# Uncertainty-Aware Evaluation of Deep Learning Object Detectors under Scarce and Evolving Test Datasets

Esla Timothy Anzaku[1,2], Mohammed Aliy Mohammed[3,4], Stefan Magez[5], Sophie Van Hoecke[3], Arnout Van Messem[6], and Wesley De Neve[1,2,3]

[1] Department of Electronics and Information Systems, Ghent University, Belgium
[2] Center for Biosystems and Biotech Data Science, Ghent University Global Campus, Incheon, Korea
[3] IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium
[4] School of Biomedical Engineering, Jimma Institute of Technology, Jimma University, Jimma, Oromia, Ethiopia
[5] Department of Bio-engineering Sciences, Vrije Universiteit Brussel, Brussel, Belgium
[6] Department of Mathematics, University of Liège, Liège, Belgium

**Abstract.** In data-scarce domains, building reliable deep learning models often requires the relabeling, merging, or expansion of existing datasets. While these steps improve dataset quality and diversity, they complicate model evaluation: changes in evaluation outcomes may arise from dataset changes rather than real model improvement, making standard evaluation protocols difficult to interpret. We demonstrate this challenge in the context of trypanosome parasite detection, where model comparisons across three dataset versions yield inconsistent conclusions. Conventional metrics such as mean average precision (mAP) fail to reveal critical reliability issues, particularly when test data evolves. To address this, we propose a complementary evaluation approach based on predictive uncertainty. By assessing how well models distinguish true from false positives, including on out-of-distribution samples, we obtain a more stable and informative signal of model quality across dataset versions. Our findings show that uncertainty-aware evaluation exposes overlooked failure modes, enables more meaningful comparisons across evolving datasets, and highlights models that maintain reliable confidence estimates under distribution shift.

**Keywords:** DNN Reliability · Neglected Tropical Diseases · Object Detection · Trypanosome Parasite Detection

## 1 Introduction

In data-scarce domains such as parasite detection for neglected tropical diseases (NTDs), building reliable deep neural networks (DNNs) often requires iterative

dataset development—relabeling, merging fragmented datasets, and incrementally expanding biological diversity. While these steps improve dataset quality, they complicate evaluation: conventional protocols assume a fixed test set and may yield ambiguous or misleading signals as test distributions shift. Apparent gains in performance may reflect changes in dataset composition rather than actual model improvement.

We illustrate this challenge through trypanosome parasite detection, a core task in AI-based microscopy diagnostics. We constructed three dataset versions: V1 comprises public microscopy images of unstained thick blood smears; V2 refines V1 via annotation corrections guided by model feedback and manual review; and V3 extends V2 with additional images from a different parasite species and staining protocol. This reflects a realistic dataset evolution trajectory, where both annotation quality and domain diversity improve. Full details are in Section 3.

As test distributions evolve, comparing model performance becomes more difficult. Even without changes to the model, enhanced annotations or new image domains can shift evaluation outcomes. Standard metrics like mean average precision (mAP), which assume a fixed distribution, offer no way to account for such changes. This undermines comparability across dataset versions and obscures true model progress. In domains where dataset curation must evolve, the absence of a stable, reliably measurable model property poses a fundamental challenge.

This ambiguity has practical implications. Diagnostic models must not only be accurate but also exhibit trustworthy uncertainty. Overconfidence on negatives and under-confidence on clear positives represent distinct failure modes that standard metrics may miss. Without a way to assess predictive reliability across dataset versions, meaningful progress becomes difficult to quantify.

We address this by proposing an uncertainty-aware evaluation framework that complements conventional object detection metrics. It leverages near-OOD samples—images from similar capture pipelines or biological domains that lack target objects—as a stable anchor for evaluating model behavior. By explicitly quantifying how well confidence separates true from false positives and distinguishes in- from near-OOD samples, the framework assesses a core property of reliable models: the ability to assign higher confidence to correct predictions. This expectation remains valid across dataset shifts, offering a robust basis for tracking reliability in evolving data settings.

### 1.1 Contributions

We make the following contributions to the evaluation of DNNs in data-scarce NTD settings:

- **Problem Formalization: Evaluation Instability Under Dataset Evolution.** We identify and formalize the issue of evaluation instability when test datasets evolve over time, and demonstrate its impact on the reliability of DNN assessment in domains where datasets are scarce and fragmented.

- **Dataset Contributions: Enhanced Annotations and Biological Diversity.** We release two new dataset versions: V2, with corrected and refined annotations, and V3, which introduces additional species and microscopy capture conditions. These improvements support richer species diversity and better reflect real-world diagnostic variability.
- **Uncertainty-Aware Evaluation approach.** We propose a complementary approach that assesses the quality of the predictive uncertainty over near-OOD samples, in addition to a selected number of standard metrics, to assess model reliability. This approach provides richer model effectiveness information across dataset revisions, especially when standard evaluation approaches fail to reflect meaningful evaluation signals.
- **Reliability Analysis: Experiments and Insights Across Evolving Datasets.** We perform comprehensive experiments by training 8 object detection models and evaluating them across 3 dataset versions. Our analysis reveals critical failure modes and limitations of standard metrics under dataset evolution.

## 2 Related Work

DNN object detectors have the potential to transform microscopy-based diagnostics by enabling faster and more accessible disease detection in low-resource settings. Several initiatives have explored the integration of AI into global health diagnostics [17, 13, 5], but their success hinges on access to diverse, well-annotated datasets that support the development of reliable models.

Such datasets are essential because DNNs can generalize from minimal task specifications (e.g., microscopy images and parasite bounding boxes) when provided with a suitable objective function. However, this same capacity raises reliability concerns, as DNNs are prone to overfitting to dataset-specific biases or spurious correlations (e.g., hospital-specific artifacts or imaging markers) that undermine generalization under distribution shifts [18, 7, 4]—a particularly serious issue in clinical contexts where failures are hard to detect. Moreover, practices like label refinement, dataset merging, and incremental expansion, while critical for improving dataset quality, can introduce instability into evaluation [14, 16]. As datasets evolve alongside model development, there is a growing need for evaluation strategies that can track model reliability and effectiveness consistently across dataset versions.

## 3 Methodology and Experimental Setup

This section details the experimental setup for evaluating object detector reliability across evolving datasets, covering dataset preparation, training, and evaluation. The goal is to capture changes in model effectiveness and uncertainty across dataset versions. As shown in Figure 1, each model is evaluated under three test settings—i.i.d., mixed i.i.d. and near-OOD, and strictly OOD—using

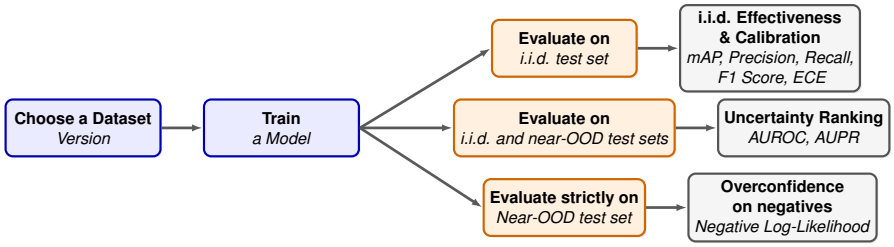



Fig. 1: Overview of our experimental pipeline. For each dataset version, models are trained and evaluated under three conditions: i.i.d., mixed, and near-OOD. Each yields distinct metrics capturing detection accuracy, uncertainty ranking, and overconfidence on negative samples.

metrics that assess accuracy, uncertainty ranking, and overconfidence on negative samples. This process is repeated for each model–dataset pair, and results are visualized through performance trends and confidence histograms. The following paragraphs describe each pipeline component in detail.

**Datasets.** The Tryp dataset [2] comprises microscopy images of *T. brucei brucei* from unstained thick smears but we found annotation issues in it. Other datasets offer point labels [13], cropped images [20, 19], or motion analysis [12], limiting their utility for detection.

Therefore, we curated three dataset versions. Version 1 (V1) comprises of Tryp as-is, that is, the train, validation and test partitions are the same. Version V2 (V2) comprises of the same images and dataset partitions in V1 (that is, the train, validation and test partitions are the same), but improves the annotation quality via model-assisted and manual relabeling. Version V3 (V3) extends V2 with re-annotated giemsa-stained images of *T. cruzi* from [13], increasing domain and species diversity. To evaluate robustness and false positive suppression, we assemble a near-OOD set from two sources: (i) microscopy images from non-infected blood samples [2, 1], and (ii) microscopy images of other infections (e.g., *Plasmodium*, *M. tuberculosis*) [1] captured under a similar imaging protocol as in [13].

Dataset statistics for validation splits of V1–V3 are shown in Table 1, with image examples in Figures 2 and 3. Full dataset details are in the cited source publications.

**Models.** We selected object detectors representative of major architectural paradigms: two-stage (Faster R-CNN [15]) and one-stage (RetinaNet [8], RTMDet [11], RT-DETRv2 [10]) object detectors. Faster-RCNN and RetinaNet are classic deep learning object detectors, while RTMDet and RT-DETRv2 are more recent detectors. Faster-RCNN, RetinaNet, and RTMDet employ a CNN-based architecture design; RT-DETRv2 adopts a transformer-based decoder with end-to-end detection via global attention. Each model is instantiated with two backbone capacities: lightweight variants (Faster-RCNN-r18,

Table 1: Summary of test datasets: V1, V2, V3 (progressively refined trypanosome detection); Near-OOD (similar-condition microscopy images without parasites).

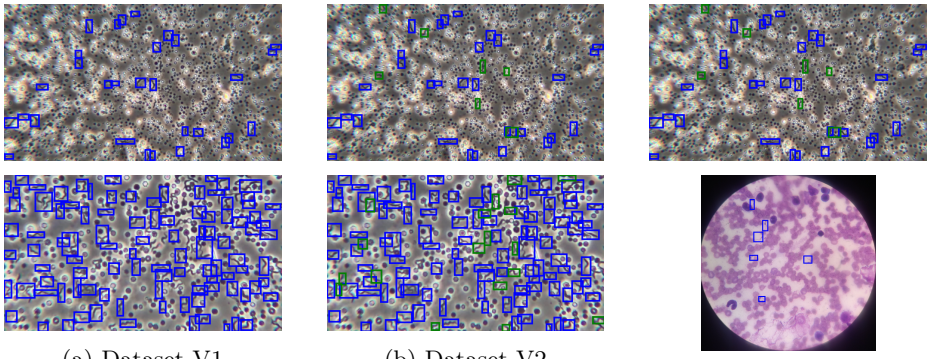
| Dataset Version | Total #Images | Total #Annotations | Parasite Species | Blood Smear | Staining Protocol |
|---|---|---|---|---|---|
| V1 | 610 | 8,697 | *T. brucei brucei* | Thick | Unstained |
| V2 | 610 | 9,914 | *T. brucei brucei* | Thick | Unstained |
| V3 | 724 | 10,474 | *T. brucei brucei*, *T. cruzi* | Thin, Thick | Stained, Unstained |
| Near-OOD | 2,467 | 0 | None (*Plasmodium*, *M. tuberculosis*) | Thin, Thick | Stained, Unstained |

RetinaNet-r18, RTMDet-small, RT-DETRv2-r18) and larger variants (Faster-RCNN-r50, RTMDet-tiny, RT-DETRv2-r50). ResNet backbones (*r18*, *r50*) follow [6], while *small* and *tiny* refer to the RTMDet variants introduced in [11].

**Training Procedure.** All models are trained on each dataset version (V1–V3) by fine-tuning models pretrained on MS-COCO dataset [9], using the default settings of publicly available versions of their implementations. Faster-RCNN, RetinaNet, and RTMDet were trained based on the MMDetection [3] implementation, while the original author's default implementation [10] was used to train RT-DETRv2. For each model, the training parameters and procedure are the same across dataset versions, and the best model checkpoint saved is based on the best mAP on the validation set during the training epochs.

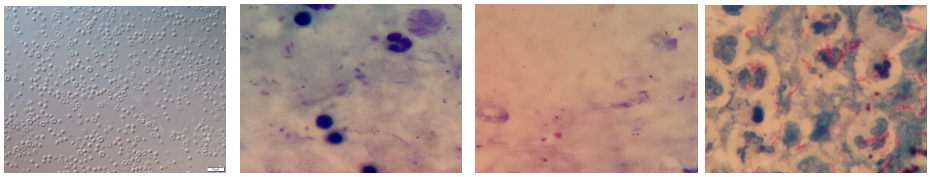
**Evaluation Methodology.** Evaluation spans both accuracy and uncertainty-based metrics to capture model behavior under distributional shift. Two test sets are used: (i) a labeled i.i.d. set representing the same distribution as the training data, and (ii) a curated near-OOD set, composed of microscopy images collected under similar conditions but confirmed to lack trypanosome parasites. While i.i.d. predictions may include true and false positives, near-OOD predictions are expected to be sparse and low-confidence.

We report standard detection metrics (precision, recall, mAP, F1 Score) on the i.i.d. test set. To assess uncertainty quality, we compute Expected Calibration Error (ECE), Area Under the ROC Curve (AUROC), Area Under Precision/Recall Curve (AUPR-in, AUPR-out), and negative log-likelihood (NLL). AUPR-in treats i.i.d. predictions as positives, while AUPR-out treats near-OOD predictions as positives to quantify separability. NLL penalizes overconfident incorrect predictions, particularly useful under distribution shifts where there are no objects of interest, and we do not expect any predictions. Together, these metrics evaluate both detection quality and the structure of predictive confidence.



(a) Dataset V1 (b) Dataset V2 (c) Dataset V3

Fig. 2: Example images from the three dataset versions. All blue and green boxes represent valid annotations. Green boxes highlight annotation differences between V1 and V2, which share the same underlying images but differ in annotation quality.



(a) Trypanosome (b) Plasmodium (c) Plasmodium (d) Tuberculosis

Fig. 3: An example image from each of the dataset sources that constitutes the near-OOD test dataset.

## 4 Results and Analyses

Lightweight and larger backbone variants show similar trends, with lightweight models slightly outperforming. We therefore report results mainly for lightweight models.

***Diverging Trends Without a Clear Winner.*** Figure 4 shows that no model consistently excels across all metrics. Standard detection metrics (mAP, Precision, Recall, F1) reveal performance trade-offs, while uncertainty metrics (AUROC, AUPR, ECE) capture differences in calibration and ranking quality, underscoring the need for multidimensional evaluation to assess reliability across i.i.d. and near-OOD scenarios. Top-row metrics generally improve across dataset versions, reflecting better in-distribution detection, but fail to capture changes in confidence behavior under near-OOD shifts. This limitation is evident in the bottom-row metrics, where several models show a noticeable decline in AUPR-out in V3, likely due to increased visual similarity between near-OOD samples

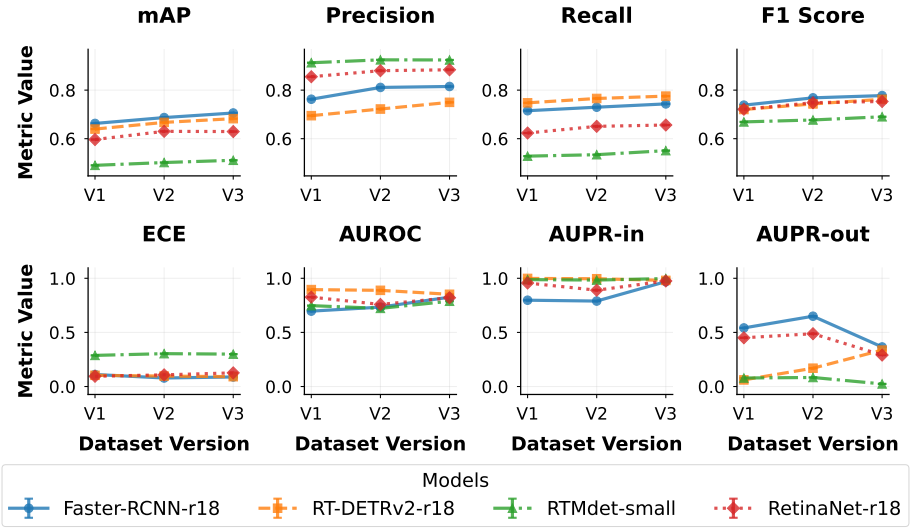



Fig. 4: Top row: standard i.i.d. evaluation metrics for detection performance, including mAP, Precision, Recall, and F1 Score. Bottom row: reliability-oriented metrics, with ECE measuring calibration on the i.i.d. set, and AUROC, AUPR-in, and AUPR-out assessing OOD detection performance. For all metrics, higher values indicate better performance, except for ECE, where lower is better.

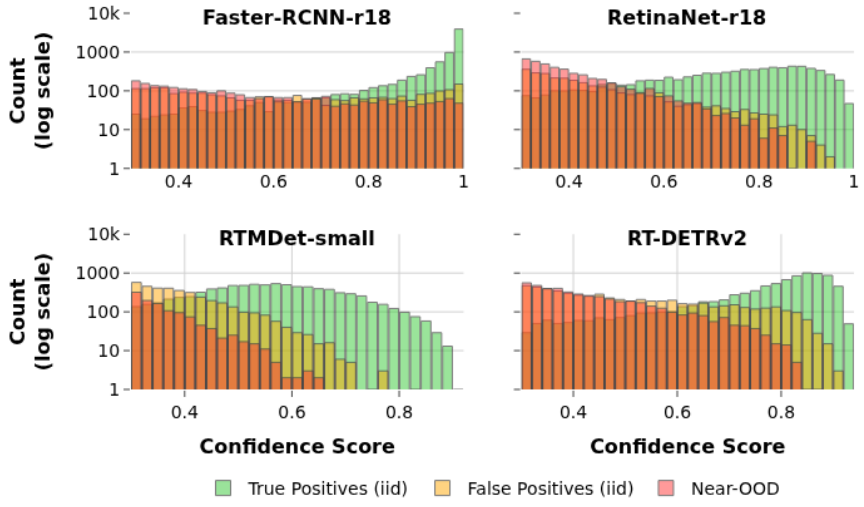
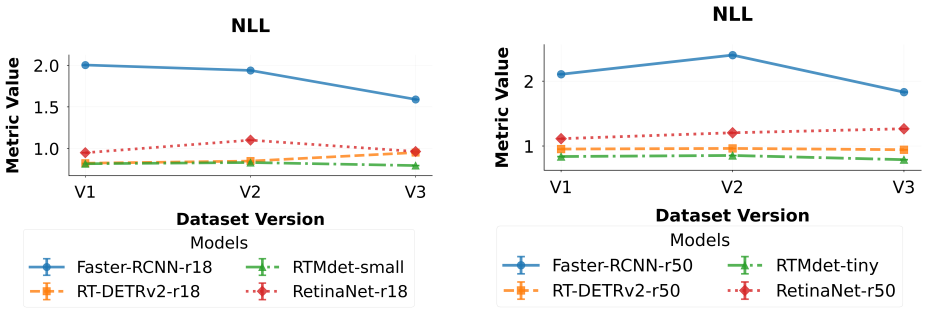



Fig. 5: Distribution of prediction confidence scores for i.i.d. true positives, i.i.d. false positives, and near-OOD predictions.




Small backbones Larger backbones

Fig. 6: NLL for all models across dataset versions. Lower values indicate better uncertainty calibration on near-OOD samples.

and *T. cruzi* images, which hinders confident rejection of negatives. These results highlight that accuracy metrics alone are insufficient for diagnosing reliability issues. While uncertainty-aware metrics help quantify confidence–correctness alignment and in-/out-distribution separability, they remain insensitive to the severity of miscalibrations and may overlook rare but critical overconfidence. This shortcoming is exemplified in Figure 5, where Faster-RCNN-r18 produces highly confident false positives on near-OOD samples despite favorable scores on mAP, F1, ECE, and AUPR-out.

***The Benefit of NLL as a Complementary Metric.*** None of the earlier reported metrics adequately penalize overconfident incorrect predictions. As shown in Figure 5, models like Faster-RCNN-r18 assign high confidence to false positives and near-OOD inputs while still achieving reasonable scores on mAP, F1 Score, and AUPR-out. To better capture this failure mode, we report the NLL on near-OOD datapoints, where no detections are expected. As shown in Figure 6, Faster-RCNN models have the highest NLL values, as NLL penalizes over-confident predictions. Based on Figure 4, Figure 5, and Figure 6, RTMDet and RT-DETRv2 models demonstrate more desirable uncertainty behavior than the other evaluated models, even though they do not consistently rank best in the in-distribution evaluations.

## 5 Conclusions

In data-scarce domains, iterative dataset improvements are often necessary but complicate model evaluation. We show that standard object detection and OOD metrics miss key reliability issues, including high-confidence in-distribution errors, poor in-/out-distribution separation, and overconfidence on negative samples. By augmenting these metrics with NLL and confidence histograms, we reveal overlooked failure modes and more fully characterize model behavior. Our

results underscore the need for broader evaluation as datasets evolve. RTMDet and RT-DETRv2 exhibit more desirable uncertainty behavior alongside strong accuracy, making them promising for real-world diagnostics. RetinaNet also balances accuracy and uncertainty effectively. Though designed for data-scarce settings, our evaluation approach generalizes to richer domains and enables more comprehensive assessment.

## References


1. Automated Blood Smear Analysis for Mobile Malaria Diagnosis. In: Karlen, W. (ed.) Mobile Point-of-Care Monitors and Diagnostic Device Design, pp. 132–149. CRC Press, 0 edn. (2018)
2. Anzaku, E.T., Mohammed, M.A., Ozbulak, U., Won, J., Hong, H., Krishnamoorthy, J., Van Hoecke, S., Magez, S., Van Messem, A., De Neve, W.: Tryp: a dataset of microscopy images of unstained thick blood smears for trypanosome detection. Scientific Data **10**(1), 716 (Oct 2023). https://doi.org/10.1038/s41597-023-02608-y
3. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open MMLab Detection Toolbox and Benchmark (2019). https://doi.org/10.48550/arXiv.1906.07155
4. Compton, R., Zhang, L., Puli, A., Ranganath, R.: When More is Less: Incorporating Additional Datasets Can Hurt Performance By Introducing Spurious Correlations. In: Proceedings of the 8th Machine Learning for Healthcare Conference. pp. 110–127 (2023)
5. Ewnetu, Y., Badu, K., Carlier, L., Vera-Arias, C.A., Troth, E.V., Mutala, A.H., Afriyie, S.O., Addison, T.K., Berhane, N., Lemma, W., Koepfli, C.: A digital microscope for the diagnosis of Plasmodium falciparum and Plasmodium vivax, including P. falciparum with hrp2/hrp3 deletion. PLOS global public health **4**(5), e0003091 (2024). https://doi.org/10.1371/journal.pgph.0003091
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
7. Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M.W., Wiens, J.: Deep Learning Applied to Chest X-Rays: Exploiting and Preventing Shortcuts. In: Proceedings of the 5th Machine Learning for Healthcare Conference. pp. 750–782 (2020)
8. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal Loss for Dense Object Detection. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2999–3007 (2017)
9. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) European Conference on Computer Vision. pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48
10. Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., Liu, Y.: RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer (2024). https://doi.org/10.48550/arXiv.2407.17140
11. Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen, K.: RTMDet: An Empirical Study of Designing Real-Time Object Detectors (2022). https://doi.org/10.48550/arXiv.2212.07784

12. Martins, G.L., Ferreira, D.S., Ramalho, G.L.B.: Collateral motion saliency-based model for *Trypanosoma cruzi* detection in dye-free blood microscopy. Computers in Biology and Medicine **132**, 104220 (2021). https://doi.org/10.1016/j.compbiomed.2021.104220
13. Morais, M.C.C., Silva, D., Milagre, M.M., Oliveira, M.T.d., Pereira, T., Silva, J.S., Costa, L.d.F., Minoprio, P., Junior, R.M.C., Gazzinelli, R., Lana, M.d., Nakaya, H.I.: Automatic detection of the parasite Trypanosoma cruzi in blood smears using a machine learning approach applied to mobile phone images. PeerJ **10**, e13470 (2022). https://doi.org/10.7717/peerj.13470
14. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet Classifiers Generalize to ImageNet? In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 5389–5400 (2019)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Advances in Neural Information Processing Systems. vol. 28 (2015)
16. Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C.S., Koitka, S., Pelka, O., Abacha, A.B., Herrera, A.G.S.d., Müller, H., Horn, P.A., Nensa, F., Friedrich, C.M.: ROCOv2: Radiology Objects in COntext Version 2, An Updated Multimodal Image Dataset (2023), https://zenodo.org/records/8333645
17. Ward, P., Dahlberg, P., Lagatie, O., Larsson, J., Tynong, A., Vlaminck, J., Zumpe, M., Ame, S., Ayana, M., Khieu, V., Mekonnen, Z., Odiere, M., Yohannes, T., Hoecke, S.V., Levecke, B., Stuyver, L.J.: Affordable artificial intelligence-based digital pathology for neglected tropical diseases: A proof-of-concept for the detection of soil-transmitted helminths and Schistosoma mansoni eggs in Kato-Katz stool thick smears. PLOS Neglected Tropical Diseases **16**(6), e0010500 (2022). https://doi.org/10.1371/journal.pntd.0010500
18. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine **15**(11), e1002683 (2018). https://doi.org/10.1371/journal.pmed.1002683
19. Zhang, C., Jiang, H., Jiang, H., Xi, H., Chen, B., Liu, Y., Juhas, M., Li, J., Zhang, Y.: Deep learning for microscopic examination of protozoan parasites. Computational and Structural Biotechnology Journal **20**, 1036–1043 (2022). https://doi.org/10.1016/j.csbj.2022.02.005
20. Zhang, Y., Jiang, H., Ye, T., Juhas, M.: Deep Learning for Imaging and Detection of Microorganisms. Trends in Microbiology **29**(7), 569–572 (2021). https://doi.org/10.1016/j.tim.2021.01.006