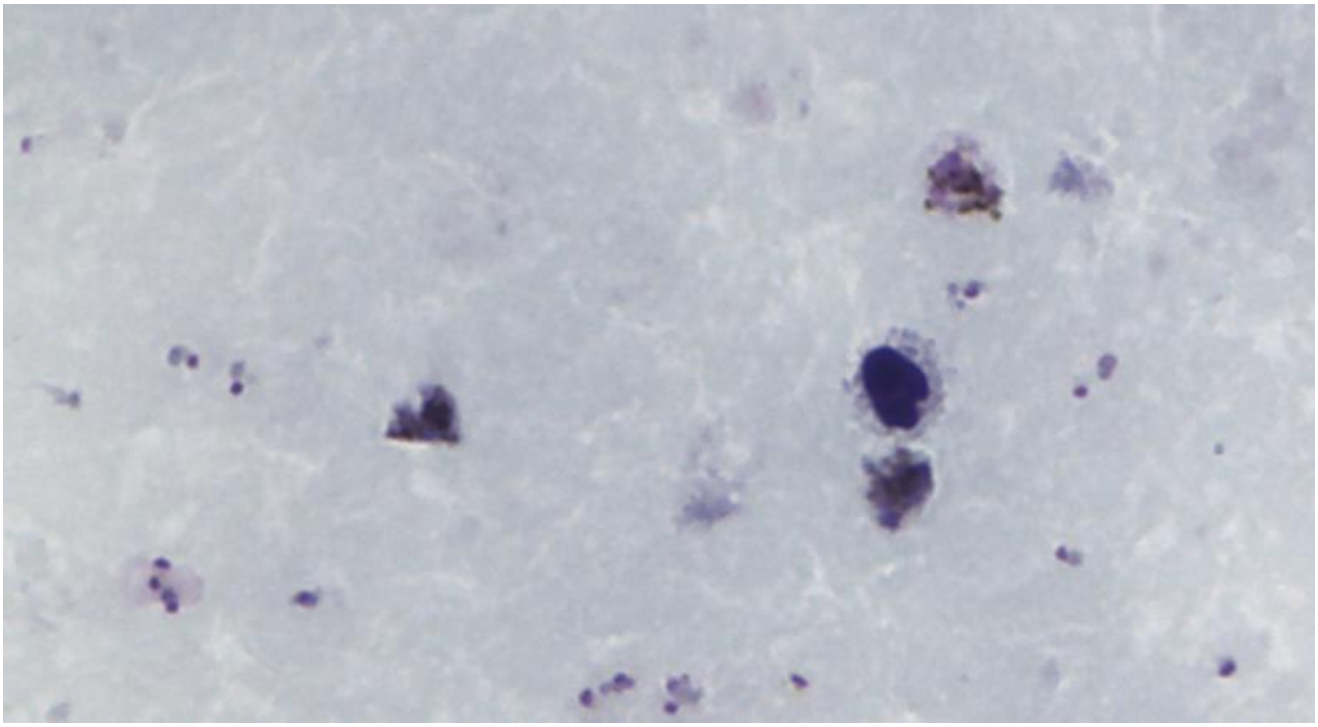
Malaria Giemsa-stained blood film images dataset release

This page will be updated as the release progresses. **Due for release in late 2026**



Statistics:

- Open access (train-val): 5000+ patients from 10+ sites, spanning Africa, South America, SE Asia, and returning travelers.
- Restricted (for challenges): up to 1000 patients, including from 13 additional sites in Africa and SE Asia.
- Thick films (as 7-layer stacks) and thin films
- Wide range of slide qualities and appearances
- 2 device types
- 13k image sets, ~12 TB.
- Each thick film contains > 0.1 uL of imaged blood (i.e. clinically relevant blood volume per patient)
- 1.2M positive Fields of View (FoVs)
- 420k positive FoVs from 1000 patients fully annotated at object level.
- Object annotations: 1.6M malaria parasites, 30k other species, 380k white and red blood cells
- 530k negative FoVs
- 6000 image tiles captured by both devices

- Enough data to fully support the current generation of ML / AI methods and models.
- Substantial holdout datasets, structured similarly to WHO evaluation protocols and including held-out sites, kept invisible to developers and used for seasonal challenges and to aid regulatory review.
- This is a greatly extended version of the dataset used to train the GH Labs models. These models are current State of the Art, evaluated in clinical studies in 16+ countries. Papers can be found at charlesdelahunt.github.io.

Clinical use cases for ML:

- Returning travelers
- Drug resistance sentinel sites
- Hospitals in endemic regions
- Reference and quality assurance labs
- Remote clinic support (upload an uncertain image for automated diagnosis/speciation assistance)
- White blood cell differential counts

Any questions, please email Charles Delahunt delahunt@uw.edu