Hierarchical Deep Learning for Malaria Diagnosis: Integrating YOLO-Based Detection and Uncertainty-Aware Cell Classification

Luciano Luz Beylouni Farias[1]*, **Mateus Balda Mota**[1]*, **Karin Becker**[1], **Mariana Recamonde-Mendoza**[1]

[1]Institute of Informatics – Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre - RS, Brazil

{llbfarias,mateus.balda,kbecker,mrmendoza}@inf.ufrgs.br

Abstract. *Malaria diagnosis through optical microscopy of blood smears remains the gold standard but is labor-intensive and subject to inter-observer variability. This study proposes a hierarchical deep learning pipeline that integrates cell detection in microscope field of view (FoV) images with classification of individual red blood cells. YOLOv8 is used to detect candidate cells and the extracted regions are classified by convolutional neural networks with uncertainty quantification and Grad-CAM explainability. MobileNetV2 achieved 95.17% accuracy in the classification of isolated cells with efficient prediction sets. However, under domain shift conditions, conformal prediction revealed a drop in reliability (coverage from 95% to approximately 86%), indicating reduced generalization. Grad-CAM analysis also revealed the Clever Hans effect, where models relied on image artifacts rather than biological features. These results highlight the importance of integrating reliability and explainability mechanisms for computer-aided malaria diagnosis.*

1. Introduction

Malaria remains one of the major global public health challenges, with approximately 263 million cases and 597 thousand deaths reported annually [World Health Organization 2025]. It is an infectious disease caused by protozoa of the genus *Plasmodium*, transmitted through the bite of infected *Anopheles* mosquitoes. Among the species responsible for human malaria, *Plasmodium vivax* is one of the most prevalent in several endemic regions. Its intraerythrocytic developmental stages include *ring*, *trophozoite*, *schizont*, and *gametocyte*, each presenting distinct morphological characteristics observable in Giemsa-stained thin blood smears [Voß et al. 2023].

Optical microscopy of stained blood smears remains the gold standard for malaria diagnosis. However, this process is highly dependent on expert analysis, requiring the manual identification of infected red blood cells (RBCs) among thousands of cells present in a single slide. This task is labor-intensive, subject to inter-observer variability, and particularly challenging in regions with limited access to trained specialists [Tangpukdee et al. 2009]. In this context, Computer-Aided Diagnosis (CAD) approaches

*Luciano Luz Beylouni Farias and Mateus Balda Mota contributed equally to this work.

based on deep learning have emerged as promising tools to support microscopic analysis [Doi 2007], although several challenges remain.

Many existing studies focus primarily on either parasite detection in microscope field of view (FoV) images or classification of individually segmented RBCs. However, the real diagnostic workflow involves multiple interdependent steps, ranging from locating candidate cells to determining infection status. Approaches that integrate these stages while also providing reliable predictions remain limited. Although some studies have proposed multi-stage pipelines combining detection and classification, most works concentrate on improving detection or classification accuracy, with limited attention to prediction reliability and uncertainty quantification. In addition, differences in staining conditions, imaging devices, and parasite species often introduce domain shifts between datasets, which may affect model reliability in practical deployments.

To address these limitations, this work proposes a hierarchical pipeline for automated malaria diagnosis in thin blood smear images. The approach consists of two complementary stages. First, slide-level detection is performed using YOLOv8 to automatically locate candidate RBCs in complete FoV smear images. Second, the detected regions of interest (ROIs) are classified using convolutional neural networks (CNNs), specifically MobileNetV2 and ResNet50V2, to determine the infection status of individual cells.

In addition to detection and classification, the proposed pipeline incorporates reliability and explainability mechanisms. Uncertainty quantification is introduced through *Conformal Prediction* [Angelopoulos and Bates 2021], which provides prediction sets with statistical coverage guarantees, while Grad-CAM highlights the image regions influencing model decisions. We also evaluate the pipeline under domain shift conditions by assessing prediction reliability when the classification model processes regions detected from a different dataset.

As its main contribution, this study presents an integrated architecture that combines detection, classification, uncertainty quantification, and explainability within a hierarchical workflow. This design enables the analysis of reliability under domain shift and supports the development of safer CAD systems.

2. Related Work

The literature on automated malaria diagnosis has explored different deep learning approaches to overcome the limitations of traditional microscopy. In the context of automatic detection in blood smear FoV images, [Yang et al. 2020] proposed a cascading system based on YOLOv2 followed by a classifier trained on false positives, achieving an increase of approximately 8% in mAP compared to the standalone detector, using 2,567 images from 171 patients, with the study being limited to the detection of P. vivax.

Regarding the classification of isolated cells, [Ramos et al. 2024] applied *transfer learning* to 6,222 images from the BBBC and FIOCRUZ datasets, achieving an AUC of 99.41% with DenseNet201. Similarly, [Asif et al. 2024] proposed MozzieNet, evaluated on the NIH dataset (27,558 images), obtaining an accuracy of 96.73% and an AUC of 99.35%, with explainability support through Grad-CAM. Despite the high performance reported, these studies focus on the classification of previously segmented cells and do not integrate the detection stage in FoV images.

From a different perspective, [Otesteanu et al. 2024] combined CNN and LSTM models to analyze temporal series of *Plasmodium berghei in vitro*, exploring spatio-temporal modeling, but without focusing on clinical diagnosis. More recently, [Ramos-Briceño et al. 2025] developed a multiclass CNN to distinguish *P. falciparum*, *P. vivax*, and leukocytes, achieving an accuracy of 99.51% .Studies using specialized architectures, such as networks with parallel soft attention (SPCNN) [Ahamed et al. 2025], have also reported precision, recall, and AUC values above 99%, reinforcing the potential of customized models with Grad-CAM-based explainability.

Overall, most studies address detection and classification separately, with limited integration into complete pipelines and scarce incorporation of formal uncertainty quantification mechanisms. In this context, the present work proposes a two-stage hierarchical approach that integrates YOLOv8-based detection, classification of individual red blood cells, uncertainty quantification, and Grad-CAM-based explainability.

3. Methodology

The proposed method follows a hierarchical design composed of two complementary stages: (i) detection of RBCs in microscopic smear images and (ii) classification of individually segmented RBCs. Each stage relies on specific datasets, preprocessing procedures, and modeling strategies. The stages are connected through the output of the YOLO detector, whose predicted bounding boxes are used to extract regions of interest that serve as input to the cell classification model. The final stage produces prediction confidence sets and Grad-CAM activation maps as outputs. The implementation of the proposed pipeline is publicly available in a repository[1]. Figure 1 provides an overview of the proposed pipeline, which will be explained in the following sections.

3.1. Datasets

Two public datasets were used in this study, each supporting a different stage of the proposed pipeline. The BBBC041 dataset was employed to train the detection model on full blood smear images, while the NIH dataset was used to train and evaluate the classification models on individual RBCs.

Detection dataset – BBBC041. For the detection stage, the Broad Institute Bioimage Benchmark Collection dataset BBBC041 (version 01) was used [Ljosa et al. 2012]. The dataset contains 1,364 images (.png and .jpg), divided into training (1,208 images) and validation (156 images), with approximately 86,000 cells. The images were obtained from three collections from Brazil and Southeast Asia and were stained using the Giemsa reagent. The dataset includes two categories of non-infected cells (RBCs and leukocytes) and four categories of infected by *Plasmodium vivax* (*gametocytes*, *rings*, *trophozoites*, and *schizonts*), in addition to a class labeled as difficult to annotate. The annotations were performed by specialists who defined bounding boxes and assigned the corresponding labels. The images have an average resolution of $1{,}219 \times 1{,}635 \pm 55 \times 104$ pixels, with bounding boxes of $123 \times 124 \pm 24 \times 24$ pixels. In total, the dataset contains 86,035 cells: 83,034 RBCs and 103 leukocytes (non-infected); 156 gametocytes, 522 rings, 1,584 trophozoites, 190 schizonts, and 446 cells labeled as

[1]https://github.com/mlab-inf-ufrgs/conformalized-xplainable-malaria

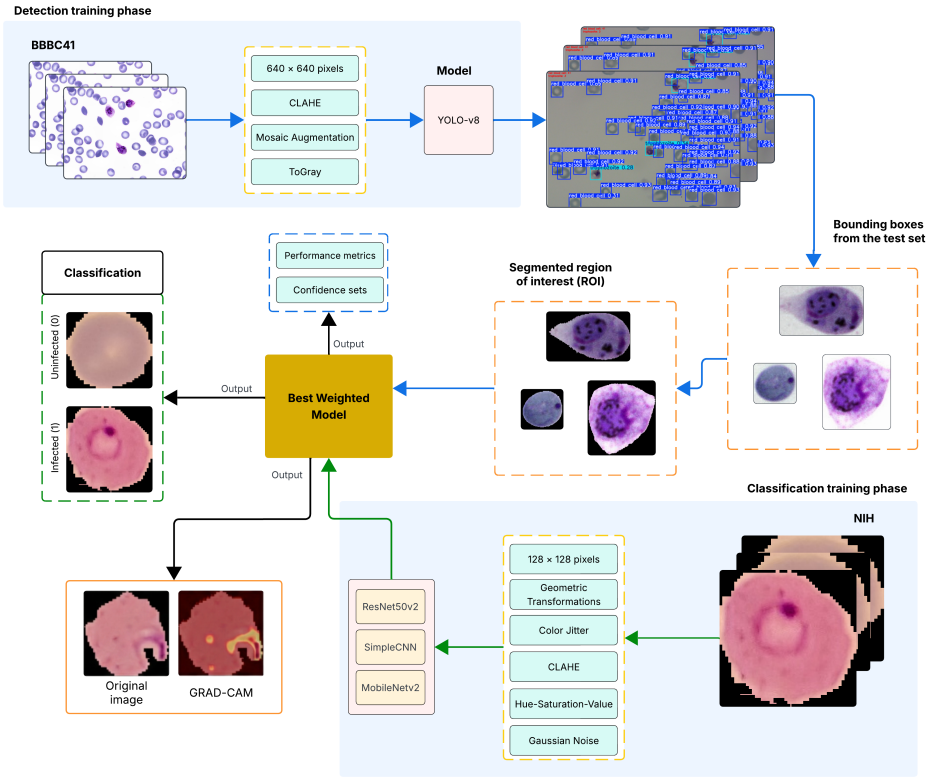


Figure 1. Overview of the proposed pipeline: the blue arrows indicate the detection and extraction of regions of interest to locate candidate RBCs, and the green arrows indicate individual RBCs classification.

difficult to annotate (infected). A strong class imbalance is observed, with approximately 3.37% infected cells and 96.63% non-infected cells.

Classification dataset – NIH. For the cell-level classification stage, a public dataset provided by the *National Institutes of Health* (NIH) was used, available through *TensorFlow Datasets*. The images were collected at the *Chittagong Medical College Hospital* in Bangladesh from thin blood smear samples of 150 patients infected with *Plasmodium falciparum* and 50 healthy patients. All samples were prepared and stained using *Giemsa*. The dataset contains 27,558 segmented images (patches) of individual RBCs obtained using a level-set algorithm followed by manual validation conducted by microscopy specialists. The classes are balanced, with 13,779 images for each class (infected and non-infected). The use of segmented images allows the analysis of morphological characteristics of the parasite while reducing background interference and cell overlap [Rajaraman et al. 2018].

3.2. Stage I: Detection of RBCs in Clinical Blood Smears

An exploratory analysis conducted on the training set of the BBBC041 dataset, composed of 1,208 images, showed that the total number of cells per image has an average of 66.31 ± 34.57, with a minimum value of 9 and a maximum of 223. The number of infected cells per image has an average of 2.63 ± 1.90, with a minimum of 1 cell and a maximum of 12 cells per image. The next paragraphs explain the steps of data preprocessing, model training, and extraction of ROIs to isolate individual cells.

Data Preprocessing and Augmentation. The images were downloaded from the *Broad Institute* website. The original annotations, provided in *.json* format, were processed and converted to the COCO (*Common Objects in Context*) format to standardize and facilitate label analysis. Subsequently, the COCO annotations were adapted to the input format required by the *YOLO* architecture. Prior to model training, the images were resized to 640×640 pixels. The *Albumentations* library was integrated into the pipeline for *data augmentation*, applying the following techniques: CLAHE (*Contrast Limited Adaptive Histogram Equalization*) with probability 0.01 to enhance edges; *Mosaic Augmentation* with probability 1.0 during training, combining four images into a single input to improve detection at different scales; and grayscale conversion (*ToGray*) with low probability to encourage the network to learn morphological characteristics. To avoid data leakage, data augmentation was applied exclusively in the training set.

Model Training and Evaluation. The model used for slide-level cell detection was YOLOv8 Small (YOLOv8s), composed of approximately 11.1 million parameters and 129 layers. Transfer learning was applied using weights from a model pre-trained on the COCO dataset (80 classes), adapted to the seven classes of the BBBC041 malaria dataset. Training was performed using the AdamW optimizer, with an initial learning rate ($lr0$) of 9.09×10^{-4}, momentum of 0.9, and weight decay of 5×10^{-4}. The model was trained for 200 epochs with a batch size of 16 images. Training was executed in deterministic mode (*deterministic=True*) with a fixed random seed. Model performance was evaluated using the mean Average Precision (mAP), specifically mAP@50 and the averaged mAP@50–95, which together compose the fitness index used for selecting the best model. During training, the optimization process considered three loss components: (1) box loss (IoU), which measures the overlap between the predicted and ground-truth bounding boxes; (2) classification loss (cls loss), computed using binary cross-entropy for each class; and (3) distribution focal loss (DFL), which refines bounding box boundaries when object edges are not clearly defined.

Extraction of Regions of Interest (ROIs). After training the YOLOv8 model, it was applied to the test set to detect bounding boxes. Only detections with confidence greater than 0.30 were retained, an empirical threshold chosen to increase recall for rare parasite classes while limiting excessive false positives. Due to severe class imbalance and lower confidence for underrepresented classes, only 5 detections of *schizont* and 10 of *gametocyte* were obtained. To reduce skew in the OOD test set, the number of samples per class was capped (*e.g.*, 20 for non-infected and 15 for infected). Without this filtering, the combination of class scarcity and detection noise would lead to unstable and unreliable uncertainty estimates. Finally, ROI segmentation and isolation were performed manually using LabelMe and Segment Anything Model 2 (SAM 2) to construct a clean *ground truth* for the OOD evaluation. This step is used only for experimental purposes and is not part of the proposed inference pipeline.

3.3. Stage II: Uncertainty-Aware Classification of Individual RBCs

Given the segmented ROIs, the second stage classifies individual cells as infected or non-infected while quantifying predictive uncertainty and providing visual explanations through Grad-CAM. Due to the characteristics of the dataset used for model training, this stage does not differentiate between the developmental stages of the parasite, although this could be implemented in the future by using multiclass datasets. The following sections

describe the steps of data preprocessing, model training and evaluation, and explainability of model outputs.

Data Preprocessing and Augmentation. The input images from the NIH Dataset, consisting of cropped patches of individual cells, were resized to dimensions of 128×128 pixels (benchmark) and 224×224 pixels (in-distribution). Image-wise normalization was applied to reduce local intensity variations. For *data augmentation*, the following techniques were applied: geometric transformations, *color jitter*, CLAHE (*Contrast Limited Adaptive Histogram Equalization*), adjustments in the HSV (*hue–saturation–value*) color space, and the addition of Gaussian noise. The dataset was divided into training (70%), validation (10%), calibration (10%), and test (10%) sets. The inclusion of a calibration set is required by the *Split Conformal Prediction* method. This set is used only to compute nonconformity scores and define confidence thresholds, ensuring statistical coverage without reusing training data.

Model Training and Evaluation. The ability of neural networks to extract morphological characteristics of parasites was evaluated through a binary classification task (infected vs. non-infected), comparing three architectures with different levels of complexity. The first architecture, SimpleCNN, is a shallow CNN developed in this study as a baseline model, composed of convolutional and max-pooling layers. The second architecture was MobileNetV2, designed for computational efficiency through depthwise separable convolutions and inverted residual blocks [Sandler et al. 2018]. Finally, ResNet50V2 was evaluated, which is a deep residual network that uses skip connections and pre-activation to mitigate gradient vanishing and capture more complex features [He et al. 2016]. Training was performed using the Adam optimizer [Kingma and Ba 2015] and binary cross-entropy loss for up to 40 epochs, with Early Stopping (patience of 5 epochs) to reduce overfitting.

Assessment of Model Reliability. To assess model reliability under different data conditions, two experimental scenarios were considered: in-distribution (ID) and out-of-distribution (OOD). In the ID setting, the classification models were trained, calibrated, and evaluated using images from the NIH dataset. In the OOD scenario, the trained models received ROIs detected by YOLO from the BBBC041 dataset. Crucially, the classification model was trained on *Plasmodium falciparum* (NIH) but tested on *Plasmodium vivax* (BBBC041). This species divergence is not an experimental flaw, but rather the intentional core of the biological *domain shift* evaluated in this study. Along with technical artifacts (such as staining protocols and cropping boundaries), morphological differences between species (e.g., cell deformation and gametocyte shape) introduce a severe domain shift, allowing for a rigorous evaluation of prediction reliability when the models operate outside their original training distribution.

Conformal Prediction and Explainability. To quantify predictive uncertainty, the *Split Conformal Prediction* method was adopted, implemented through the *Crepes* library [Boström 2022]. In contrast to standard classification, which returns a single class, this method produces prediction sets with a marginal coverage guarantee of 95% ($1 - \alpha = 0.95$). This allows the intrinsic reliability of model predictions to be assessed [Angelopoulos and Bates 2021]. Furthermore, explainability was analyzed using the *Gradient-weighted Class Activation Mapping* (Grad-CAM) technique [Selvaraju et al. 2017]. This method uses the gradients of the final convolutional layer

to generate heatmaps that highlight image regions that contribute most to the model's decision, allowing verification of whether the network focuses on parasite-related structures.

3.4. Experimental Setup

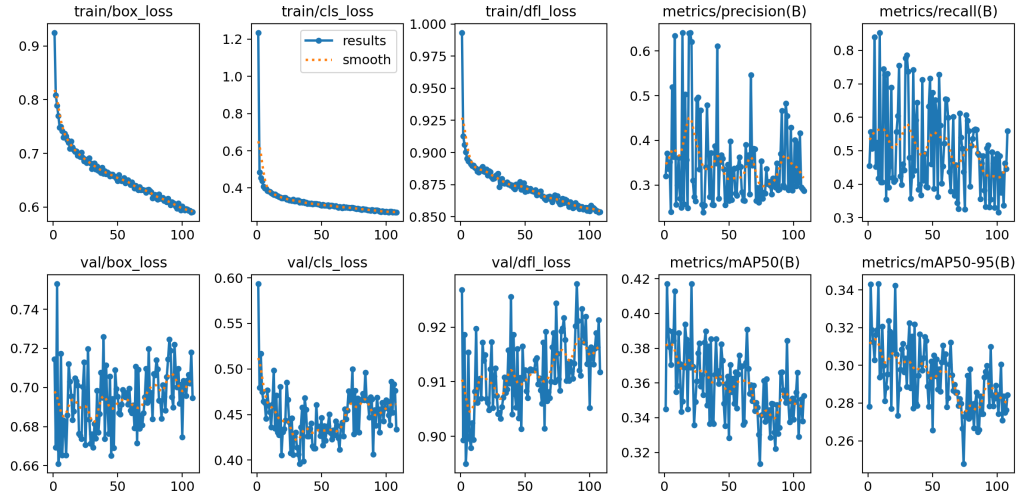
The experiments were conducted on a 12® generation Intel™ Core i7 processor (20 cores) equipped with an NVIDIA GeForce RTX™ 3060 GPU, as well as on an Apple M3 chip with 8 cores. Python version 3.12.7 was used for implementation. The main libraries included *crepes* (v. 0.9.0), *PyTorch*, *Ultralytics YOLOv8*, and *Grad-CAM* for Keras. Additional libraries such as *NumPy*, *Pandas*, *Matplotlib*, *Scikit-learn*, and NVIDIA CUDA were used to support data processing and model training.

4. Results and Discussion

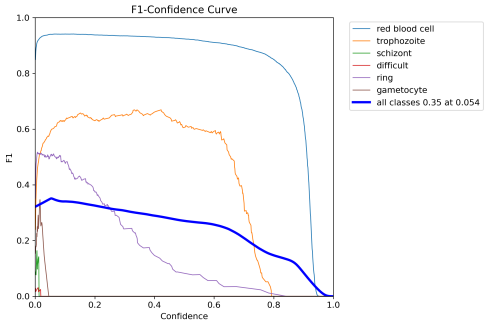
The results obtained are presented according to the hierarchical approach. First, the results of cells detection in clinical blood smears using YOLO are discussed, as described in Subsection 4.1. Next, the results of individual cell classification with CNNs, adopting uncertainty quantification and model explainability, are presented in Subsection 4.2.

4.1. Detection of Candidate Cells in Clinical Blood Smears with YOLO

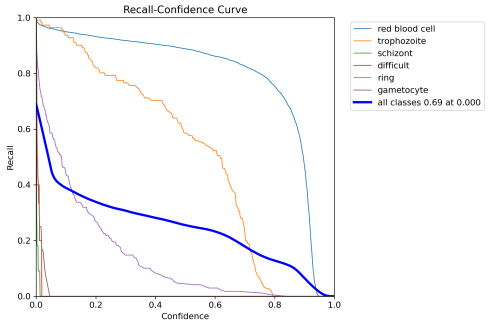
The training curves presented in Figure 2a indicate stable convergence of the YOLOv8 model throughout the process. The loss functions associated with the bounding box, the



(a) Overall metrics of the loss functions and mAP@50 and mAP@50–95.



(b) F1-score.



(c) Recall.

Figure 2. Performance analysis of the YOLO model in the validation set.

classification, and the distribution focal loss showed a consistent decreased during training, suggesting that the model progressively learned relevant spatial and semantic patterns from the dataset. In contrast, the validation losses exhibit higher variability across epochs, which may be related to the limited size and heterogeneity of the dataset. Regarding detection performance, the model achieved moderate values of mAP@50 and mAP@50–95, with peak values occurring during the early training epochs followed by a slight decline, which may indicate sensitivity to the validation set composition.

The class-wise performance curves shown in Figures 2b and 2c further highlight the effect of class imbalance in the dataset. RBCs achieved substantially higher F1-score and recall values when compared to parasite-related classes, indicating that the detector learn to identify RBC instances more reliably. The limited performances associated to parasite stages such as *ring*, *schizont*, and *gametocyte* is likely due to their limited representation in the training data and greater morphological variability. These results suggest that the YOLO model is effective in locating candidate cells in blood smear images but less reliable for directly discriminating parasite stages. This observation motivates the second stage of the proposed pipeline, which focuses on the classification of individual RBCs in terms of presence of absence of infection using dedicated CNNs.

4.2. Individual RBC Classification and Reliability Analysis

After the detection stage, the regions of interest corresponding to individual RBCs are extracted and used as input for the classification models. This stage focuses on determining the infection status of each RBC while also assessing prediction reliability through uncertainty quantification and visual explainability techniques.

4.2.1. Performance and Reliability Benchmark (ID vs. OOD)

The overall performance benchmark experiment on individual cells (Table 1) demonstrated a relationship between classification accuracy and efficiency in uncertainty quantification. **MobileNetV2** presented the most consistent performance, with the highest validation accuracy (95.17%) and the highest efficiency in conformal prediction (average set size = 1.0054). This value close to 1 indicates that, in order to guarantee 95% confidence, the model rarely needed to include both classes (infected and non-infected) simultaneously, suggesting well-defined decision margins.

In the *in-distribution* experiment (*i.e.*, NIH dataset), with *data augmentation* applied to prepare the model for the *domain shift* present in the pipeline, the models maintained robustness and coverage close to 95%. MobileNetV2 again showed the best performance, reaching 96.81% accuracy. However, when receiving the ROIs detected by YOLO in the *out-of-distribution* scenario, reliability showed a statistically significant decrease. The conformal prediction coverage dropped from approximately 95% to 86.67% for MobileNetV2 and to 81.33% for the other networks. This loss of calibration indicates a violation of the exchangeability assumption, caused by the *domain shift* resulting from morphological differences between *Plasmodium* species and preprocessing artifacts.

In traditional classification pipelines, this loss of generalization would likely remain unnoticed, since the *softmax* function tends to produce high-confidence point predictions even when operating on OOD data. In this context, conformal prediction acted as

a quantitative indicator of uncertainty, revealing the deviation in performance. In clinical decision-support scenarios where statistical guarantees are desirable (*e.g.*, a 95% coverage), the reduction in coverage signals that the model is operating outside its optimal domain, allowing uncertain cases to be directed to specialized human review.

Table 1. Comparative performance of the evaluated architectures: benchmark results and reliability analysis under In-Distribution (ID) and Out-of-Distribution (OOD) conditions.

	Benchmark				Reliability: ID vs OOD				
Model	**Val Loss**	**Val Acc**	**Cov. 95%**	**Set Size**	**Val Acc (ID)**	**Cov. (ID)**	**Size (ID)**	**Cov. (OOD)**	**Size (OOD)**
SimpleCNN	0.1710	93.94%	95.36%	1.0399	96.15%	95.32%	0.9935	81.33%	0.8933
MobileNetV2	0.1615	**95.17%**	95.28%	**1.0054**	**96.81%**	95.25%	0.9775	**86.67%**	0.9467
ResNet50V2	**0.1581**	94.52%	95.07%	1.0138	96.08%	95.21%	0.9909	81.33%	**0.9600**

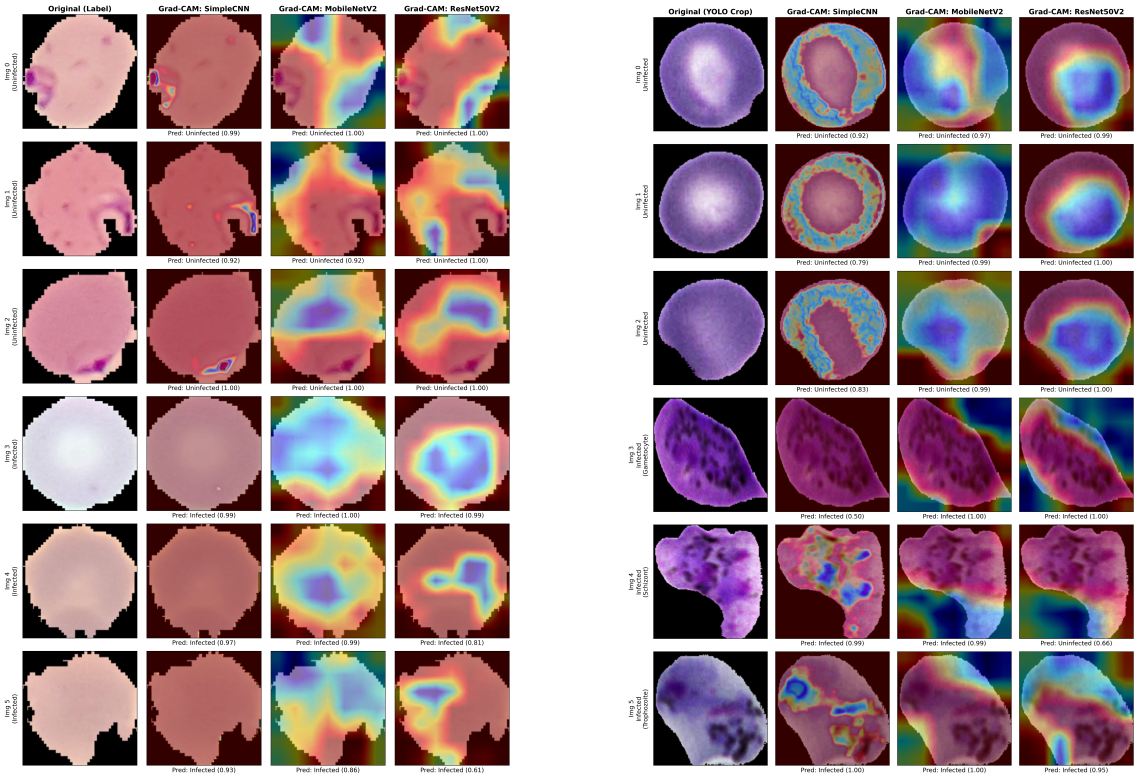
Val Loss = validation loss; *Val Acc* = validation accuracy; *Cov.* = conformal prediction coverage ($\alpha = 0.05$); *Set Size* = average prediction set size.

4.2.2. Explainability with Grad-CAM and the Clever Hans Effect

The qualitative analysis using Grad-CAM (Figure 3a) corroborated the quantitative metrics and supported the hypothesis of robustness under the ideal scenario. Differences in attention patterns were observed and appeared to be related to the depth of each network. For non-infected RBCs, SimpleCNN shows a predominantly homogeneous activation across the entire cell, whereas MobileNetV2 and ResNet50V2 distribute attention in a more diffuse manner. In infected RBCs, different behaviors are observed. Due to its shallow architecture, SimpleCNN still produces a homogeneous activation pattern, capturing a global "texture" of infection without isolating the parasite. MobileNetV2 focuses on broader regions, while ResNet50V2 shows greater spatial precision by concentrating activation directly on the area of intraerythrocytic infection.

In Figure 3b, the visual analysis helps explain the decrease in reliability in the *out-of-distribution* scenario. The *Clever Hans* effect, often reported in medical imaging, occurs when models rely on shortcuts in the dataset or processing pipeline rather than on biologically relevant features [Vásquez-Venegas et al. 2025]. SimpleCNN, due to its smaller receptive field, tends to focus on local textures. In healthy cells detected by YOLO, the network concentrates attention on the transition between the edge of the red blood cell and the black background, forming a ring-shaped activation pattern. In this case, the cropping artifact appears to be interpreted as the main factor for the prediction. For infected cells, although a global activation is observed, the network does not recognize the texture associated with *P. vivax*.

In contrast, deeper networks such as MobileNetV2 and ResNet50V2, which have a larger receptive field, tend to capture more global contextual information, including illumination patterns and the background of the original smear image. When receiving OOD crops generated by YOLO, these architectures exhibit a strong semantic mismatch. In infected cells, the activations shift toward the corners of the uniform black background, ignoring the central pathogen. This lack of biological correspondence, amplified by preprocessing effects, visually illustrates the impact of the *domain shift*. This behavior was captured by the safety mechanism provided by conformal prediction.



(a) In-Distribution (ID) **(b) Out-of-Distribution (OOD)**

Figure 3. Grad-CAM maps for the three evaluated architectures. In (a), the models correctly highlight morphological patterns in In-Distribution data. In (b), the *Clever Hans* effect appears in OOD data.

5. Threats to Validity

Although the results demonstrate the technical feasibility of the proposed approach, some limitations related to the experimental design and datasets may restrict its generalization to real clinical scenarios. First, the study relied exclusively on *Giemsa*-stained blood smears, which may limit performance when applied to slides prepared with different staining protocols or laboratory conditions. In addition, the images represent a single optical view and do not capture variations in focus or depth commonly present in manual microscopy. Regarding morphological representativeness, some samples exhibit low texture or faint staining, making the extraction of discriminative *features* more challenging, particularly in early parasite stages. Furthermore, the static nature of the NIH dataset prevents evaluating the model across the full progression of the intraerythrocytic cycle, during which cell morphology changes substantially. Finally, potential bias in the annotation (*ground truth*) process must be considered, as labels provided by human specialists are subject to subjectivity and possible errors, which may introduce uncertainty into the learned decision boundaries.

6. Conclusion and Future Work

This study presented a hierarchical deep learning pipeline for automated malaria diagnosis in blood smear images, integrating cell detection in full FoV images with classification of individual RBCs. The proposed architecture combines YOLOv8 for ROI detection and

CNNs for cell-level classification, complemented by uncertainty quantification through Conformal Prediction and visual interpretability via Grad-CAM.

The experimental results demonstrated strong performance in the classification of isolated RBCs, with MobileNetV2 achieving 95.17% accuracy and efficient conformal prediction sets. In the detection stage, YOLOv8 effectively identified candidate RBCs in full FoV images, although parasite-related classes remained challenging due to severe class imbalance in the dataset. More importantly, the reliability analysis revealed a substantial reduction in conformal coverage under out-of-distribution conditions, exposing the impact of domain shift between datasets. In this scenario, Conformal Prediction acted as a safety mechanism by signaling when the model operated outside its expected reliability regime, highlighting limitations that would likely remain unnoticed in conventional accuracy-based evaluations. This behavior illustrates how reliability-aware methods can help identify hidden failure modes in clinical AI systems.

These findings emphasize the importance of incorporating reliability-aware evaluation and interpretability mechanisms into computer-aided diagnostic pipelines for medical imaging. As future work, strategies to mitigate class imbalance in the detection stage and to reduce domain shift effects, such as domain adaptation techniques or training with multiple clinical datasets, may improve model generalization. Additionally, extending the classification stage to multiclass prediction of parasite developmental stages may enable more detailed clinical analyses and support disease monitoring.

Acknowledgments

This study was partially funded by CAPES - Finance Code 001, FAPERGS [Proj. No. 22/2551-0000390-7] and CNPq [Proj. No. 308075/2021-8].

References

Ahamed, M. F., Nahiduzzaman, M., Mahmud, G., Shafi, F. B., Ayari, M. A., Khandakar, A., Abdullah-Al-Wadud, M., and Islam, S. R. (2025). Improving malaria diagnosis through interpretable customized cnns architectures. *Scientific Reports*, 15(1):6484.

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Asif, S., Khan, S. U. R., Zheng, X., and Zhao, M. (2024). MozzieNet: A deep learning approach to efficiently detect malaria parasites in blood smear images. *International Journal of Imaging Systems and Technology*, 34(1):e22953.

Boström, H. (2022). crepes: a python package for generating conformal regressors and predictive systems. In *Conformal and Probabilistic Prediction with Applications*, pages 24–41. PMLR.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, pages 630–645. Springer.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Ljosa, V., Sokolnicki, K. L., and Carpenter, A. E. (2012). Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9(7):637.

Otesteanu, C. F., Caldelari, R., Heussler, V., and Sznitman, R. (2024). Machine learning for predicting plasmodium liver stage development in vitro using microscopy imaging. *Computational and Structural Biotechnology Journal*, 24:334–342.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568.

Ramos, J. d. S., Vieira, I. H. P., Rocha, W. S., Esquerdo, R. P., Watanabe, C. Y. V., and Zanchi, F. B. (2024). A transfer learning approach to identify plasmodium in microscopic images. *PLoS Computational Niology*, 20(8):e1012327.

Ramos-Briceño, D. A., Flammia-D'Aleo, A., Fernández-López, G., Carrión-Nessi, F. S., and Forero-Peña, D. A. (2025). Deep learning-based malaria parasite detection: convolutional neural networks model for accurate species identification of plasmodium falciparum and plasmodium vivax. *Scientific Reports*, 15(1):3746.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Tangpukdee, N., Duangdee, C., Wilairatana, P., and Krudsood, S. (2009). Malaria diagnosis: a brief review. *The Korean journal of parasitology*, 47(2):93.

Voß, Y., Klaus, S., Guizetti, J., and Ganter, M. (2023). Plasmodium schizogony, a chronology of the parasite's cell cycle in the blood stage. *PLoS Pathogens*, 19(3):e1011157.

Vásquez-Venegas, C., Wu, C., Sundar, S., Prôa, R., Beloy, F. J., Medina, J. R., McNichol, M., Parvataneni, K., Kurtzman, N., Mirshawka, F., Aguirre-Jerez, M., Ebner, D. K., and Celi, L. A. (2025). Detecting and mitigating the clever hans effect in medical imaging: A scoping review. *J Imaging Inform Med*.

World Health Organization (2025). Malaria. `https://www.who.int/news-room/fact-sheets/detail/malaria`. Accessed: 2025-12-02.

Yang, F., Quizon, N., Yu, H., Silamut, K., Maude, R. J., Jaeger, S., and Antani, S. (2020). Cascading YOLO: automated malaria parasite detection for plasmodium vivax in thin blood smears. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 404–410. SPIE.