | 0.999 |

Source: Prepared by the authors (2026). Note: Mean of the 15 runs; threshold chosen by MCC on the synthetic validation set. Values are not comparable across generators.

**Table S6 –** Specificity and sensitivity in Dataset 1 (factor 1.0) by architecture, at the frozen threshold (mean ± SD of the three seeds).

| Architecture | Metric | baseline | domain_rand. | ctx_controlado | ctx_controlado_dr | ctx_controlado_escala |
|---|---|---|---|---|---|---|
| ResNet-50 | Specificity | 0.424 ± 0.057 | 0.435 ± 0.042 | 0.777 ± 0.276 | 0.645 ± 0.060 | 0.944 ± 0.050 |
| ResNet-50 | Sensitivity | 1.000 ± | 1.000 ± | 0.607 ± | 0.940 ± | 0.203 ± |



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 0.000 | 0.000 | 0.324 | 0.000 | 0.248 |
| EfficientNet-B3 | Specificity | 0.516 ± 0.343 | 0.517 ± 0.040 | 0.971 ± 0.009 | 0.834 ± 0.052 | 0.973 ± 0.027 |
| EfficientNet-B3 | Sensitivity | 0.983 ± 0.029 | 1.000 ± 0.000 | 0.517 ± 0.042 | 0.880 ± 0.092 | 0.147 ± 0.119 |
| ConvNeXt-Tiny | Specificity | 0.394 ± 0.109 | 0.524 ± 0.052 | 0.843 ± 0.186 | 0.618 ± 0.536 | 0.925 ± 0.068 |
| ConvNeXt-Tiny | Sensitivity | 1.000 ± 0.000 | 1.000 ± 0.000 | 0.730 ± 0.355 | 0.840 ± 0.160 | 0.500 ± 0.433 |
| Swin-T | Specificity | 0.475 ± 0.046 | 0.443 ± 0.105 | 0.832 ± 0.111 | 0.738 ± 0.193 | 0.864 ± 0.031 |
| Swin-T | Sensitivity | 1.000 ± 0.000 | 1.000 ± 0.000 | 0.887 ± 0.040 | 0.937 ± 0.055 | 0.347 ± 0.168 |
| ViT-B/16 | Specificity | 0.289 ± 0.126 | 0.256 ± 0.129 | 0.467 ± 0.064 | 0.454 ± 0.079 | 0.340 ± 0.116 |
| ViT-B/16 | Sensitivity | 1.000 ± 0.000 | 1.000 ± 0.000 | 0.993 ± 0.012 | 1.000 ± 0.000 | 0.997 ± 0.006 |

Source: Prepared by the authors (2026). Note: With controlled context, specificity was higher and sensitivity was lower or equal in all architectures.

**Table S7 –** Ensembles of 15 networks (14 in the restricted real variant): AUROC and confusion matrices at the threshold chosen on validation and frozen.

| **Dataset · condition** | **AUROC** | **Threshold** | **TP** | **FN** | **TN** | **FP** | **Sens. (95% CI)** | **Spec. (95% CI)** | **MCC** |
|---|---|---|---|---|---|---|---|---|---|
| Dataset 1 · baseline | 0.978 | 0.784 | 100 | 0 | 588 | 312 | 1.000 (0.963–1.000) | 0.653 (0.622–0.684) | 0.398 |
| Dataset 1 · domain_randomization | 0.980 | 0.729 | 100 | 0 | 529 | 371 | 1.000 (0.963–1.000) | 0.588 (0.555–0.619) | 0.353 |
| Dataset 1 · ctx_controlado | 0.964 | 0.695 | 63 | 37 | 880 | 20 | 0.630 (0.532–0.718) | 0.978 (0.966–0.986) | 0.661 |
| Dataset 1 · ctx_controlado_dr | 0.957 | 0.726 | 89 | 11 | 814 | 86 | 0.890 (0.814–0.937) | 0.904 (0.883–0.922) | 0.627 |
| Dataset 1 · ctx_controlado_escala | 0.891 | 0.659 | 8 | 92 | 876 | 24 | 0.080 (0.041–0.150) | 0.973 (0.961–0.982) | 0.091 |
| Dataset 1 · Real (full) | 0.960 | 0.591 | 100 | 0 | 87 | 813 | 1.000 (0.963–1.000) | 0.097 (0.079–0.118) | 0.103 |
| Dataset 1 · Real (restricted) | 0.867 | 0.584 | 2 | 98 | 900 | 0 | 0.020 (0.006–0.070) | 1.000 (0.996–1.000) | 0.134 |
| Makerere · baseline | 0.841 | 0.784 | 919 | 298 | 27 | 10 | 0.755 (0.730–0.778) | 0.730 (0.570–0.846) | 0.187 |
| Makerere · domain_randomization | 0.843 | 0.729 | 854 | 363 | 30 | 7 | 0.702 (0.675–0.727) | 0.811 (0.658–0.905) | 0.187 |
| Makerere · ctx_controlado | 0.862 | 0.695 | 599 | 618 | 35 | 2 | 0.492 (0.464–0.520) | 0.946 (0.823–0.985) | 0.148 |



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Makerere · ctx_controlado_dr | 0.906 | 0.726 | 904 | 313 | 33 | 4 | 0.743 (0.718–0.767) | 0.892 (0.753–0.957) | 0.240 |
| Makerere · ctx_controlado_escala | 0.891 | 0.659 | 715 | 502 | 34 | 3 | 0.588 (0.560–0.615) | 0.919 (0.787–0.972) | 0.173 |
| Makerere · Real (full) | 0.782 | 0.591 | 1087 | 130 | 16 | 21 | 0.893 (0.875–0.909) | 0.432 (0.287–0.591) | 0.172 |
| Makerere · Real (restricted) | 0.557 | 0.584 | 0 | 1217 | 37 | 0 | 0.000 (0.000–0.003) | 1.000 (0.906–1.000) | 0.000 |

Source: Prepared by the authors (2026). Note: Dataset 1, factor 1.0 (100 positives and 900 negatives); Makerere, factor 0.25 (1,217 and 37). TP, true positives; FN, false negatives; TN, true negatives; FP, false positives. Wilson 95% CI.

**Table S8 –** Decision thresholds of the 15 runs of each condition, selected by maximizing the MCC on the synthetic validation set.

| Condition | Runs | Median | Interquartile range | Minimum–maximum |
|---|---|---|---|---|
| baseline | 15 | 0.990 | 0.944–0.999 | 0.051–>0.999 |
| domain_randomization | 15 | 0.978 | 0.728–0.994 | 0.083–>0.999 |
| ctx_controlado | 15 | 0.995 | 0.732–0.999 | 0.136–>0.999 |
| ctx_controlado_dr | 15 | 0.989 | 0.925–0.997 | 0.195–>0.999 |
| ctx_controlado_escala | 15 | 0.979 | 0.896–0.999 | 0.027–>0.999 |

Source: Prepared by the authors (2026).

**Table S9 –** Sensitivity in the internal set with → without the 14 images with label inconsistency (number of evaluable images in parentheses).

| Condition | Factor 0.25 (627 → 613) | Factor 1.0 (623 → 609) | Calibrated (595 → 581) |
|---|---|---|---|
| baseline | 0.698 → 0.710 | 0.830 → 0.847 | 0.905 → 0.923 |
| domain_randomization | 0.645 → 0.657 | 0.874 → 0.890 | 0.887 → 0.903 |
| ctx_controlado | 0.638 → 0.650 | 0.383 → 0.391 | 0.757 → 0.772 |
| ctx_controlado_dr | 0.657 → 0.668 | 0.518 → 0.527 | 0.873 → 0.890 |
| ctx_controlado_escala | 0.601 → 0.613 | 0.491 → 0.502 | 0.691 → 0.706 |
| Real (restricted) | 0.370 → 0.374 | 0.527 → 0.534 | 0.453 → 0.459 |

Source: Prepared by the authors (2026). Note: Mean of the runs, same frozen threshold.

**Table S10 –** Sensitivity of the 15-network ensemble in Makerere (factor 0.25, frozen threshold) by number of annotated bacilli in the image, with Wilson 95% CI.

| Condition | 1 (n = 70) | 2–3 (n = 230) | 4–7 (n = 374) | ≥ 8 (n = 543) |
|---|---|---|---|---|
| baseline | 0.514 (0.400–0.628) | 0.622 (0.558–0.682) | 0.738 (0.691–0.780) | 0.855 (0.822–0.882) |
| domain_randomization | 0.400 (0.293–0.517) | 0.535 (0.470–0.598) | 0.684 (0.636–0.730) | 0.823 (0.789–0.853) |
| ctx_controlado | 0.229 (0.146–0.340) | 0.348 (0.289–0.411) | 0.449 (0.400–0.500) | 0.617 (0.575–0.657) |
| ctx_controlado_dr | 0.571 (0.455– | 0.578 (0.514– | 0.730 (0.683– | 0.843 (0.810– |



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*

| | | | | |
|---|---|---|---|---|
| | 0.681) | 0.640) | 0.772) | 0.872) |
| ctx_controlado_escala | 0.343 (0.242–0.460) | 0.448 (0.385–0.512) | 0.551 (0.500–0.600) | 0.703 (0.664–0.740) |
| Real (full) | 0.714 (0.599–0.807) | 0.826 (0.772–0.870) | 0.877 (0.840–0.907) | 0.956 (0.935–0.970) |

Source: Prepared by the authors (2026). Note: The restricted real variant did not detect any positive at the frozen threshold and was omitted.

**Table S11 –** AUROC by architecture in the external datasets (mean ± SD of the three seeds; Dataset 1 at factor 1.0 and Makerere at factor 0.25).

| Dataset · architecture | baseline | domain_rand. | ctx_controlado | ctx_controlado_dr | ctx_controlado_escala | Real (full) | Real (restricted) |
|---|---|---|---|---|---|---|---|
| Dataset 1 · ResNet-50 | 0.859 ± 0.015 | 0.903 ± 0.044 | 0.820 ± 0.119 | 0.916 ± 0.010 | 0.765 ± 0.054 | 0.902 | 0.773 |
| Dataset 1 · EfficientNet-B3 | 0.760 ± 0.279 | 0.920 ± 0.035 | 0.937 ± 0.010 | 0.928 ± 0.026 | 0.749 ± 0.157 | 0.883 | 0.518 |
| Dataset 1 · ConvNeXt-Tiny | 0.970 ± 0.007 | 0.967 ± 0.025 | 0.882 ± 0.092 | 0.804 ± 0.227 | 0.767 ± 0.231 | 0.814 | 0.852 |
| Dataset 1 · Swin-T | 0.927 ± 0.015 | 0.925 ± 0.029 | 0.936 ± 0.021 | 0.941 ± 0.021 | 0.793 ± 0.015 | 0.913 | 0.708 |
| Dataset 1 · ViT-B/16 | 0.949 ± 0.046 | 0.890 ± 0.135 | 0.952 ± 0.030 | 0.952 ± 0.015 | 0.926 ± 0.012 | 0.660 | 0.677 |
| Makerere · ResNet-50 | 0.805 ± 0.041 | 0.797 ± 0.016 | 0.755 ± 0.052 | 0.831 ± 0.003 | 0.827 ± 0.014 | 0.741 | 0.456 |
| Makerere · EfficientNet-B3 | 0.591 ± 0.125 | 0.668 ± 0.268 | 0.615 ± 0.161 | 0.720 ± 0.229 | 0.640 ± 0.149 | 0.515 | 0.515 |
| Makerere · ConvNeXt-Tiny | 0.857 ± 0.006 | 0.814 ± 0.056 | 0.776 ± 0.105 | 0.786 ± 0.173 | 0.751 ± 0.185 | 0.626 | 0.542 |
| Makerere · Swin-T | 0.818 ± 0.015 | 0.797 ± 0.037 | 0.806 ± 0.064 | 0.861 ± 0.027 | 0.836 ± 0.041 | 0.856 | 0.601 |
| Makerere · ViT-B/16 | 0.880 ± 0.011 | 0.768 ± 0.092 | 0.887 ± 0.017 | 0.868 ± 0.018 | 0.879 ± 0.022 | 0.521 | 0.527 |

Source: Prepared by the authors (2026). Note: For the synthetic conditions, means and standard deviations across three seeds are reported for each architecture. For the real-image reference models, only means are presented. In the restricted real-image variant, ConvNeXt-Tiny was evaluated with two seeds because one run was interrupted. Therefore, the overall mean for this variant was calculated across 14 runs and may differ from the unweighted mean of the five architecture-specific means. The ranking of experimental conditions varied across architectures.

## REFERENCE

HANLEY, J. A.; McNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, v. 143, n. 1, p. 29-36, 1982. DOI: https://doi.org/10.1148/radiology.143.1.7063747.

*Recebido em: 20/09/2026*
*Aprovado em: 12/10/2026*
*Publicado em: 04/11/2026*



*A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)*