# SGMCE: Segment-Grounded Morphological Concept Explanation for Malaria Parasite Species Identification in Thick Blood Smears

Ahmed Tahiru Issah[1], Charles B. Delahunt[2], and Carine Mukamakuza[1⋆]

[1] Carnegie Mellon University, Kigali, Rwanda
aissah@alumni.cmu.edu, cmukamak@andrew.cmu.edu
[2] University of Washington, Seattle, Washington
delahunt@uw.edu

**Abstract.** Malaria diagnosis in endemic regions depends on species-level identification of *Plasmodium* parasites in thick blood smears, but deep learning detectors classify detections without providing morphological evidence for their predictions, limiting the ability of microscopists to audit those predictions at the case level. We present SGMCE (Segment-Grounded Morphological Concept Explanation), a post-hoc explanation framework that requires no additional training, no morphological annotations, and no labelled explanation data, yet produces per-detection natural-language explanations anchored in thick-smear morphology. For each detection, SGMCE extracts mask-guided crop thumbnails, computes fourteen handcrafted computer-vision morphological features (shape, colour, chromatin, haemozoin pigment) using adaptive within-mask thresholds, and queries GPT-4o with both visual evidence and computed measurements, conditioned on a thick-smear-specific knowledge base compiled from the World Health Organization bench aids. The primary output is a structured explanation identifying which morphological features support the detected species and why the competing species are excluded. Explanations are validated by four automatic metrics: Knowledge-Base Consistency (KBC), CV-Claim Faithfulness (CCF), Discriminativeness Score (DS), and LLM-as-Judge (LLMj). A sentence-level semantic scoring rule with species-aware negation filtering resolves the vocabulary mismatch between clinical prose and knowledge-base terms. Across 737 detections from 139 thick-smear images spanning four *Plasmodium* species and white blood cells, parasite-class mean KBC is 0.91, mean DS is 0.99, and mean CCF is 0.97, while a per-rule CCF breakdown confirms that the CV-grounded claims made by the vision-language model are consistent with the measurements they cite.

**Keywords:** Malaria · Explainable AI · Thick blood smear · Species identification · Morphological analysis · Vision-language models · Medical image understanding

⋆ Corresponding author.

## 1 Introduction

Malaria, caused by *Plasmodium* parasites transmitted by female *Anopheles* mosquitoes, remains one of the most significant preventable causes of mortality worldwide, with roughly 249 million cases and 608,000 deaths estimated by the World Health Organization in 2022, and more than 90 % of global mortality concentrated in Sub-Saharan Africa [16]. Species-level identification is clinically important because treatment regimens differ across species: *P. falciparum* (Pf) requires artemisinin-based combination therapy and, if untreated, can progress to fatal cerebral malaria within hours; *P. vivax* (Pv) and *P. ovale* (Po) additionally require primaquine to clear dormant liver hypnozoites and prevent relapse; and *P. malariae* (Pm) typically responds to chloroquine alone [15, 2]. Thick blood smear (thick-film) microscopy remains the diagnostic gold standard in endemic regions because of its high sensitivity at low parasitaemia [9], but it requires expert microscopists, is time-consuming, and suffers from inter-observer variability [3].

Deep learning object detectors and segmentation models have demonstrated strong object-level accuracy on parasite detection in blood smears [10, 1, 4]. Object-level detection, locating and classifying individual parasites in a full field of view, is a key step in patient-level diagnosis, which requires aggregating object-level counts across a slide to estimate parasitaemia and infer infection status [3, 4]. A separate requisite for clinical utility, orthogonal to raw accuracy, is to ensure user trust by explaining *why* a detection is assigned to a particular species. This matters for microscopist-in-the-loop review, where a diagnostician auditing an automated prediction needs to know the logic behind the prediction (*e.g.,* whether the chromatin dots are single or double), and why the observed morphology excludes competing species.

Existing explainability methods such as Grad-CAM [12], LIME [11], and SHAP [6] all produce pixel-level saliency maps or attribution scores. These identify regions of interest but not the morphological reasoning behind a prediction. A saliency heatmap cannot indicate whether a bright region is a chromatin dot or a staining artefact, and it produces no differential argument against competing species. Concept-based approaches such as TCAV [5] explain model behaviour in terms of human-defined concepts, but require trained probes and do not generate per-detection free-form explanations. Closest to our setting, MorphXAI [17] couples parasite detection with fine-grained morphological analysis on *Leishmania* and *Trypanosoma* species and fills in a templated text, demonstrating that structured morphological attributes (shape, dot count, developmental stage) can be integrated into detection pipelines to produce explanations closer to the evidence clinicians rely on than pure saliency. Our work starts from the same premise - that morphology, rather than pixel intensity, is the right vocabulary for parasitology explanations, and extends it to *Plasmodium* speciation in thick films, replacing fixed template sentences with free-form reasoning from a vision-language model in a post-hoc, knowledge-base-grounded pipeline. We use GPT-4o [8] as the reasoning backbone. Recent work [7] reports strong zero-shot performance of modern vision-language models on medical image understanding tasks when

they are appropriately prompted with structured domain knowledge, which suits the per-detection explanation setting we target.

A practical caveat is that reliable non-falciparum speciation on thick film is acknowledged to be difficult even for expert microscopists. The standard WHO bench aids note that *Pv*, *Po*, and *Pm* are often indistinguishable on thick film and that confirmatory speciation is usually performed on a thin film [14]. The YOLOv12n detector used in this study was trained on thick-smear images with clinical species labels, so its predictions reflect patterns learnable from thick film under the labelling conventions of that laboratory. SGMCE explains those predictions using thick-smear-appropriate morphological evidence. It does not claim to resolve cases that would be ambiguous even under expert review.

We introduce **SGMCE** (Segment-Grounded Morphological Concept Explanation), a hybrid computer-vision (CV) plus vision-language model (VLM) post-hoc framework. For each detected parasite, SGMCE *(i)* identifies species-discriminating morphological features supporting the classification; *(ii)* provides differential reasoning explicitly rejecting competing species; *(iii)* describes the evidence in clinical vocabulary; and *(iv)* self-validates through four metrics.

*Contributions.*

1. A post-hoc LLM-based explanation pipeline for malaria speciation that requires no model retraining, no morphological training annotations, and is applicable to any segmentation use case where the model outputs instance masks.
2. A thick-smear-specific morphological knowledge base encoding per-species hallmarks and pairwise differentiators, compiled from the WHO bench aids for the diagnosis of malaria infections [14].
3. A handcrafted CV morphological feature extractor computing fourteen clinically grounded measurements from mask-guided crops using adaptive within-mask thresholds suited to variable thick-smear staining.
4. Four automatic explanation-validation metrics, including a sentence-level semantic embedding scoring rule with species-aware negation filtering that resolves vocabulary mismatch between clinical prose and knowledge-base keyword strings.
5. Empirical evaluation on 737 detections from 139 held-out thick-smear images across four *Plasmodium* species, showing parasite-class mean KBC of 0.91, DS of 0.99, and CCF of 0.97, with a per-rule CCF breakdown that supports the faithfulness of the VLM's visual claims.

## 2 SGMCE Framework

### 2.1 Pipeline Overview

Figure 1 illustrates the five-stage SGMCE pipeline. Given a trained segmentation model (here, a Yolov12) and a thick-smear image, SGMCE produces for each

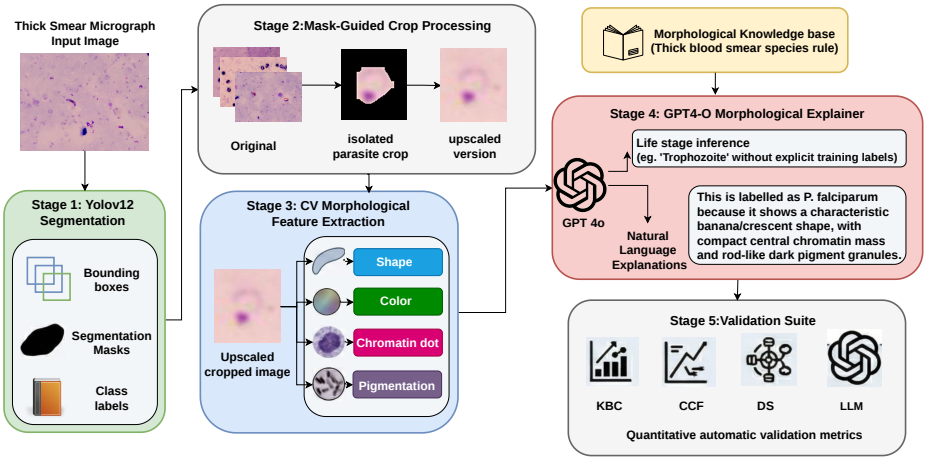


**Fig. 1.** SGMCE pipeline for thick-film malaria explanation. Each detected object passes through five stages. GPT-4o receives two upscaled crop images (visual path) and a structured text rendering of the fourteen CV measurements (measurement path), conditioned on a thick-smear-specific morphological knowledge base (KB) compiled from WHO guidelines.

detected object: two mask-guided cropped thumbnails, fourteen CV morphological measurements, a species-discriminating natural-language explanation with differential reasoning, and four automatic validation scores. Stages 1–2 (detection and crop preprocessing) are standard components adapted for thick-smear constraints. Stages 3–5 (handcrafted morphological feature extraction, knowledge-base-grounded GPT-4o reasoning, and the automatic validation suite) are the novel contributions of this work. All stages are post-hoc: no model weights are read or modified at any stage.

### 2.2 YOLOv12n Segmentation Detector

We use a YOLOv12n-seg model [13] trained on approximately 14,000 thick-smear images (after class-balancing augmentation) from a national biomedical centre in Sub-Saharan Africa, classifying five categories: *Pf*, *Pv*, *Po*, *Pm*, and white blood cells (WBCs). Training runs for 70 epochs at resolution 2048×2048 with SGD (momentum 0.937, cosine learning-rate schedule). At inference, we apply a confidence threshold of 0.25. Each detection yields a bounding box $(x_1, y_1, x_2, y_2)$, an instance mask $M \in \{0,1\}^{H \times W}$, a class label $c$, and a confidence $\hat{p}$. All detections above the confidence threshold - including any debris, staining artefacts, or distractor objects that the detector classifies as parasites - are passed through the subsequent SGMCE stages unchanged; low-confidence detections are flagged diagnostically but not suppressed (Sect. 3.3).

### 2.3 Mask-Guided Crop Preprocessing

Raw detections span only 10–60 px in our acquisition setting, which is too small for reliable visual analysis by a VLM. The *CropPreprocessor* extracts two complementary thumbnails per detection (examples in Figure 1, stage 2 box).
**1. Masked Crop.** The bounding-box region is extracted and pixels outside mask $M$ are replaced with the image's estimated background colour (sampled from the image corners, yielding the characteristic pinkish Giemsa-stained smear background), isolating the parasite from surrounding lysed red-blood-cell debris.
**2. Context Crop.** The bounding box is symmetrically padded by 50 % of its width and height before extraction, preserving surrounding context (nearby parasites, co-detected WBCs, pigment distribution).

Both crops are upscaled to a minimum dimension of 200 px (4–12× magnification) via bicubic interpolation, which preserves edge sharpness at these magnification factors and is important for chromatin-dot boundary delineation. Crops are encoded as JPEG images for transmission to GPT-4o.

### 2.4 Handcrafted Morphological Feature Extractor

The *CVFeatureExtractor* computes fourteen handcrafted morphological measurements from the masked crop (Table 1), grounded in clinical practice [14]. Thresholds are adaptive, derived from within-mask pixel statistics, to accommodate the variable staining intensity typical of field-prepared thick smears.
**Shape Analysis.** Shape is computed from the binary mask contour. Let $A$ denote pixel area and $P$ the contour perimeter. *Circularity* is $C = 4\pi A/P^2 \in [0, 1]$, where $C = 1$ is a perfect circle and $C > 0.65$ is characteristic of ring-stage parasites and WBCs. Fitting the minimum-area enclosing ellipse yields major/minor semi-axes $a, b$ with aspect ratio $\mathrm{AR} = a/b$ and eccentricity $e = \sqrt{1 - (b/a)^2}$; $\mathrm{AR} > 2.5$ and $e > 0.90$ flag the characteristic falciform *Pf* gametocyte. *Solidity* $S = A/A_{\mathrm{hull}}$ (where $A_{\mathrm{hull}}$ is the convex-hull area) captures amoeboid cytoplasmic extensions typical of *Pv* growing trophozoites when $S < 0.82$.
**Colour Analysis.** The masked crop is converted from BGR to CIE LAB and HSV colour spaces, and statistics are computed over mask pixels only. *Blue-purple fraction* BP is the proportion of mask pixels with OpenCV HSV hue $h_i \in [100, 150]$ (equivalent to $[200°, 300°]$ in standard convention), capturing Giemsa nucleic-acid staining. *Dark fraction* DF is the proportion of mask pixels with HSV value $V_i < 80$, indicating haemozoin or densely packed chromatin.
**Chromatin Detection.** Chromatin appears as small bright reddish dots under Giemsa staining and is most prominent on the CIE LAB $a^*$ axis – the red-to-green opponent channel, where positive values indicate red. Let $\mu_a$ and $\sigma_a$ be the within-mask mean and standard deviation of $a^*$. Candidate chromatin pixels satisfy

$$a_i^* > \mu_a + \max(1.2\,\sigma_a,\ 8.0), \tag{1}$$

where the absolute offset 8.0 ensures a meaningful threshold when $\sigma_a$ is low. Connected components whose area is between 0.5 % and 15 % of the mask area

**Table 1.** The fourteen handcrafted CV morphological features computed by `CVFeatureExtractor`. All threshold-derived flags are adaptive relative to within-mask pixel statistics.

| Feature | Type | Morphological interpretation |
|---|---|---|
| `circularity` | Float $[0, 1]$ | Compactness; high ($> 0.65$) for rings or WBCs |
| `aspect_ratio` | Float | Elongation; $> 2.5$ flags Pf gametocyte |
| `eccentricity` | Float $[0, 1]$ | Ellipse axis ratio; high for crescent shape |
| `solidity` | Float $[0, 1]$ | Convexity; $< 0.82$ flags amoeboid Pv |
| `is_round` | Bool | $C > 0.65$ |
| `is_elongated` | Bool | AR $> 2.0$ |
| `is_amoeboid` | Bool | $S < 0.82$ |
| `blue_purple_fraction` | Float | Giemsa nucleic-acid stain uptake |
| `dark_fraction` | Float | Dark pixel fraction (haemozoin / chromatin) |
| `chromatin_count` | Int | Chromatin dots $N_c$ |
| `has_double_chromatin` | Bool | $N_c \geq 2$; key Pf ring indicator |
| `pigment_count` | Int | Haemozoin granules $N_p$ |
| `pigment_distribution` | String | Central / peripheral / scattered |
| `relative_size_vs_wbc` | Float | Parasite-to-WBC area ratio (when WBC available) |

yield the chromatin count $N_c$. The *double-chromatin flag* ($N_c \geq 2$) is characteristic of Pf ring-stage parasites.

**Pigment (Haemozoin) Detection.** Haemozoin appears as dark brownish-black granules with species-specific colour, count, and distribution. Detection uses the CIE $L^*$ (luminance) channel; pixels satisfying $L_i^* < \mu_L - \max(\sigma_L,\ 10.0)$ are haemozoin candidates. Components whose area is at least 0.5 % of the mask area contribute to the granule count $N_p$; the normalised mean distance of granule centroids from the mask centroid classifies the distribution as *central* ($< 35\,\%$ of the maximum radius), *peripheral* ($> 60\,\%$), or *scattered*.

**Size Reference via Co-detected WBC.** WBCs are reliably abundant in thick films, and lie in a well-characterised size range, typically 10–14 µ, which projects to roughly 60–120 px in our 2048×2048 acquisitions, whereas ring-stage parasites span 10–30 px. When a WBC is co-detected in the same image, the parasite-to-WBC area ratio is reported per detection as `relative_size_vs_wbc`. Typical values are 0.03–0.15 for ring-stage parasites, 0.2–0.6 for larger trophozoites or schizonts, and close to 1.0 for co-detected WBCs themselves. This ratio gives GPT-4o a grounded size cue that does not rely on intact red blood cells, which are absent in thick films due to lysing.

### 2.5 Thick-Smear Morphological Knowledge Base

The knowledge base (KB) is compiled from the WHO bench aids for the diagnosis of malaria infections [14], the standard thick-smear parasitology reference used

for laboratory training in endemic regions. The KB is deliberately thick-film-specific: features that depend on intact red blood cells are excluded, because RBC enlargement (Pv) and Ziemann's stippling (Pm) are not visible after RBC lysis. Schüffner's dots, visible as discrete pink stipples on thin films, appear on thick films as a diffuse pinkish cloud around the parasite; the KB describes this thick-film appearance explicitly to avoid priming the VLM with thin-film terminology.

The KB contains one entry per class (*Pf, Pv, Po, Pm, WBC*). Each entry records `thick_film_notes` (general appearance); `life_stages` (per-stage feature lists for ring, growing trophozoite, mature trophozoite, schizont, and gametocyte); `key_differentiators` (three to five most discriminative thick-film features); `pigment` (haemozoin colour, distribution, count); and `differential_vs` (pairwise feature lists against each alternative species). Serialised as structured plain text, the KB is embedded verbatim in the GPT-4o system prompt for every API call, ensuring consistent grounding across all detections. The full system prompt (KB plus thick-smear context preamble, reasoning guidance, and output JSON schema) occupies approximately 1,500 tokens and contains no embedded images.

### 2.6 GPT-4o Morphological Explainer

The *MorphologicalExplainer* queries GPT-4o through the OpenAI Chat Completions API. Each call comprises a structured system prompt, a per-detection user prompt, and two base64-encoded JPEG images (masked and context crops).

**System Prompt.** Four parts: *(i)* a thick-smear context preamble that directs GPT-4o away from thin-smear reasoning patterns; *(ii)* the full KB from Sect. 2.5; *(iii)* reasoning guidance that prioritises species-discriminating morphological evidence and that requires an explicit rejection of each alternative species before reporting uncertainty; *(iv)* the output JSON schema, specifying every output field's name, type, and content guidance, with `supporting_features` and `differential_reasoning` designated the primary outputs. A separate `WBC_SYSTEM_PROMPT` is used for WBC detections. It focuses on nuclear architecture (neutrophil, lymphocyte, monocyte, eosinophil) and explicitly prohibits parasite-feature discussion, preventing the VLM from over-reading morphology on non-parasite objects.

**User Prompt.** Three items are passed per detection, in addition to the two images: *(i)* the YOLO class label $c$ and confidence $\hat{p}$, presented as a prior belief rather than ground truth; *(ii)* the fourteen CV measurements rendered as a structured text block, pairing each numeric value with a human-readable interpretive label so that GPT-4o can treat it as a natural-language cue rather than a raw number, *e.g.,*:

```
Circularity: 0.82 (round), AR: 1.04, Chromatin dots: 2 (DOUBLE -
notable in Pf ring stage), Pigment granules: 1 (central),
Blue-purple fraction: 0.47, Size vs WBC: 0.06 (parasite is very
small relative to co-detected WBC).
```

*(iii)* a per-species hallmark checklist, asking GPT-4o to respond `observed`, `not_observed`, or `uncertain` for each thick-film hallmark of the detected class (*e.g.,* for Pf: "double chromatin dot", "small delicate ring", "falciform gametocyte").
**Structured Output.** A JSON object with primary fields: `supporting_features` (feature–observation–significance triples grounding specific claims in visual evidence); `differential_reasoning` with explicit per-species keys (`vs_pf`, `vs_pv`, `vs_po`, `vs_pm`), each stating why the alternative species is excluded; `hallmark_checklist` (per-hallmark verdict); `natural_language_explanation` (a free-form clinical description, 2–4 sentences); and `explanation_confidence`. Secondary fields include `estimated_life_stage` and `caveats`. A partial example for a Pm schizont:

```
"natural_language_explanation":
"The observed morphology is consistent with Plasmodium
malariae, primarily due to the presence of a rosette
schizont formation and central pigment granule. These
features are pathognomonic for Pm and exclude other
species.",
"differential_reasoning": {
"vs_pf": "Pf is excluded due to the absence of very
small rings and double chromatin dots, and the
presence of a rosette formation which is not seen
in Pf.",
"vs_pv": "Pv is excluded because the amoeboid shape
is not typical for Pm, and there is no
golden-brown pigment.",
"vs_po": "Po is excluded due to the absence of early
prominent stippling and the presence of a rosette
formation, which is not characteristic of Po." }
```

We use temperature 0.05 and up to 1,800 output tokens; life-stage inference is reported as a secondary field but is not evaluated quantitatively here, because the training labels encode species only.

### 2.7 Automatic Validation Suite

Four metrics assess explanation quality without expert re-annotation and without any manual spot-checking of individual explanations. KBC, CCF, and DS are newly introduced in this work; LLMj adapts the general LLM-as-judge paradigm of Zheng *et al.* [18] to the per-detection setting.
**Knowledge-Base Consistency (KBC).** KBC measures the proportion of species-appropriate thick-film morphological concepts that are semantically present in explanation $E$. For each species we curate a positive concept set $\mathcal{P} = \{p_1, \ldots, p_m\}$ drawn verbatim from the species' `key_differentiators` (*e.g.,* for Pf: "double chromatin dot", "delicate ring"; for Pv: "amoeboid cytoplasm", "golden-brown pigment"). $E$ is split into sentences $\{s_1, \ldots, s_n\}$ and each sentence and concept is independently embedded with OpenAI `text-embedding-3-small`

($d$=1536). The sentence-level maximum cosine similarity is

$$f(E, p_k) = \max_{j \in [n]} \cos\big(\phi(s_j),\, \phi(p_k)\big), \tag{2}$$

which avoids the signal dilution that a whole-document embedding would introduce. KBC is then

$$\text{KBC} \;=\; \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\big[f(E, p_k) \geq \theta\big], \qquad \theta = 0.40. \tag{3}$$

Competing-species concepts (*e.g.,* Pv hallmarks evaluated against a Pf explanation) are computed and logged for diagnostic inspection but are excluded from the score, because sentence-level embeddings are negation-blind: a sentence such as "unlike the amoeboid morphology of *P. vivax*, this parasite is compact" scores high cosine similarity to the Pv amoeboid concept even though it correctly rejects it. Penalising such sentences would actively discourage differential reasoning, one of the primary SGMCE outputs.

**Discriminativeness Score (DS).** DS is computed identically to KBC (Eqs. 2–3) but over a curated hallmark subset $\mathcal{H} \subseteq \mathcal{P}$ of species-hallmark phrases most reliably distinguishing the detected species from all others (*e.g.,* "rosette schizont" for Pm, "banana-shaped crescent gametocyte" for Pf, "amoeboid fragmented cytoplasm" for Pv):

$$\text{DS} \;=\; \frac{1}{q} \sum_{k=1}^{q} \mathbb{1}\big[f(E, h_k) \geq \theta\big]. \tag{4}$$

**CV-Claim Faithfulness (CCF).** CCF checks whether the explanation's visual claims are consistent with the computer vision (CV) measurements. A set of seven handcrafted correspondence rules $\mathcal{R}$ (Table 2) maps each CV measurement to the clinical vocabulary a microscopist would use for it. Each rule activates only when its trigger keywords appear in the explanation; it then verifies that the CV measurement satisfies the expected condition. To avoid the same negation issue as in KBC, CCF applies the rules only to the `natural_language_explanation` and `supporting_features` fields (not the `differential_reasoning` field), and passes each candidate clause through a species-aware negation filter that skips clauses of the form "unlike Pv…" or "no amoeboid cytoplasm". CCF is the fraction of activated rules that agree with the corresponding CV measurement.

**LLM-as-Judge (LLMj).** LLMj adapts the LLM-as-judge paradigm [18] by prompting an independent GPT-4o instance (blind to the original system prompt) to score factual accuracy (FA), differential specificity (SP), and clinical utility (CU), each on $[0, 1]$, yielding LLMj = (FA+SP+CU)/3. LLMj costs approximately \$0.01–0.03 per explanation and is reported as a complementary diagnostic; the main validation in Sect. 3 relies on the three deterministic metrics KBC, CCF, and DS.

**Table 2.** The seven CV-Claim Faithfulness (CCF) correspondence rules. Each rule activates when its trigger keywords appear in the explanation; the CV condition is then checked. Negation-scoped clauses are skipped before activation.

| Trigger keywords | CV feature | Condition |
|---|---|---|
| elongated, banana, crescent, sausage | `aspect_ratio` | $> 2.0$ |
| round, circular, compact shape | `circularity` | $> 0.65$ |
| amoeboid, irregular, fragmented | `solidity` | $< 0.85$ |
| double chromatin, two dots, two chromatin | `chromatin_count` | $\geq 2$ |
| single chromatin, one dot, single dot | `chromatin_count` | $= 1$ |
| dark pigment, hemozoin, pigment granule | `has_pigment` | True |
| blue cytoplasm, purple cytoplasm | `blue_purple_fraction` | $> 0.15$ |

## 3 Experiments and Results

### 3.1 Dataset and Configuration

**Training Data.** The YOLOv12n-seg detector is trained on approximately 7,000 original thick blood smear images from Rwanda Biomedical Center, with species-specific offline augmentation and a noise-injection pass that expands the effective training volume to roughly 14,000 samples. All images are acquired under Giemsa staining at high magnification (oil-immersion). Exact per-image pixel-to-micron calibration and numerical aperture are not uniformly recorded in the acquisition metadata; the characteristic object sizes at acquisition time ($\sim$10–14 µm WBCs $\rightarrow$ 60–120 px, ring parasites $\rightarrow$ 10–30 px on $2048^2$ frames) are consistent with a 100$\times$ oil-immersion objective, following standard thick-film microscopy practice.
**Test Set and Evaluation Sample.** The held-out test set contains 410 thick-smear images (134 Po, 125 Pm, 125 Pf, 26 Pv). For SGMCE evaluation we use an adaptive per-species stratified sample: images are first ranked by detector density, and the per-species budget is chosen so that fewer images are needed for dense species (Pf) and more for sparse species (Pv, Pm). The Pv target of 95 detections is reached by including the entire available Pv test partition (26 images); remaining budgets are allocated proportionally. The resulting sample comprises 139 images and 737 detections at confidence threshold $\hat{p} \geq 0.25$ (Pf: 233, WBC: 209, Pm: 105, Po: 97, Pv: 93). All detections above the threshold (including confidence-based false-positive candidates) are passed through the full pipeline.
**Configuration.** YOLOv12n inference runs on an NVIDIA GPU; all subsequent stages are CPU and API-bound. *CropPreprocessor* targets a minimum crop dimension of 200 px with 50 % bounding-box padding for context. GPT-4o is queried at temperature 0.05 with up to 1,800 output tokens. The *ValidationSuite* uses `text-embedding-3-small` at cosine-similarity threshold $\theta = 0.40$ with a shared embedding cache across all detections.

**Table 3.** Per-class validation scores and keyword-vs-semantic comparison across 737 detections from 139 thick-smear images, using sentence-level semantic scoring with $\theta = 0.40$. Left: overall KBC, CCF, DS per class. Right: keyword matching vs. semantic scoring for KBC and DS on the same GPT-4o explanations; the semantic rule with species-aware negation filtering recovers most of the score lost to paraphrase. KBC and DS are not applicable to WBC since the KB encodes no parasite concepts for this class.

| | | | | | KBC | | DS | |
|---|---|---|---|---|---|---|---|---|
| Species | Detections | KBC | CCF | DS | Keyword | Semantic | Keyword | Semantic |
| *P. falciparum* (Pf) | 233 | 0.81 | 1.00 | 0.96 | 0.02 | 0.81 | 0.57 | 0.96 |
| *P. vivax* (Pv) | 93 | 1.00 | 0.90 | 1.00 | 0.01 | 1.00 | 0.34 | 1.00 |
| *P. ovale* (Po) | 97 | 0.88 | 1.00 | 1.00 | 0.15 | 0.88 | 0.52 | 1.00 |
| *P. malariae* (Pm) | 105 | 0.98 | 0.97 | 0.99 | 0.34 | 0.98 | 0.74 | 0.99 |
| WBC | 209 | n/a | 0.94 | n/a | | | | |
| Parasite mean | 528 | 0.91 | 0.97 | 0.99 | 0.13 | 0.91 | 0.54 | 0.99 |

### 3.2 Quantitative Results

Table 3 reports per-class mean validation scores. Table 3 (right panel) compares sentence-level semantic scoring against a keyword-matching baseline that tests only for verbatim KB strings, on the same 737 cached explanations. Semantic scoring recovers most of the score lost to vocabulary mismatch: GPT-4o paraphrases KB terms rather than quoting them verbatim (*e.g.,* "falciform" for "banana-shaped", "two chromatin dots" for "double chromatin").

### 3.3 False Positive Analysis

Because SGMCE passes every above-threshold detection to the explanation pipeline, distractors and weakly supported detections are processed alongside genuine parasites. We therefore report a confidence-based false-positive (FP) flag as a diagnostic: any detection with $0.25 < \hat{p} \leq 0.30$ (i.e., a detection retained by the model, but only marginally above the detection threshold) is flagged as a *suspected* FP on the grounds that the detector was least certain it corresponds to a real parasite. The flag is intentionally conservative about the reverse direction where a low confidence does not prove the object is a distractor, and some flagged detections are genuine parasites the detector found visually difficult. The flagged subset provides a short queue of candidates most worth a human reviewer's attention. Across the 737 detections, 34 (4.6 %) were flagged (Table 4). Flag rates were highest for Pf (8.6 %) and Pm (6.7 %), and lowest for Pv (1.1 %) and Po (1.0 %). For flagged detections, the GPT-4o explainer tends to qualify its natural-language output with caveats (the `caveats` field of the JSON schema), which a human reviewer can use to decide whether to audit the detection manually.

**Table 4.** Confidence-based false-positive flag per class ($\hat{p} \leq 0.3$). The flag is a diagnostic aid, not a definitive FP label: some low-confidence detections are genuine parasites that the detector was simply less certain about.

| Class | Low-conf. | High-conf. | Low-conf. rate |
|---|---|---|---|
| Pf | 20 | 213 | 8.6 % |
| Pv | 1 | 92 | 1.1 % |
| Po | 1 | 96 | 1.0 % |
| Pm | 7 | 98 | 6.7 % |
| WBC | 5 | 204 | 2.4 % |
| All | 34 | 703 | 4.6 % |

**Table 5.** Per-rule CCF diagnostic. For each species we show the six most-fired CCF rules, with the number of detections on which the rule activated and the fraction that agreed with the CV measurement. "–" denotes rule-species pairs with zero activations.

| Rule / Species | Pf | Pv | Po | Pm | Agreement rate |
|---|---|---|---|---|---|
| pigment present | 174 | 18 | 30 | 85 | 100 % (all) |
| round / compact | 32 | 66 | 80 | 54 | 90.7 % (Pm); 100 % (Pf, Pv, Po) |
| single chromatin | 102 | 2 | 12 | 13 | 100 % (all) |
| double chromatin | 50 | 3 | 6 | – | 100 % (all) |
| elongated shape | 10 | 2 | 1 | – | 100 % (all) |
| irregular outline | – | 10 | – | 3 | 0 % (Pv); 66.7 % (Pm) |

### 3.4 Per-Rule CCF Diagnostic

Aggregate CCF hides which individual correspondence rules are most active and where the VLM's claims disagree with the CV measurements. Table 5 reports, for each species and each activated rule, the number of detections where the rule fired (the trigger keywords appeared in the non-negated portion of the text) and the fraction for which the CV measurement satisfied the rule condition. The "pigment present", "single chromatin", "double chromatin", "round/compact", and "elongated shape" rules fire on hundreds of detections and agree in $\geq 99$ % of cases across all four species, giving direct per-claim evidence that GPT-4o's visual descriptions are grounded in the measured morphology rather than being generic species labels. The lowest per-rule agreement is "irregular outline"→`solidity`$< 0.85$ for Pv, which fires on 10 detections and agrees on 0. Inspection shows these are rings whose solidity exceeds 0.85 (CV measurement: compact) but whose GPT-4o prose nonetheless describes them as "slightly irregular". This is a disagreement the rule correctly catches, and is the primary driver of Pv's CCF = 0.90.

### 3.5 Qualitative Example

Figure 2 shows an example SGMCE report for a thick-smear field in which the detector identified one Pm schizont ($\hat{p} = 0.912$) and one co-detected WBC

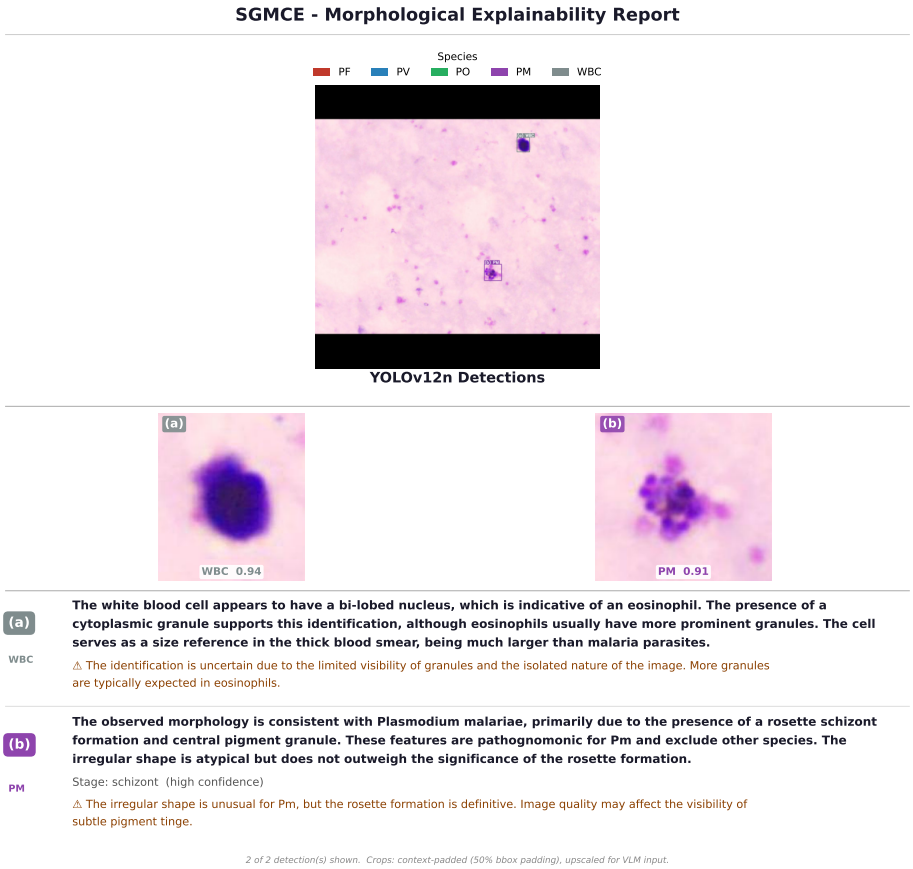


**Fig. 2.** Example SGMCE report on a thick-smear field with a Pm schizont and a co-detected WBC. For each detection, the report lists the detected class, the supporting morphological features, the free-form natural-language explanation, and the per-species differential reasoning.

($\hat{p} = 0.935$). For the Pm detection, SGMCE's supporting features cite the rosette-pattern schizont and the central dark pigment granule, both pathognomonic for Pm under the KB, and the differential block rejects Pf ("no very small rings or double chromatin dots"), Pv ("no amoeboid shape or golden-brown pigment"), and Po ("no early prominent stippling; rosette not characteristic of Po"). The WBC detection, driven by the dedicated WBC prompt, is described as a probable eosinophil (bi-lobed nucleus, scant granules) and explicitly avoids parasite morphology claims. The report contains no hand-curated text: the species call, the supporting features, and the per-species rejections are all produced automatically for each detection. This approach to explanation is arguably more generalizable to other use cases than the fixed text templates of MorphXAI.

### 3.6 Discussion and Limitations

*CCF Analysis.* CCF is $\geq 0.90$ for all parasite classes and 0.94 for WBCs (Table 3). The per-rule diagnostic (Table 5) shows that the high-volume rules (`pigment present`, chromatin-count rules, `round/compact`) fire on hundreds of detections and agree at $\geq 99\,\%$, so the aggregate score is not carried by a few low-activation rules. The single visible disagreement mode is the "irregular outline" $\rightarrow$ solidity rule for Pv (Sect. 3.4), which correctly flags a systematic over-description by GPT-4o. Read together with the raw VLM text, CCF functions as a per-claim auditor rather than a summary score.

*KBC Analysis.* KBC under verbatim keyword matching is low (0.01–0.34 across species; Table 3, right panel), not because the explanations omit thick-film morphology but because GPT-4o paraphrases KB terms in natural clinical prose. Sentence-level semantic scoring with species-aware negation filtering recovers this substantially, raising parasite-class mean KBC from 0.13 to 0.91. The score is now dominated by the paraphrase coverage of the explanation rather than by surface-form mismatch.

*DS Analysis.* DS improves across all four species under semantic scoring (Pf: $0.57 \rightarrow 0.96$; Pv: $0.34 \rightarrow 1.00$; Po: $0.52 \rightarrow 1.00$; Pm: $0.74 \rightarrow 0.99$). The improvement is largest for species whose hallmarks are vocabulary-rich but rarely quoted verbatim, notably the falciform Pf gametocyte and the Pv amoeboid trophozoite.

*Thick-Film Ambiguity and Honest Scope.* The evaluation in this paper measures whether SGMCE's explanations are consistent with the detector's predictions and with the thick-film KB. It does not claim to resolve cases that would be ambiguous even under expert review. The WHO bench aids themselves note that Pv, Po, and Pm are often indistinguishable on thick film [14], and confirmatory speciation is usually performed on a thin film. SGMCE inherits this limit.

*Negation Handling.* The CCF negation filter skips clauses such as "unlike *P. vivax*, no amoeboid cytoplasm" before rule activation, and KBC/DS deliberately exclude competing-species concepts from the score. These rules remove the most visible inflation and deflation modes of embedding-based metrics, but they do not replace a full natural-language inference model.

*Life-Stage Inference.* SGMCE exposes an `estimated_life_stage` field as a secondary output. We do not evaluate per-stage accuracy quantitatively because the training labels encode species only; reported life stages are provided as a microscopist-facing aid but should not be interpreted as having been validated against stage ground truth.

*What This Evaluation Does Not Cover.* No prospective expert-microscopist validation was conducted for this study; the four automatic metrics are the primary evaluation. A human-subjects study comparing unassisted microscopist

review with SGMCE-assisted review on real clinical slides is the natural next step. A separate extension is applying the same pipeline to thin-film images, where intact RBCs enable richer size-reference features.

*Cascade Dependency on the Detector.* Every SGMCE claim depends on a correct model detection and mask. Three failure modes follow: a missed parasite produces no record; a confident false positive on a distractor hands the VLM a class label it will rationalise against that species' expected morphology; and a disconnected or background-heavy mask feeds the CV features a distorted region. The confidence-based FP flag catches the low-confidence portion of the first two, but not all, and is silent on the third. The per-rule CCF diagnostic remains the most useful automatic signal when the VLM's claims diverge from the CV measurements.

*Practical Deployment Concerns.* The default GPT-4o backend is a commercial API: it raises data-privacy concerns for clinical images and is unusable offline. Both can be addressed by swapping the cloud VLM for a locally deployed open-weights model, with the rest of the pipeline remaining unchanged.

*Beyond Malaria.* The SGMCE template is not specific to *Plasmodium*. Its components (mask-guided crop preprocessing, handcrafted morphological features, a thin-text domain KB, VLM reasoning, and an automatic validation suite) fit any detection task where expert diagnosis is driven by structured morphological features and a concise KB can be written: other blood-film parasites, cytology, and digital pathology ROI triage are candidate domains.

## 4 Conclusion

We presented SGMCE, a post-hoc explanation framework for malaria species identification in thick blood smears that generates clinically grounded natural-language morphological explanations without retraining or additional annotation. By coupling mask-guided crop preprocessing, fourteen handcrafted adaptive CV features, and GPT-4o reasoning conditioned on a WHO-derived thick-smear knowledge base, SGMCE produces, for every detection, a species call with supporting features and an explicit differential against each competing species. The KBC and DS metrics indicate that the explanations carry the expected clinical content, and the per-rule CCF diagnostic confirms that the VLM's visual claims align with the CV measurements rather than reading as generic species labels. By anchoring algorithmic decisions in verifiable domain concepts, SGMCE offers a route to explanations that increase user trust and support the responsible adoption of automated decision-making systems.

**Acknowledgments.** We thank the Rwanda Biomedical Center for helping us with data collection and Afretec Network for funding the project.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References


1. Boit, S., Patil, R.: An Efficient Deep Learning Approach for Malaria Parasite Detection in Microscopic Images. Diagnostics **14**(23), 2738 (Jan 2024). https://doi.org/10.3390/diagnostics14232738
2. Centers for Disease Control and Prevention: Treatment of Malaria: Guidelines for Clinicians (United States) (2023), https://stacks.cdc.gov/view/cdc/131221
3. Delahunt, C.B., Gachuhi, N., Horning, M.P.: Metrics to guide development of machine learning algorithms for malaria diagnosis. Frontiers in Malaria **2** (Apr 2024). https://doi.org/10.3389/fmala.2024.1250220
4. Delahunt, C.B., Jaiswal, M.S., Horning, M.P., Janko, S., Thompson, C.M., Kulhare, S., Hu, L., Ostbye, T., Yun, G., Gebrehiwot, R., Wilson, B.K., Long, E., Proux, S., Gamboa, D., Chiodini, P., Carter, J., Dhorda, M., Isaboke, D., Ogutu, B., Oyibo, W., Villasis, E., Tun, K.M., Bachman, C., Bell, D., Mehanian, C.: Fully-automated patient-level malaria assessment on field-prepared thin blood film microscopy images. In: 2019 IEEE Global Humanitarian Technology Conference (GHTC). pp. 1–8 (Oct 2019). https://doi.org/10.1109/GHTC46095.2019.9033083, iSSN: 2377-6919
5. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: Proceedings of the 35th International Conference on Machine Learning. pp. 2668–2677 (2018)
6. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
7. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023). https://doi.org/10.1038/s41586-023-05881-4
8. OpenAI: GPT-4o System Card (Oct 2024). https://doi.org/10.48550/arXiv.2410.21276, arXiv:2410.21276 [cs]
9. Poostchi, M., Silamut, K., Maude, R.J., Jaeger, S., Thoma, G.: Image analysis and machine learning for detecting malaria. Translational Research **194**, 36–55 (Apr 2018). https://doi.org/10.1016/j.trsl.2017.12.004
10. Ramos-Briceño, D.A., Flammia-D'Aleo, A., Fernández-López, G., Carrión-Nessi, F.S., Forero-Peña, D.A.: 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 (Jan 2025). https://doi.org/10.1038/s41598-025-87979-5
11. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier (Aug 2016). https://doi.org/10.48550/arXiv.1602.04938, arXiv:1602.04938 [cs]
12. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (Oct 2017). https://doi.org/10.1109/ICCV.2017.74, iSSN: 2380-7504
13. Tian, Y., Ye, Q., Doermann, D.: YOLOv12: Attention-Centric Real-Time Object Detectors (Feb 2025). https://doi.org/10.48550/arXiv.2502.12524, arXiv:2502.12524 [cs] version: 1
14. World Health Organization: Bench aids for the diagnosis of malaria infections, 2nd ed. (2000), https://iris.who.int/handle/10665/42195

15. World Health Organization: Guidelines for the treatment of malaria, third edition (2015)
16. World Health Organization: World malaria report 2023 (2023)
17. Yousaf, A., Win, S.S., Coffee, M., Olufowobi, H.: MorphXAI: An Explainable Framework for Morphological Analysis of Parasites in Blood Smear Images (Jan 2026). https://doi.org/10.48550/arXiv.2601.18001, arXiv:2601.18001 [cs] version: 1
18. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track (2023)