# Multimodal Artificial Intelligence for Aided Diagnosis of Neglected Tropical Diseases: A Comparative Study of Visual Architectures

**Mário de Araújo Carvalho**[1], **Allison Oliveira Miranda**[2], **Celso Soares Costa**[3], **Wesley Nunes Gonçalves**[1]

[1]Faculdade de Computação, Universidade Federal de Mato Grosso do Sul (UFMS)
Campo Grande - MS - Brasil

[2]Santa Casa - Hospital Vale do Guaporé
Pontes e Lacerda - MT - Brasil

[3]Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso do Sul (IFMS)
Ponta Porã - MS - Brasil

{mario.carvalho, wesley.goncalves}@ufms.br

***Abstract.*** *Neglected Tropical Diseases (NTDs) impose a disproportionate burden on vulnerable populations worldwide and challenge public health systems in their efforts to provide prompt, accurate diagnoses. Artificial Intelligence (AI) and computer vision have emerged as promising tools for epidemiological surveillance and clinical decision support in resource-limited settings. This work compares recent visual architectures, ranging from traditional Convolutional Neural Networks (CNNs) to multimodal foundation models, for classifying parasitic pathogens in microscopy images. Five architectures were evaluated: ResNet-50, EfficientNet-B3, ViT-B/16, CLIP, and SigLIP. The dataset contains eight classes: Babesia, Leishmania, Plasmodium, Toxoplasma, Trichomonad, Trypanosome, Leukocytes, and Red Blood Cells (RBCs), and was split into training, validation, and test subsets following a stratified 70/15/15 partition. The analysis focuses on formally recognized NTDs, namely Leishmania and Trypanosome. Vision Transformers and multimodal baselines achieved near-perfect performance, with ViT-B/16 achieving 99.63% accuracy and 99.19% F1-Macro. The best-performing models were also deployed in a functional Progressive Web App (PWA) prototype, the CADTN Classifier, demonstrating cross-platform feasibility on desktop and mobile browsers. These results indicate that advanced computational models can strengthen health information systems and support clinical diagnostics in underserved environments. The source code is available at:* `https://github.com/MarioCarvalhoBr/multimodal-ntd-classifier`.

***Resumo.*** *As Doenças Tropicais Negligenciadas (DTNs) impõem um ônus desproporcional às populações vulneráveis em todo o mundo e desafiam os sistemas de saúde pública quanto ao diagnóstico oportuno e preciso. A Inteligência Artificial (IA) e a visão computacional surgiram como ferramentas promissoras para a vigilância epidemiológica e o apoio à decisão clínica em ambientes com recursos limitados. Este trabalho compara arquiteturas visuais recentes, abrangendo Redes Neurais Convolucionais (CNNs) tradicionais*

*e modelos de fundação multimodais, para classificar patógenos parasitários em imagens de microscopia. Cinco arquiteturas foram avaliadas: ResNet-50, EfficientNet-B3, ViT-B/16, CLIP e SigLIP. O conjunto de dados contém oito classes: Babesia, Leishmania, Plasmodium, Toxoplasma, Trichomonad, Trypanosome, Leucócitos e Hemácias (RBCs), e foi dividido em subconjuntos de treino, validação e teste segundo uma partição estratificada de 70/15/15. A análise enfatiza as DTNs formalmente reconhecidas: Leishmania e Trypanosome. Os Vision Transformers e os baselines multimodais alcançaram desempenho próximo ao perfeito, com o modelo ViT-B/16 atingindo 99,63% de acurácia e 99,19% de F1-Macro. Os modelos de melhor desempenho também foram disponibilizados em um protótipo funcional de Progressive Web App (PWA), o CADTN Classifier, demonstrando viabilidade multiplataforma em navegadores desktop e mobile. Esses resultados indicam que os modelos computacionais avançados podem fortalecer os sistemas de informação em saúde e apoiar o diagnóstico clínico em ambientes com recursos escassos. O código-fonte está disponível em:* `https://github.com/MarioCarvalhoBr/multimodal-ntd-classifier`.

## 1. Introduction

Neglected Tropical Diseases (NTDs) form a heterogeneous group of communicable infections that persist mainly in tropical and subtropical regions, disproportionately affect impoverished populations, and produce severe social, economic, and public health consequences [Organization et al. 2022]. According to the World Health Organization (WHO), over 1.6 billion people require interventions for at least one NTD annually, with the burden concentrated in regions where sanitation, surveillance, and access to specialized healthcare remain critically constrained [Organization et al. 2022]. Tackling these diseases demands transdisciplinary strategies in which AI, data science, and computational methods play an increasingly central role, supporting epidemiological surveillance, diagnosis, patient monitoring, and the design of public health policies [Soares and Chiavegatto Filho 2026, Sutton et al. 2020].

Microscopy remains the reference standard for diagnosing several NTDs and blood-borne parasitic infections, particularly malaria and trypanosomatid-borne diseases such as leishmaniasis and Chagas disease [Maturana et al. 2022,Organization et al. 2022]. Manual microscopic analysis, however, is laborious, time-consuming, and highly dependent on the expertise of trained microscopists, who are scarce in endemic and resource-constrained regions [Maturana et al. 2022]. Even when expert reading is available, sensitivity and specificity at peripheral health facilities fall well below those achieved in reference laboratories, especially for low-parasitemia samples and for species other than *Plasmodium falciparum* [Maturana et al. 2022]. The need for rapid and accurate diagnosis thus motivates automated computational systems that assist healthcare professionals and improve outcomes at the population level [Kumar et al. 2024].

Deep learning has advanced considerably in medical image analysis [Litjens et al. 2017], yet much of the literature still concentrates on disease-specific models or CNNs alone. Recent advances in foundation models and Vision Transformers (ViTs) have demonstrated strong capacity for image understanding and feature extraction across medical domains, as self-attention captures global context that CNNs struggle to model due

to their inherently local receptive fields [Shamshad et al. 2023]. Despite this progress, the literature lacks systematic comparisons between traditional visual models and recent multimodal or transformer-based architectures for multi-class identification of NTD pathogens in microscopy [Shamshad et al. 2023, Kumar et al. 2024]. Within this gap, the present study uses a publicly available eight-class dataset [Li 2020, Alrefaei 2023] in which *Leishmania*, the causative agent of leishmaniasis, and *Trypanosome*, the agent of Chagas disease and human African trypanosomiasis, are formally classified by the WHO as NTDs [Organization et al. 2022]. *Plasmodium* (malaria), *Babesia*, *Toxoplasma*, and *Trichomonad* fall under broader frameworks for infectious diseases but share analogous epidemiological profiles [Maturana et al. 2022, Organization et al. 2022], while Leukocytes and RBCs were included as host-cell classes to force discrimination from the cellular background of real clinical smears [Litjens et al. 2017]. In Brazil, strengthening the Unified Health System (SUS) through technological innovation is a strategic priority for expanding diagnostic care in regions with scarce specialized personnel [Soares and Chiavegatto Filho 2026], and AI-driven Clinical Decision Support Systems (CDSS) for NTDs align with this agenda when grounded in validated models [Sutton et al. 2020, Litjens et al. 2017].

The objective of this study is to compare five visual architectures, ResNet-50, EfficientNet-B3, ViT-B/16, CLIP, and SigLIP, on a unified training, partition, and evaluation protocol over the eight target classes. The analysis emphasizes the reliable identification of the formally recognized NTDs *Leishmania* and *Trypanosome*, including precision/recall trade-offs and error modes. The best-performing model is then deployed as a functional cross-platform PWA, the CADTN Classifier, as a proof of concept for clinical decision support in resource-limited settings. Section 2 reviews related work; Sections 3 and 4 present the methodology and results; Section 5 present Web-based CDSS prototype; Section 6 discusses them; and Section 7 concludes.

## 2. Related Work

Among the works that use the same eight-class corpus of 34,298 microscopy images, [Kumar et al. 2024] evaluated ten deep transfer-learning architectures (VGG19, InceptionV3, ResNet variants, EfficientNet variants, MobileNetV2, Xception, DenseNet169, and InceptionResNetV2), tuning SGD, RMSprop, and Adam optimizers. The authors reported 99.96% accuracy with InceptionResNetV2 and Adam after a pipeline of RGB-to-grayscale conversion, Otsu thresholding, and watershed-based region-of-interest extraction; the dataset split is not explicitly specified, although the reported confusion matrices suggest a 90/10, with 90% for training and validation and 10% for the test partition. [Hosen et al. 2024] combined swarm intelligence with deep learning on the same dataset, framing parasite identification as a multi-class problem that benefits from pretrained CNN backbones and from optimization heuristics that push CNN-based classifiers toward high-accuracy regimes.

Beyond the 34,298-image benchmark, other studies have explored adjacent corpora and methodological variations. [Rajinikanth 2024] combined MobileNet with the Firefly metaheuristic for *Plasmodium* detection in blood-cell images for resource-constrained deployment. [Vijayakumar et al. 2025] fused features from EfficientNet variants to discriminate normal cells from malaria-infected ones, while [Abed and Waleed 2024] proposed a CNN-based pipeline for general microorganism classification. [Pa-

mungkas et al. 2026] examined a ViT for recognizing parasitic types in microscopy, reinforcing the recent shift toward transformer-based architectures, and [Vassalotti et al. 2026] demonstrated AI-powered blood-parasite detection on smartphone hardware, underscoring the translational relevance of the field.

Table 1 summarizes these representative contributions. Two methodological gaps emerge. First, the literature remains anchored on CNN-based pipelines, with few systematic evaluations of pure ViTs and even fewer of multimodal vision-language models such as CLIP and SigLIP in parasitic microscopy. Second, when transformers are explored, they are typically compared against a narrow set of CNN baselines rather than against multiple recent families under a unified protocol. This work addresses both gaps and couples the analysis with a functional cross-platform PWA prototype that delivers the best-performing model as a CDSS, an artifact not reported by prior studies on this corpus to the best of the authors' knowledge.

**Table 1. Representative recent works on AI-based parasitic microscopy classification. Methods refer to the principal architectural family evaluated by each study; reported metrics correspond to the best-performing configuration disclosed by the authors.**

| Study | Method (best model) | Dataset | Best Acc. |
|---|---|---|---|
| [Kumar et al. 2024] | Deep transfer learning (InceptionResNetV2 + Adam) | 34,298 imgs, 8 classes | 99.96% |
| [Hosen et al. 2024] | Four pretrained CNNs (best: XceptionNet) | 34,298 imgs, 10 classes | 94.89% |
| [Rajinikanth 2024] | MobileNet + Firefly Algorithm | *Plasmodium* blood cells | 99.50% |
| [Abed and Waleed 2024] | CNN-based pipeline | Microorganism imagery | 98.95% |
| [Vijayakumar et al. 2025] | EfficientNet with fused features | Blood-cell (malaria) imagery | 98.33% |
| [Pamungkas et al. 2026] | ViT for parasitic recognition | Parasitic microscopy, 8 classes | 99.70% |
| [Vassalotti et al. 2026] | Smartphone-deployed AI detector | Blood parasite imagery | 98.24% |
| **This work** | **ViT-B/16 + multimodal (CLIP, SigLIP) under unified protocol** | **34,298 imgs, 8 classes** | **99.63%** |

## 3. Materials and Methods

### 3.1. Dataset and Preprocessing Pipeline

This study uses the publicly available "Parasite Dataset: Leishmania, Plasmodium & Babesia" [Li 2020, Alrefaei 2023], containing 34,298 microscopic blood smear images across eight classes (*Babesia*, *Leishmania*, *Plasmodium*, *Toxoplasma*, *Trichomonad*, *Trypanosome*, Leukocytes, and RBCs) at 400X or 1000X magnification, captured under vary-

ing illumination, magnification, and staining conditions. Figure 1 shows representative samples from each class.

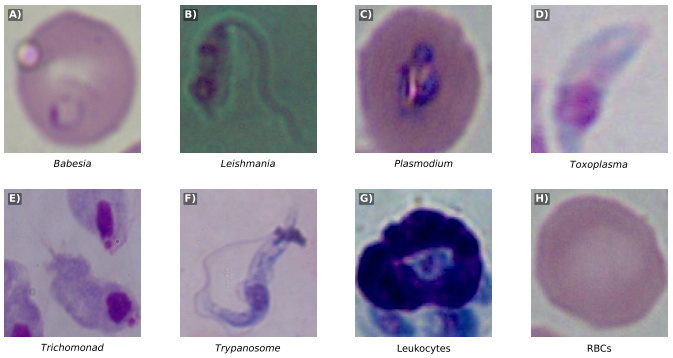



**Figure 1. Representative microscopic samples from each of the eight target classes, showing the morphological diversity across NTD pathogens (*Leishmania*, *Trypanosome*), other blood-borne parasites, and host cellular components (Leukocytes and RBCs).**

For unbiased evaluation, the dataset was partitioned into training, validation, and test subsets following a stratified 70/15/15 split that preserves the original class proportions (Table 2). An integrity audit identified no corrupted, empty, or otherwise invalid images. The dataset does not provide patient-level or slide-level identifiers, so the stratified split was performed at the image level. This follows the practice of recent works on the same benchmark [Kumar et al. 2024, Hosen et al. 2024, Pamungkas et al. 2026]; its methodological implications are discussed in Section 7.

**Table 2. Per-class distribution across training, validation, and test subsets under the stratified 70/15/15 split.**

| Class | Train | Validation | Test | Total |
|---|---|---|---|---|
| *Babesia* | 821 | 176 | 176 | 1,173 |
| *Leishmania* | 1,891 | 405 | 405 | 2,701 |
| *Plasmodium* | 590 | 126 | 127 | 843 |
| *Toxoplasma* | 4,684 | 1,004 | 1,003 | 6,691 |
| *Trichomonad* | 7,094 | 1,520 | 1,520 | 10,134 |
| *Trypanosome* | 1,670 | 358 | 357 | 2,385 |
| Leukocytes | 963 | 206 | 207 | 1,376 |
| RBCs | 6,296 | 1,349 | 1,350 | 8,995 |
| **Total** | **24,005** | **5,144** | **5,149** | **34,298** |

All images were processed through a standardized pipeline. They were converted to the RGB color space to eliminate inconsistencies arising from varying channel depths (such as hidden alpha channels or native grayscale formats), resized to $224 \times 224$ pixels (the standard input resolution for the ViT and CNN models adopted here), and finally normalized through model-specific *AutoImageProcessor* configurations from the Hugging Face library [Wolf et al. 2020], which applies Z-score scaling on pixel values.

### 3.2. Architectures Under Evaluation

Five architectures were compared, grouped into two categories. The first comprises traditional CNN baselines: ResNet-50 [He et al. 2016, Shafiq and Gu 2022] and EfficientNet-B3 [Tan and Le 2019], both widely recognized for hierarchical feature extraction in medical imaging. The second comprises recent foundation models: ViT-B/16 [Dosovitskiy et al. 2021], selected to assess the contribution of self-attention and global contextual processing without the inductive biases of convolution; and the multimodal CLIP (Contrastive Language-Image Pretraining) [Radford et al. 2021] and SigLIP (Sigmoid Loss for Language Image Pre-Training) [Zhai et al. 2023], whose visual encoders provide general feature representations advantageous for complex pattern recognition. Transfer learning was applied uniformly: the pretrained classification heads were replaced by a dense linear layer with eight-class output, and fine-tuning was restricted to the final layers.

### 3.3. Performance Evaluation

Models were assessed through Accuracy, F1-Macro score, and Cross-Entropy Loss. F1-Macro was prioritized because it weights all classes equally, regardless of frequency. The metrics are defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad F1 - Macro = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i}, \tag{1}$$

$$Loss = -\sum_{c=1}^{N} y_{o,c} \log(p_{o,c}), \tag{2}$$

where $TP$, $TN$, $FP$, and $FN$ denote true positives, true negatives, false positives, and false negatives; $N$ is the number of classes; and the loss is computed via Cross-Entropy. Confusion Matrices were also generated to inspect inter-class misclassification patterns.

### 3.4. Hardware, Software, and Training Dynamics

Experiments ran on an Ubuntu 22.04.5 LTS workstation with an AMD Ryzen 9 5950X CPU (16 cores / 32 threads), 128 GiB RAM, and 2×NVIDIA RTX 3080 GPUs (10 GB each, CUDA 12.4 / driver 550.120). The pipeline was implemented in PyTorch [Paszke et al. 2019] integrated with the Hugging Face *Transformers* library [Wolf et al. 2020]. All training runs used the Adam optimizer with a learning rate of $1 \times 10^{-3}$ and Cross-Entropy loss, with deterministic data loading for reproducibility. Restricting trainable parameters to the final classification layers, while keeping pretrained backbones frozen, led to rapid convergence within the first ten epochs, as shown in Figures 2 and 3. Training spanned 30 epochs in total, giving any latent overfitting tendency sufficient time to surface in the validation curves.

A joint inspection of both curves provides empirical evidence that the high performance reported here stems from genuine generalization rather than overfitting. Validation loss decreases monotonically and stabilizes at low values for all architectures, with no upward inflection that would signal memorization [Goodfellow et al. 2016]; the gap between training and validation accuracy remains narrow throughout, typically below 1.5 percentage points for the transformer-based models, a widely accepted indicator of low variance and adequate generalization [Ying 2019]; and the validation accuracy curves plateau after

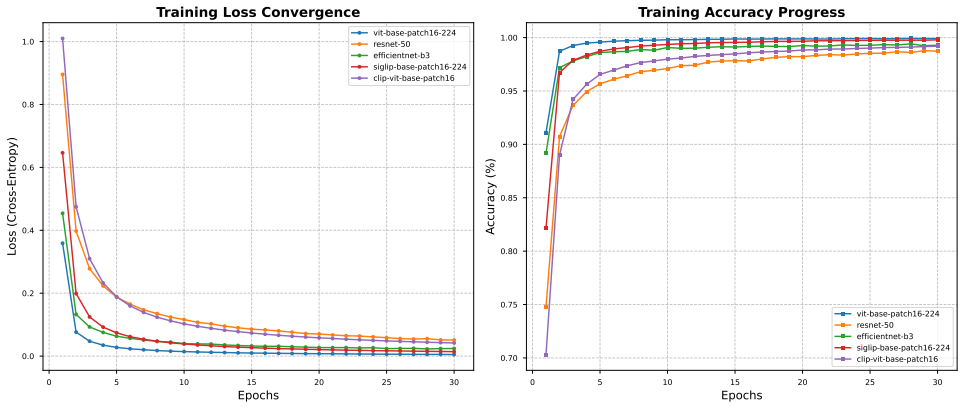



**Figure 2. Training dynamics across all architectures over 30 epochs. Left: Training Loss curves. Right: Training Accuracy evolution.**

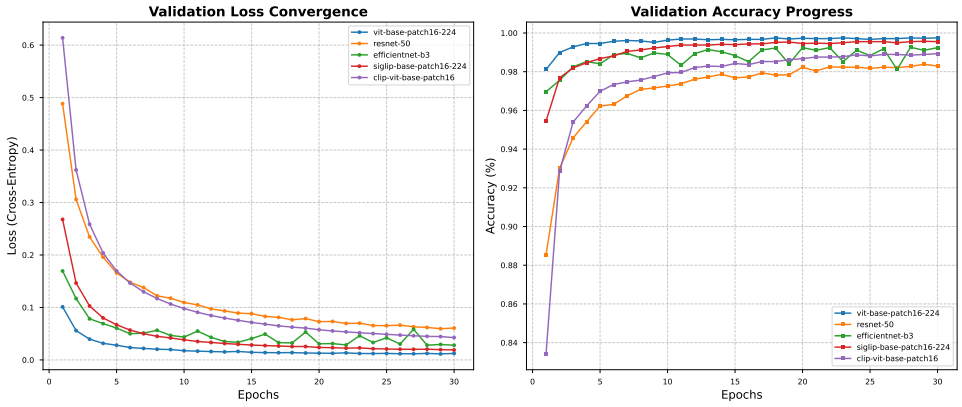



**Figure 3. Validation dynamics across all architectures over 30 epochs. Left: Validation Loss curves. Right: Validation Accuracy evolution. The convergence patterns mirror those of the training set without divergence, indicating effective generalization.**

roughly ten epochs rather than descend. Additional safeguards reinforce this interpretation: pretrained backbones with frozen weights and a single trainable linear head substantially limit the effective capacity available for fitting noise, since the number of trainable parameters is several orders of magnitude smaller than the total [Kornblith et al. 2019]; and the dataset is sufficiently large (34,298 samples) and well-balanced across partitions to mitigate statistical artifacts that often inflate metrics in smaller datasets.

## 4. Experimental Results

### 4.1. Ablation and Quantitative Results

The ablation analysis focused on the effect of freezing the backbone weights while training only the classification heads. Combining large-scale pretrained weights (ImageNet for CNNs and ViTs, LAION for CLIP and SigLIP) with a fixed learning rate of $1 \times 10^{-3}$ produced immediate training stability. ViT-B/16 and SigLIP converged fastest, reaching an accuracy plateau within the first three epochs, indicating that global features extracted via self-attention align with the morphological patterns of blood-borne parasites such as *Plasmodium* and *Trypanosome*. Ablating the multimodal text encoder, retaining only the

visual encoders of CLIP and SigLIP, did not degrade performance, indicating that visual embeddings alone are sufficiently rich for clinical microscopy [Radford et al. 2021, Zhai et al. 2023].

Table 3 reports performance on the independent test set (5,149 images, corresponding to 15% of the total data). Transformer-based models consistently outperform the convolutional baselines: ViT-B/16 achieves 99.63% accuracy and 99.19% F1-Macro, closely followed by SigLIP at 99.02% F1-Macro. Both surpass the ResNet-50 baseline (97.51% F1-Macro), indicating that attention-based architectures capture subtle visual differences between parasitic species more effectively.

**Table 3. Test-set classification performance across multimodal and visual architectures, evaluated over 30 training epochs.**

| Model | Accuracy (%) | F1-Macro (%) |
|---|---|---|
| google/vit-base-patch16-224 | 99.63 | 99.19 |
| google/siglip-base-patch16-224 | 99.57 | 99.02 |
| google/efficientnet-b3 | 99.13 | 98.47 |
| openai/clip-vit-base-patch16 | 99.05 | 98.31 |
| microsoft/resnet-50 | 98.39 | 97.51 |

### 4.2. Per-Class Performance and NTD Analysis

A closer inspection of per-class performance confirms that ViT-B/16 was the most robust model across the eight target categories, maintaining near-perfect separation between pathogens and host cellular structures. Table 4 reports precision, recall, F1-score, and support per class on the independent test set.

**Table 4. Per-class precision, recall, and F1-score of ViT-B/16 on the test set.**

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| *Trichomonad* | 1.0000 | 1.0000 | 1.0000 | 1,521 |
| *Babesia* | 1.0000 | 0.9943 | 0.9972 | 176 |
| RBCs | 0.9941 | 0.9993 | 0.9967 | 1,350 |
| *Toxoplasma* | 0.9941 | 0.9990 | 0.9965 | 1,004 |
| ***Trypanosome*** | 0.9972 | 0.9916 | 0.9944 | 358 |
| ***Leishmania*** | 0.9975 | 0.9901 | 0.9938 | 406 |
| Leukocytes | 0.9951 | 0.9903 | 0.9927 | 207 |
| *Plasmodium* | 0.9836 | 0.9449 | 0.9639 | 127 |
| **Macro avg** | **0.9952** | **0.9887** | **0.9919** | **5,149** |

F1-scores range from 0.9639 to a perfect 1.0000, with seven of the eight classes above 0.99. *Trichomonad* reached perfect precision and recall, consistent with its large training support and distinctive flagellated morphology. RBCs and Leukocytes were also recovered with F1-scores above 0.99, indicating reliable separation of pathological agents from the cellular background. *Plasmodium* is the most challenging class (F1 = 0.9639, recall = 0.9449), which is both the smallest in absolute terms (127 test images) and the

one most morphologically similar to *Babesia*; the resulting confusion is morphologically expected, given that both parasites occupy an intraerythrocytic ring stage [Maturana et al. 2022].

For the formally classified NTDs, ViT-B/16 reached F1-scores of 0.9944 for *Trypanosome* and 0.9938 for *Leishmania*, with precision values of 0.9972 and 0.9975, respectively. The high precision is particularly relevant for clinical decision support: when the model flags an image as one of these two NTDs, the prediction is almost always correct, minimizing the false-positive burden that would otherwise translate into unnecessary downstream investigations. The slightly lower recall (0.9916 and 0.9901) indicates that a small fraction of true cases is still missed; in real deployment, this trade-off can be managed by lowering the decision threshold or routing low-confidence samples to a human microscopist. Figure 4 shows representative cases: panels (A) and (B) display correctly classified *Leishmania* and *Trypanosome* samples with the typical morphological pattern clearly visible, while panels (C) and (D) show residual errors in which the parasite was confused with *Toxoplasma*, driven by atypical staining and out-of-focus acquisition (C) and degraded contrast that obscures the kinetoplast (D). These error modes point toward image-quality control as the most impactful intervention for closing the remaining performance gap.

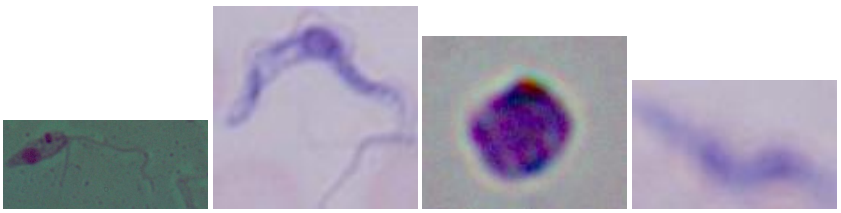



**Figure 4. Representative NTD samples. (A) *Leishmania* and (B) *Trypanosome* correctly identified by ViT-B/16. (C) *Leishmania* and (D) *Trypanosome* misclassified as *Toxoplasma*, with degraded image quality.**

Figure 5 shows the confusion matrix of ViT-B/16. The pronounced diagonal dominance confirms its high recall across all categories. The only systematic confusion involves *Babesia* and *Plasmodium*, expected given the intraerythrocytic ring-stage morphology shared by both parasites [Maturana et al. 2022]. Even so, ViT-B/16 reduced these errors more effectively than convolutional baselines such as ResNet-50, indicating that global contextual representations discriminate subtle inter-species differences better than local convolutional filters [Shamshad et al. 2023].

## 5. Prototype Clinical Decision Support Interface

To illustrate the practical applicability of the evaluated models, a web-based CDSS prototype, the CADTN Classifier, was developed using PWA technology. The hybrid nature of PWAs allows the application to run on any operating system with a browser (GNU/Linux, Android, iOS, Windows, and macOS) and to be deployed across computers, tablets, and smartphones. The interface lets clinicians and researchers submit a microscopic blood smear image and select any trained architecture from a dropdown menu; the system then performs real-time inference and returns a Top-5 probability ranking, with each prediction shown as a labeled confidence score and a horizontal bar. Each predicted class also

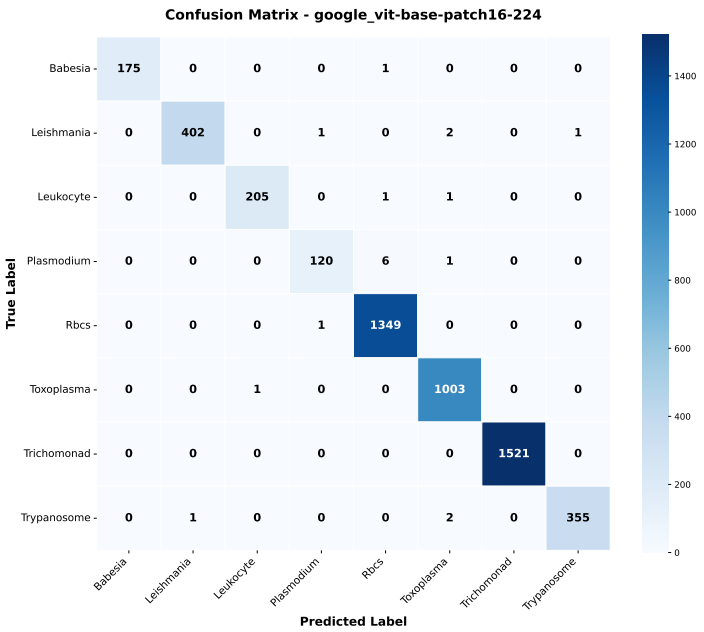


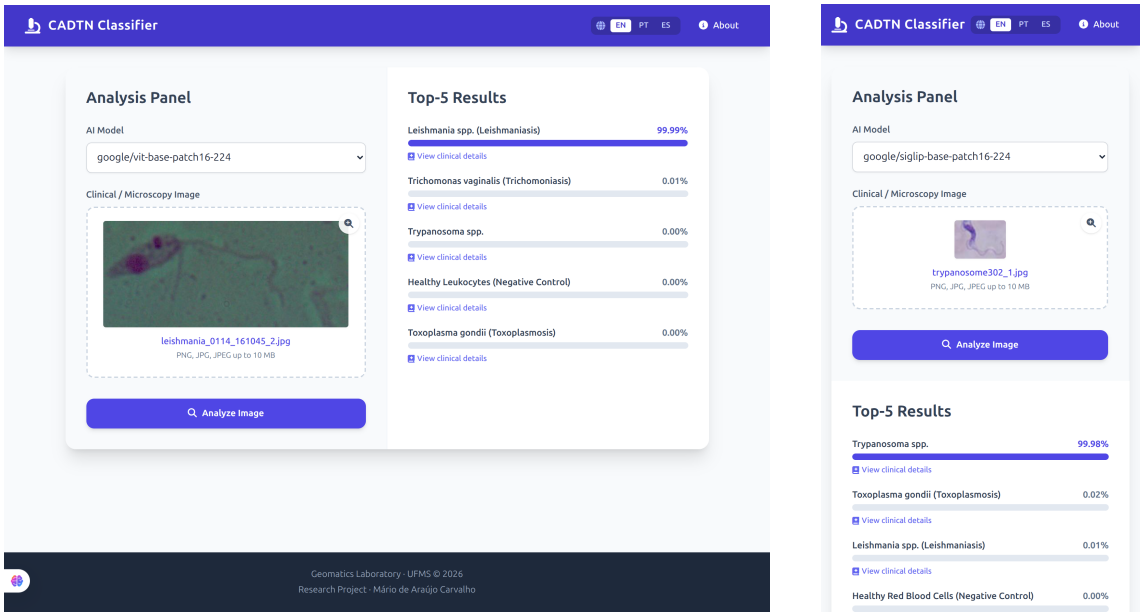

**Figure 5. Confusion matrix for ViT-B/16, the best-performing architecture. The near-perfect diagonal alignment reflects the model's high sensitivity and specificity for both formally classified NTDs and common blood-borne pathogens.**

provides supplementary clinical details to support diagnostic reasoning. Figure 6 shows both deployment modes: the desktop version (A) correctly classified a *Leishmania* sample with 99.99% confidence using ViT-B/16, and the mobile version (B) correctly classified a *Trypanosome* sample with 99.98% confidence using SigLIP, with negligible probability assigned to the remaining categories in both cases.

## 6. Discussion

The transition from CNNs to transformer-based architectures represents a substantive advance in automated blood-borne pathogen identification. ViT-B/16 and SigLIP both exceeded 99% accuracy and displayed greater training stability and faster convergence than ResNet-50. This gap is especially relevant for distinguishing formally classified NTDs such as *Leishmania* and *Trypanosome*, whose recognition requires capturing elongated flagellated bodies and a prominent kinetoplast, spatially extended features that benefit from the global context modeled by self-attention. The uniformly high F1-Macro scores confirm that the models managed both class imbalance and morphological nuance across all eight categories.

When set against the recent literature (Table 1), the 99.63% accuracy of ViT-B/16 is close to the 99.96% reported by [Kumar et al. 2024] for InceptionResNetV2 on the same dataset; the narrow margin indicates that, at this performance ceiling, architectural family matters less than the unification of preprocessing, partitioning, and metric reporting. ViT-B/16 reached this regime without the grayscale conversion and watershed-based region-of-interest extraction of [Kumar et al. 2024], suggesting that transformer-based encoders absorb part of the inductive work otherwise externalized in handcrafted preprocessing. Compared to [Hosen et al. 2024], which evaluated four CNNs, and [Ra-




**(a) Desktop version.** **(b) Mobile version.**

**Figure 6. Web-based prototype of the CADTN Classifier system, deployed as a Progressive Web App: (A) desktop browser layout and (B) mobile browser layout.**

jinikanth 2024], which examined MobileNet variants, this study spans CNNs, a pure ViT, and multimodal foundation models within a single protocol. The fact that SigLIP reaches 99.57% accuracy, within 0.06 percentage points of ViT-B/16, extends earlier observations on transformer effectiveness [Pamungkas et al. 2026] to the multimodal setting and indicates that visual encoders learned through contrastive image-text objectives transfer competitively to clinical microscopy, opening a path toward future integration of textual clinical metadata that complements emerging mobile-deployment efforts such as [Vassalotti et al. 2026].

For public health systems, particularly the SUS, integrating high-accuracy, transformer-based models into clinical workflows can act as a force multiplier in endemic regions where specialized microscopists are scarce, enabling faster decisions and more efficient epidemiological surveillance. By easing the diagnostic bottleneck, these approaches may bring earlier NTD treatment and, in turn, reduced morbidity and mortality [Organization et al. 2022]. The robustness of ViT-B/16 to staining artifacts and variable illumination further suggests viability for integration into low-cost digital microscopy setups, an avenue that the PWA prototype reported here begins to materialize.

## 7. Conclusion and Future Directions

This study compared five recent visual architectures for classifying parasitic pathogens and host cells from microscopy images. ViT-B/16 was the most effective model for the task, reaching 99.63% accuracy and 99.19% F1-Macro, and the best-performing models were made available through the CADTN Classifier, a functional cross-platform PWA prototype. Reliable identification of the formally classified NTDs *Leishmania* and *Trypanosome*, with F1-scores above 0.99 and precision close to 1.00, highlights the capacity of transformer-based deep learning to reshape diagnostic workflows in tropical medicine.

Several limitations deserve explicit acknowledgment. Sporadic misclassifications of *Babesia* as Leukocytes or RBCs, and of *Plasmodium* as *Babesia*, indicate that morphological mimicry remains a challenge, especially under extreme staining variation or overlapping cellular structures. The evaluation relied on static captured images rather than real-time video streams from operational microscopes, and the multi-class framework did not model co-infections involving multiple parasite species in the same sample. A further methodological concern is the potential risk of patient-level or slide-level data leakage between training and test partitions [Litjens et al. 2017, Shamshad et al. 2023]; partitioning was performed at the image level because the public benchmark [Li 2020, Alrefaei 2023] does not provide patient or slide identifiers, a constraint structural to the benchmark and inherited by all studies that adopt it [Kumar et al. 2024, Hosen et al. 2024, Pamungkas et al. 2026]. The absolute metrics reported across this body of literature, including those obtained here, should therefore be read as upper-bound estimates rather than direct predictors of clinical performance on unseen patient populations; the relative comparison across architectures, the primary methodological contribution of this work, remains valid since all five models were evaluated on identical partitions.

Future work will focus on three directions: adopting datasets with patient- and slide-level identifiers to enable group-aware validation and reduce data leakage risks [Litjens et al. 2017]; strengthening statistical validation through stratified cross-validation and appropriate post-hoc tests for classifier comparison; and assessing real-world deployment on mobile and edge devices, including active learning strategies and the integration of clinical metadata into the multimodal framework.

## Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This research was financed by the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq) - Finance Code (304052/2019–1, 310517/2020-6, and 405997/2021-3), the Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul - Brasil (FUNDECT), and CAPES PrInt (p: 88881.311850/2018–01, p: 88887.821508/2023-00 and 88887.710512/2022-00). The authors acknowledge the support of the UFMS (Federal University of Mato Grosso do Sul).

## Use of Generative Artificial Intelligence

The authors disclose that generative AI was used only in a supporting role during manuscript preparation and selected software implementation stages, while the scientific contribution remained under full authorial control. For textual preparation, Anthropic's Claude Opus 4.7 and Grammarly Proofreader supported grammar correction, syntactic refinement, lexical adjustments, and improvements in clarity, conciseness, and academic register. For software implementation, Claude Code and GitHub Copilot were used through prompt engineering to generate small, well-scoped code segments, all of which were analyzed, revised, integrated, and tested by the authors.

No generative AI tool was used to formulate the research questions, design the experimental protocol, run experiments, produce figures or tables, analyze data, or generate scientific findings. Medical and methodological content was additionally reviewed by

one author, a medical doctor, to ensure consistency with current biomedical knowledge. All conceptual, methodological, and interpretative content was conceived, executed, validated, and assumed as the authors' full responsibility.

## References


Abed, B. A. and Waleed, J. (2024). A developed deep learning-based approach for microorganism classification. In *2024 First International Conference on Software, Systems and Information Technology (SSITCON)*, pages 1–7. IEEE.

Alrefaei, A. (2023). Parasite dataset: Leishmania, plasmodium babesia.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hosen, M. D., Mohiuddin, A. B., Sarker, N., Sakib, M. S., Al Sakib, A., Dip, R. H., Ahmed, M. R., and Haque, R. (2024). Parasitology unveiled: Revolutionizing microorganism classification through deep learning. In *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, pages 1163–1168. IEEE.

Kornblith, S., Shlens, J., and Le, Q. V. (2019). Do better imagenet models transfer better? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2656–2666.

Kumar, Y., Garg, P., Moudgil, M. R., Singh, R., Woźniak, M., Shafi, J., and Ijaz, M. F. (2024). Enhancing parasitic organism detection in microscopy images through deep learning and fine-tuned optimizer. *Scientific Reports*, 14(1):5753.

Li, S. (2020). Microscopic Images of Parasites Species.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Maturana, C. R., De Oliveira, A. D., Nadal, S., Bilalli, B., Serrat, F. Z., Soley, M. E., Igual, E. S., Bosch, M., Lluch, A. V., Abelló, A., López-Codina, D., Suñé, T. P., Clols, E. S., and Joseph-Munné, J. (2022). Advances and challenges in automated malaria diagnosis using digital microscopy imaging with artificial intelligence tools: A review. *Frontiers in Microbiology*, 13:1006659.

Organization, W. H. et al. (2022). *Ending the neglect to attain the sustainable development goals: a rationale for continued investment in tackling neglected tropical diseases 2021–2030*. World Health Organization.

Pamungkas, Y., Triandini, E., Karim, A., and Sangsawang, T. (2026). A vision transformer architecture for automated recognition of parasitic types in microscopic images. *International Journal of Robotics and Control Systems*, 6(1):383–400.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Rajinikanth, V. (2024). Plasmodium detection from blood-cell images using mobilenet and firefly algorithm. In *2024 International Conference on Advances in Technology and Computing (ICATC)*, pages 1–5. IEEE.

Shafiq, M. and Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Applied sciences*, 12(18):8972.

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., and Fu, H. (2023). Transformers in medical imaging: A survey. *Medical image analysis*, 88:102802.

Soares, M. Q. and Chiavegatto Filho, A. D. P. (2026). Artificial intelligence in brazilian public health: potential, challenges, and ethical implications for the brazilian unified health system. *Revista Brasileira de Epidemiologia*, 29:e260006.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., and Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Vassalotti, R., Cervantes, J., D'Amario, V., and Colwell, S. (2026). P-1777. ai-powered blood parasite detection: A smartphone diagnostic tool for resource-limited settings. In *Open Forum Infectious Diseases*, volume 13, pages ofaf695–1947. Oxford University Press US.

Vijayakumar, K., Sivaraman, A., Narayanan, M., Rajinikanth, V., Jebastine, J., and Srivastava, A. (2025). Examination of blood cell images to identify nannal/malaria with efficientnet model and fused features. In *2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pages 1–6. IEEE.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2):022022.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.