# **Dataset Descriptions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Characteristic** | [Hung & Carpenter Dataset](https://openaccess.thecvf.com/content_cvpr_2017_workshops/w8/papers/Hung_Applying_Faster_R-CNN_CVPR_2017_paper.pdf#page=2.96)[**d**ata](https://datasetninja.com/malaria-bounding-boxes) **D2** | [Davidson et al. Dataset](https://www.cambridge.org/core/services/aop-cambridge-core/content/view/8573173B4952D45CA7618E548977EB50/S2633903X21000015a.pdf/automated-detection-and-staging-of-malaria-parasites-from-cytological-smears-using-convolutional-neural-networks.pdf)  [**d**ata](https://data.mendeley.com/datasets/j55fyhtxn4/1) **D1** | [Guemas et al. RT-DETR Dataset](https://journals.asm.org/doi/full/10.1128/spectrum.01440-23)  [data](https://zenodo.org/records/8358829) **D3** |
| **Species** | P. vivax | P. falciparum | *P. falciparum, P. ovale, P. vivax, P. malariae, Trypanosoma brucei, Babesia divergens* |
| **Source** | Ex vivo patient samples | Laboratory cultures | Patient samples from 6 French university hospitals |
| **Number of patients/sources** | 5 patients (4 for training, 1 for testing) | Multiple laboratory cultures across 6 research centers | 475 thin blood smears (training/validation), 170 smears for testing (170 patients) |
| **Parasite strains** | Not specified | 3D7, NF54, DD2, D10 | Clinical isolates (not specified) |
| **Total images** | 1,300 microscopy images | 399 | 29,228 (24,720 training/validation, 4,508 testing) |
| **Total cells** | ~100,000 cells | ~38,000 RBCs | 2,002,597 labels (training/validation), 358,825 labels (testing) |
| **Infected cells percentage** | ~3% | ~6% | ~6% (I calculated from label distribution here) |
| **Staining method** | Giemsa | Giemsa (standardized protocol) | May Grunwald–Giemsa (MGG) |
| **Microscopy** | 100% oil immersion | 100% oil immersion with 100× objective | ×1,000 magnification |
| **Classes** | 7 labels (RBC, leukocyte, gametocyte, ring, trophozoite, schizont, difficult) | 3 main stages (ring, trophozoite, schizont) with early/late substages + gametocytes | 9 categories: WBC, RBC, platelets, T. brucei, and RBCs infected by P. falciparum, P. ovale, P. vivax, P. malariae, and B. divergens |
| **Annotations** | Single expert | 10 experts from different research centers | Senior parasitologists with PCR confirmation for recent samples |
| **Annotation method** | Direct annotation | Model-assisted labeling via LabelBox | Manual initially, then automatically with manual correction using CVAT software |
| **Class imbalance** | Severe (97% uninfected RBCs) | Significant (94% uninfected RBCs) | Severe (94% uninfected RBCs) |
| **Annotator agreement** | Not assessed | 60% unanimous agreement on training set | Not assessed |
| **Data format** | Full-sized images with crops | Full images and 70×70 pixel crops | Full microscopy field images |
| **Data augmentation** | Rotation, contrast changes, brightness changes | Rotation, flipping, contrast changes, brightness changes, RGB randomization | Not explicitly mentioned |
| **Test data approach** | Separate patient | Separate research center | Multicentric test set from 6 different hospitals |

\*\*Micro Avereage

# Complete Quality-Guided Focal Loss Multi-Architecture Experimental Plan

## 1. Research Objectives

### Primary Research Questions

1. **Architecture Generalization**: Does Quality-Guided Focal Loss (QGFL) improve minority class detection across modern object detection architectures beyond RetinaNet?
2. **Multi-Class Extension**: Can QGFL be effectively adapted for hierarchical medical classification tasks (species identification and staging)?
3. **Foundation Model Enhancement**: Do modern foundation models (RedDino) create synergistic improvements when combined with QGFL?

### Research Contributions

* First systematic evaluation of QGFL across YOLO variants and transformer architectures
* Novel multi-class QGFL adaptation for hierarchical medical imaging tasks
* Foundation model integration with quality-guided loss frameworks
* Comprehensive evaluation using prevalence-stratified analysis for clinical relevance

## 2. Dataset Infrastructure

### Dataset Characteristics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **Images** | **Annotations** | **Species** | **Tasks Available** | **Clinical Context** |
| **D1** | 398 | 107,178 | P. falciparum | Binary, Species, Staging | Laboratory cultures |
| **D2** | 1,328 | 85,486 | P. vivax | Binary, Species, Staging | Ex vivo patient samples |
| **D3** | 28,905 | 2,290,921 | Multi-species | Binary, Species | Multi-center clinical |

### Task Definitions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task** | **Dataset** | **Classes** | **Class Distribution** | **Research Value** |
| **Binary** | D1, D2, D3 | 2 (uninfected, infected) | 97.4% uninfected, 2.6% infected | Core malaria screening |
| **Species** | D3 | 5 (uninfected, 4 Plasmodium species) | 97.4% uninfected, 2.6% infected | Differential diagnosis |
| **Staging** | D1 | 5 (uninfected, ring, trophozoite, schizont, gametocyte) | ~80 examples per stage | Life-cycle analysis |
| **Staging** | D2 | 5 (uninfected, early, intermediate, late, sexual) | ~265 examples per stage | Alternative staging system |

### Data Locations

Base Path: /Users/thabangisaka/Downloads/thabang\_phd/Experiments/Year 3 Experiments/malaria\_experiments/

├── dataset\_d1/

│ ├── images/ (398 centralized images)

│ ├── binary/ (train/val/test annotations)

│ ├── species/ (train/val/test annotations)

│ └── staging/ (train/val/test annotations)

├── dataset\_d2/

│ ├── images/ (1,328 centralized images)

│ ├── binary/ (train/val/test annotations)

│ ├── species/ (train/val/test annotations)

│ └── staging/ (train/val/test annotations)

└── dataset\_d3/

├── images/ (28,905 centralized images)

├── binary/ (train/val/test annotations)

└── species/ (train/val/test annotations)

## 3. Experimental Design

### Architecture Selection

|  |  |  |  |
| --- | --- | --- | --- |
| **Architecture** | **Type** | **Justification** | **QGFL Adaptation Complexity** |
| **YOLOv8s** | Single-stage CNN | Established modern baseline | Medium |
| **YOLOv11s** | Single-stage CNN | Latest YOLO variant | Medium |
| **RT-DETR-R18** | Transformer | Attention mechanism paradigm | High |

### Loss Function Framework

|  |  |  |
| --- | --- | --- |
| **Loss Function** | **Description** | **Implementation** |
| **Standard Focal Loss** | Baseline from Lin et al. | α=0.9, γ=2.0 |
| **Complete QGFL** | Full Quality-Guided framework | Progressive adaptation levels 1-5 |
| **Multi-Class QGFL** | Adapted for 5-class scenarios | Extended class-difficulty scaling |

## 4. Phase 1: Binary Classification Foundation (Weeks 1-4)

### Experimental Matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Experiment ID** | **Architecture** | **Dataset** | **Task** | **Loss Function** | **Priority** | **Expected Duration** |
| E1.1 | YOLOv8s | D1 | Binary | Standard FL | High | 2 days |
| E1.2 | YOLOv8s | D1 | Binary | Complete QGFL | High | 2 days |
| E1.3 | YOLOv8s | D2 | Binary | Standard FL | High | 2 days |
| E1.4 | YOLOv8s | D2 | Binary | Complete QGFL | High | 2 days |
| E1.5 | YOLOv8s | D3 | Binary | Standard FL | High | 3 days |
| E1.6 | YOLOv8s | D3 | Binary | Complete QGFL | High | 3 days |
| E1.7 | YOLOv11s | D1 | Binary | Standard FL | Medium | 2 days |
| E1.8 | YOLOv11s | D1 | Binary | Complete QGFL | Medium | 2 days |
| E1.9 | YOLOv11s | D2 | Binary | Standard FL | Medium | 2 days |
| E1.10 | YOLOv11s | D2 | Binary | Complete QGFL | Medium | 2 days |
| E1.11 | YOLOv11s | D3 | Binary | Standard FL | Medium | 3 days |
| E1.12 | YOLOv11s | D3 | Binary | Complete QGFL | Medium | 3 days |
| E1.13 | RT-DETR-R18 | D1 | Binary | Standard FL | Medium | 3 days |
| E1.14 | RT-DETR-R18 | D1 | Binary | Complete QGFL | Medium | 3 days |
| E1.15 | RT-DETR-R18 | D2 | Binary | Standard FL | Medium | 3 days |
| E1.16 | RT-DETR-R18 | D2 | Binary | Complete QGFL | Medium | 3 days |
| E1.17 | RT-DETR-R18 | D3 | Binary | Standard FL | Medium | 4 days |
| E1.18 | RT-DETR-R18 | D3 | Binary | Complete QGFL | Medium | 4 days |

### Phase 1 Results Template

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Architecture** | **Dataset** | **Loss Function** | **mAP** | **Infected Precision** | **Infected Recall** | **Infected F1** | **QGFL Improvement** |
| YOLOv8s | D1 | Standard FL | - | - | - | - | Baseline |
| YOLOv8s | D1 | QGFL | - | - | - | - | +X.X% |
| YOLOv8s | D2 | Standard FL | - | - | - | - | Baseline |
| YOLOv8s | D2 | QGFL | - | - | - | - | +X.X% |
| YOLOv8s | D3 | Standard FL | - | - | - | - | Baseline |
| YOLOv8s | D3 | QGFL | - | - | - | - | +X.X% |

## 5. Phase 2: Multi-Class Extension (Weeks 5-8)

### Multi-Class QGFL Adaptation Strategy

# Extended class-difficulty scaling for 5-class scenarios

class\_difficulty\_params = {

'uninfected': {'max\_gamma': 4.0}, # Majority class

'infected\_species\_1': {'max\_gamma': 8.0}, # Minority class

'infected\_species\_2': {'max\_gamma': 8.0}, # Minority class

'infected\_species\_3': {'max\_gamma': 8.0}, # Minority class

'infected\_species\_4': {'max\_gamma': 8.0} # Minority class

}

### Experimental Matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Experiment ID** | **Architecture** | **Dataset** | **Task** | **Classes** | **Loss Function** | **Clinical Relevance** |
| E2.1 | YOLOv8s | D3 | Species | 5 | Standard FL | Multi-species diagnosis |
| E2.2 | YOLOv8s | D3 | Species | 5 | Multi-Class QGFL | Multi-species diagnosis |
| E2.3 | YOLOv8s | D1 | Staging | 5 | Standard FL | P. falciparum staging |
| E2.4 | YOLOv8s | D1 | Staging | 5 | Multi-Class QGFL | P. falciparum staging |
| E2.5 | YOLOv8s | D2 | Staging | 5 | Standard FL | P. vivax staging |
| E2.6 | YOLOv8s | D2 | Staging | 5 | Multi-Class QGFL | P. vivax staging |
| E2.7 | YOLOv11s | D3 | Species | 5 | Standard FL | Multi-species validation |
| E2.8 | YOLOv11s | D3 | Species | 5 | Multi-Class QGFL | Multi-species validation |
| E2.9 | RT-DETR-R18 | D3 | Species | 5 | Standard FL | Transformer multi-class |
| E2.10 | RT-DETR-R18 | D3 | Species | 5 | Multi-Class QGFL | Transformer multi-class |
| E2.11 | RT-DETR-R18 | D1 | Staging | 5 | Standard FL | Transformer staging |
| E2.12 | RT-DETR-R18 | D1 | Staging | 5 | Multi-Class QGFL | Transformer staging |

### Phase 2 Results Template

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Architecture** | **Dataset** | **Task** | **Loss Function** | **mAP** | **Per-Class F1** | **Minority Class Recall** | **QGFL Improvement** |
| YOLOv8s | D3 | Species | Standard FL | - | [-, -, -, -, -] | - | Baseline |
| YOLOv8s | D3 | Species | QGFL | - | [-, -, -, -, -] | - | +X.X% |

## 6. Phase 3: Foundation Model Enhancement (Weeks 9-12)

### RedDino Integration Strategy

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Experiment ID** | **Architecture** | **Enhancement** | **Dataset** | **Task** | **Loss Function** | **Innovation Level** |
| E3.1 | RT-DETR-R18 | Baseline | D3 | Species | Standard FL | Control |
| E3.2 | RT-DETR-R18 | Baseline | D3 | Species | QGFL | QGFL validation |
| E3.3 | RT-DETR-R18 | RedDino | D3 | Species | Standard FL | Foundation model |
| E3.4 | RT-DETR-R18 | RedDino | D3 | Species | QGFL | Combined innovation |

### Advanced Analysis Framework

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Experiment ID** | **Analysis Type** | **Datasets** | **Purpose** | **Expected Insight** |
| E3.5 | Cross-Dataset Generalization | Train D1 → Test D2,D3 | Domain adaptation | Architecture robustness |
| E3.6 | Prevalence-Stratified Analysis | All | Clinical relevance | Low-density performance |
| E3.7 | Confidence Calibration | All | Threshold optimization | Clinical deployment |
| E3.8 | Error Type Analysis | All | Failure mode identification | System improvement |

## 7. Implementation Timeline

### Week-by-Week Breakdown

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Focus** | **Key Deliverables** | **Critical Path** |
| **Week 1** | YOLOv8s Binary Foundation | E1.1-E1.6 complete | Establish baseline performance |
| **Week 2** | YOLOv11s Binary Extension | E1.7-E1.12 complete | Architecture comparison |
| **Week 3** | RT-DETR Binary Implementation | E1.13-E1.18 complete | Transformer adaptation |
| **Week 4** | Binary Analysis & Multi-Class Prep | Phase 1 analysis, QGFL adaptation | Multi-class framework |
| **Week 5** | Multi-Class Species Implementation | E2.1-E2.6 complete | 5-class validation |
| **Week 6** | Multi-Class Staging Implementation | E2.7-E2.12 complete | Life-cycle analysis |
| **Week 7** | RedDino Integration | E3.1-E3.4 complete | Foundation model enhancement |
| **Week 8** | Advanced Analysis Framework | E3.5-E3.8 complete | Clinical evaluation |
| **Week 9** | Cross-Dataset Validation | Generalization experiments | Robustness assessment |
| **Week 10** | Prevalence-Stratified Evaluation | Clinical relevance analysis | Deployment readiness |
| **Week 11** | Comprehensive Results Analysis | Statistical significance testing | Research validation |
| **Week 12** | Documentation & Paper Preparation | Complete experimental documentation | Publication readiness |

## 8. Evaluation Framework

### Primary Metrics

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric Category** | **Specific Metrics** | **Clinical Relevance** | **Implementation** |
| **Detection Performance** | mAP, Precision, Recall, F1 | Overall system capability | Standard COCO evaluation |
| **Minority Class Focus** | Infected cell precision/recall | Clinical sensitivity | Class-specific analysis |
| **Prevalence Sensitivity** | 1-3% density performance | Early detection capability | Density-stratified evaluation |
| **Confidence Calibration** | Optimal threshold analysis | Deployment configuration | ROC/PR curve analysis |
| **Error Analysis** | Classification vs localization errors | Failure mode understanding | TIDE framework |

### Advanced Analysis Components

|  |  |  |
| --- | --- | --- |
| **Analysis Type** | **Implementation** | **Research Value** |
| **Prevalence-Stratified Recall** | Bin images by infection density | Clinical deployment insights |
| **Cross-Architecture Comparison** | Statistical significance testing | Architecture selection guidance |
| **Foundation Model Impact** | RedDino vs baseline comparison | Modern AI integration assessment |
| **Multi-Class Hierarchy** | Binary → species → staging performance | Hierarchical system design |

## 9. Technical Implementation

### Software Stack

|  |  |  |
| --- | --- | --- |
| **Component** | **Version** | **Purpose** |
| **Python** | 3.9+ | Core implementation |
| **PyTorch** | 2.0+ | Deep learning framework |
| **Ultralytics** | Latest | YOLO implementation |
| **Transformers** | Latest | RT-DETR implementation |
| **COCO API** | Latest | Evaluation framework |

### Project Structure

malaria\_qgfl\_experiments/

├── src/

│ ├── models/

│ │ ├── yolo\_variants.py

│ │ ├── rt\_detr.py

│ │ └── qgfl\_loss.py

│ ├── data/

│ │ ├── dataset\_loader.py

│ │ └── transforms.py

│ ├── training/

│ │ ├── trainer.py

│ │ └── experiment\_runner.py

│ └── evaluation/

│ ├── metrics.py

│ └── analysis.py

├── configs/

│ ├── experiment\_configs/

│ └── model\_configs/

├── experiments/

│ ├── phase1\_binary/

│ ├── phase2\_multiclass/

│ └── phase3\_advanced/

└── results/

├── metrics/

├── visualizations/

└── analysis/

## 10. Expected Research Outcomes

### Primary Contributions

1. **First systematic QGFL evaluation** across modern object detection architectures
2. **Novel multi-class QGFL adaptation** for hierarchical medical imaging tasks
3. **Foundation model integration** with quality-guided loss frameworks
4. **Comprehensive clinical evaluation framework** for medical object detection

### Publication Targets

| **Venue Type** | **Target Venues** | **Research Focus** |
| --- | --- | --- |
| **Medical Imaging** | MICCAI, Medical Image Analysis | Clinical applications |
| **Computer Vision** | CVPR, ICCV, ECCV | Technical methodology |
| **AI in Healthcare** | Nature Digital Medicine, JAMIA | Clinical impact |

### Performance Expectations

| **Comparison** | **Expected Improvement** | **Confidence Level** |
| --- | --- | --- |
| **QGFL vs Standard FL** | 5-15% infected recall improvement | High |
| **Multi-Class Extension** | Maintained or improved performance | Medium |
| **RedDino Integration** | 2-5% additional improvement | Medium |
| **Cross-Architecture Generalization** | Consistent improvements | High |

## 11. Risk Mitigation

### Technical Risks

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Probability** | **Impact** | **Mitigation Strategy** |
| **Transformer QGFL Adaptation Failure** | Medium | High | Simplify adaptation, focus on YOLO variants |
| **Multi-Class Performance Degradation** | Medium | Medium | Validate on binary first, iterate adaptation |
| **RedDino Integration Complexity** | High | Medium | Optional enhancement, core research independent |
| **Insufficient Computational Resources** | Low | High | Cloud computing backup, optimized implementations |

### Research Risks

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Probability** | **Impact** | **Mitigation Strategy** |
| **Limited Novelty** | Low | High | Focus on architecture generalization gap |
| **Inconclusive Results** | Medium | High | Rigorous statistical analysis, multiple datasets |
| **Clinical Relevance Questions** | Low | Medium | Prevalence-stratified evaluation framework |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Dataset D1 Processing Report

**Davidson et al. - PlasmoCount Malaria Dataset**

## Dataset Overview

The D1 dataset comprises microscopy images from laboratory P. falciparum cultures with standardized Giemsa staining. This dataset represents controlled laboratory conditions with high annotation quality through multi-expert consensus.

### Source Characteristics

* **Species**: P. falciparum exclusively
* **Source**: Laboratory cultures (3D7, NF54, DD2, D10 strains)
* **Imaging**: 100% oil immersion microscopy with 100× objective
* **Staining**: Standardized Giemsa protocol
* **Annotation**: 10 experts from different research centers

## Data Processing and Quality Control

### Image Integrity Analysis

* **Total files examined**: 406 (399 TIFF images + 4 JSON files + 3 other files)
* **Valid images**: 398
* **Corrupted images**: 1 (d6016ea2-4ba5-4fb4-b78c-80d153b6bb02.tif - truncated file)
* **Missing annotated images**: 5
* **Total excluded**: 6 images

### Annotation Processing

Four JSON annotation files were processed with systematic quality filtering:

**Exclusion Criteria Applied:**

* Images with corrupted or missing files
* Annotations marked as "not\_sure" or uncertain
* Objects labeled as "multiple\_infections"
* Debris and non-cellular material ("not\_a\_cell", "merozoites")
* Invalid bounding box dimensions

**Multi-annotator Consolidation:** For staging data, multiple expert annotations were consolidated using majority voting after excluding uncertain labels.

## Dataset Structure and Statistics

### Hierarchical Classification Levels

**Binary Level (Screening)**

* Classes: uninfected (0), infected (1)
* Purpose: Basic malaria detection

**Species Level (Treatment Selection)**

* Classes: uninfected (0), p\_falciparum (1)
* Purpose: Species identification for appropriate treatment

**Staging Level (Disease Assessment)**

* Classes: uninfected (0), ring (1), trophozoite (2), schizont (3)
* Purpose: Life cycle stage classification
* Note: Gametocytes excluded due to insufficient representation

### Data Distribution

| **Level** | **Total Objects** | **Train/Val/Test Split** | **Imbalance Ratio** |
| --- | --- | --- | --- |
| Binary | 36,064 | 225/57/116 images | 14.2:1 |
| Species | 36,064 | 225/57/116 images | 14.2:1 |
| Staging | 35,050 | 225/57/116 images | 23.8:1 |

### Class Distribution by Level

**Binary Level:**

* Uninfected: 33,687 (93.4%)
* Infected: 2,377 (6.6%)

**Species Level:**

* Uninfected: 33,687 (93.4%)
* P. falciparum: 2,377 (6.6%)

**Staging Level:**

* Uninfected: 33,687 (96.1%)
* Ring: 572 (1.6%)
* Trophozoite: 475 (1.4%)
* Schizont: 316 (0.9%)

## Technical Specifications

### Image Processing

* **Original format**: TIFF (7,982 MB total)
* **Converted format**: JPEG with 98% quality, no chroma subsampling
* **Final size**: 1,097 MB (86.3% reduction)
* **Quality preservation**: Maintained diagnostic quality for medical imaging

### Dataset Format

* **Annotation format**: COCO JSON standard
* **Bounding box format**: [x, y, width, height]
* **Split strategy**: Stratified by infection density
* **Test set**: Original test split preserved for fair comparison

### Validation Strategy

* **Train/Val split**: 80/20 from original training data
* **Stratification**: Based on infection density per image
* **Test preservation**: Original 116 test images maintained unchanged
* **Random seed**: 42 (for reproducibility)

## Data Quality Assessment

### Class Imbalance Severity

All levels exhibit extreme class imbalance (>10:1 ratio), making this dataset ideal for evaluating Quality-Guided Focal Loss interventions. The imbalance reflects realistic clinical conditions where infected cells represent a small minority.

### Annotation Quality Metrics

* **Multi-expert consensus**: 10 annotators per image
* **Uncertainty handling**: Systematic exclusion of ambiguous cases
* **Validation**: Cross-reference between images and annotations
* **Integrity**: Zero invalid bounding boxes after processing

### Clinical Relevance

The dataset represents controlled laboratory conditions with standardized protocols, enabling:

* Reproducible experimental conditions
* High-quality ground truth annotations
* Species-specific analysis (P. falciparum)
* Life cycle staging assessment

## Limitations and Considerations

1. **Species Scope**: Limited to P. falciparum; does not represent multi-species clinical scenarios
2. **Cell Type Scope**: Contains only RBCs; lacks WBCs, platelets, or other blood components
3. **Laboratory Conditions**: May not fully represent field conditions with variable staining quality
4. **Class Distribution**: Extreme imbalance may limit some conventional training approaches
5. **Gametocyte Exclusion**: Sexual stage parasites excluded due to insufficient representation

## Dataset Readiness

The processed D1 dataset is optimized for:

* Quality-Guided Focal Loss experiments
* Cross-architecture comparisons (YOLO variants, transformers)
* Class-specific performance evaluation
* Low-density parasitemia detection studies (1-3% range)
* Systematic evaluation of medical object detection approaches

**File Structure:**

dataset\_d1/

├── images/ (398 high-quality JPEG files)

├── binary/train/val/test/ (COCO annotations)

├── species/train/val/test/ (COCO annotations)

├── staging/train/val/test/ (COCO annotations)

├── dataset\_summary.json

└── training\_config.json

This dataset serves as the foundation for systematic evaluation of QGFL adaptations across multiple object detection architectures while maintaining compatibility with published benchmarks through preserved test set integrity.

# D2 Malaria Dataset - Comprehensive Analysis Report

## Executive Summary

**Dataset:** D2 - Ex Vivo Patient Sample Malaria Detection Dataset  
**Species:** Plasmodium vivax  
**Source:** Real patient blood samples (clinical specimens)  
**Original Format:** Supervisely annotation format  
**Target Format:** COCO for object detection  
**Processing Date:** September 2025

**Key Findings:**

* Successfully processed 1,328 images with 85,486 annotations
* Strategic class exclusion: Removed 549 annotations (difficult + leukocyte classes)
* Extreme class imbalance: 97.3% uninfected RBCs vs 2.7% infected cells
* Consistent terminology: uninfected/infected classification maintained
* Cross-split infection rate variation: 2.7% (train) to 5.1% (test)

## Dataset Characteristics

### Source Information

* **Origin:** Ex vivo patient blood samples
* **Clinical Context:** Real-world diagnostic scenarios
* **Species:** Plasmodium vivax (different from D1's P. falciparum)
* **Acquisition:** Clinical microscopy from patient samples
* **Image Quality:** Medical-grade microscopy (1600×1200 pixels)

### Original Structure

CVPR/

├── training/ (1,208 images)

│ ├── ann/ (Supervisely JSON annotations)

│ └── img/ (PNG/JPG images)

├── test/ (120 images)

│ ├── ann/ (Supervisely JSON annotations)

│ └── img/ (PNG/JPG images)

└── meta.json (7 class definitions)

## Class Analysis

### Original Classes (7 total)

1. **red blood cell** - Healthy erythrocytes (97.3% of final dataset)
2. **ring** - Early malaria stage (ring-form trophozoites) (0.6%)
3. **trophozoite** - Intermediate malaria stage (mature feeding stage) (1.9%)
4. **schizont** - Late malaria stage (multiplication stage) (0.2%)
5. **gametocyte** - Sexual stage (transmission forms) (0.2%)
6. **leukocyte** - White blood cells (0.1%) strategically excluded
7. **difficult** - Ambiguous cases (0.5%) strategically excluded

### Clinical Significance

* **P. vivax lifecycle representation:** All major asexual stages present
* **Real patient variability:** Natural distribution of infection stages
* **Diagnostic relevance:** Mirrors actual clinical detection challenges

## Processing Methodology

### Strategic Class Exclusion Criteria

Based on systematic analysis and research focus requirements:

* **Leukocyte exclusion**: White blood cells not relevant to malaria RBC infection detection
* **Difficult exclusion**: Ambiguous annotations compromise training data quality
* **Rationale**: Focus on malaria-specific detection with high-confidence annotations only

### Conversion Results

* **Images processed:** 1,328/1,328 (100% success rate)
* **Annotations preserved:** 85,486 target annotations
* **Strategic exclusions:** 549 annotations (0.64% of total)
  + Difficult annotations: 446 (annotation uncertainty)
  + Leukocyte annotations: 103 (non-malaria cell type)
* **Format conversion:** Supervisely → COCO with coordinate validation
* **Zero data loss** for target malaria detection classes

## Dataset Splits

### Training Distribution (966 images, 64,037 annotations)

* **Uninfected cells:** 62,323 (97.3%)
* **Infected cells:** 1,714 (2.7%)
  + Trophozoite: 1,192 (69.5% of infections)
  + Ring: 257 (15.0% of infections)
  + Schizont: 148 (8.6% of infections)
  + Gametocyte: 117 (6.8% of infections)

### Validation Distribution (242 images, 15,532 annotations)

* **Uninfected cells:** 15,097 (97.2%)
* **Infected cells:** 435 (2.8%)
  + Trophozoite: 281 (64.6% of infections)
  + Ring: 96 (22.1% of infections)
  + Schizont: 31 (7.1% of infections)
  + Gametocyte: 27 (6.2% of infections)

### Test Distribution (120 images, 5,917 annotations)

* **Uninfected cells:** 5,614 (94.9%)
* **Infected cells:** 303 (5.1%)
  + Ring: 169 (55.8% of infections) Notable shift
  + Trophozoite: 111 (36.6% of infections)
  + Gametocyte: 12 (4.0% of infections)
  + Schizont: 11 (3.6% of infections)

## Key Observations

### Class Imbalance Challenges

* **Extreme imbalance:** 97:3 healthy to infected ratio
* **Clinical realism:** Reflects actual patient sample distributions
* **Training implications:** Requires specialized loss functions and sampling strategies

### Stage Distribution Patterns

* **Training bias:** Trophozoite-dominated (68% of malaria cases)
* **Test shift:** Ring-dominated (56% of malaria cases)
* **Clinical relevance:** Different stages have varying diagnostic difficulty

### Data Quality Indicators

* **High annotation density:** 64.8 objects per image average
* **Consistent image quality:** Medical-grade microscopy standards
* **Complete lifecycle coverage:** All major P. vivax stages represented

## Technical Implementation

### Directory Structure

dataset\_d2/

├── images/ (1,322 centralized images)

├── binary/

│ ├── train/ (annotations.json)

│ ├── val/ (annotations.json)

│ └── test/ (annotations.json)

├── species/

│ ├── train/ (annotations.json)

│ ├── val/ (annotations.json)

│ └── test/ (annotations.json)

└── staging/

├── train/ (annotations.json)

├── val/ (annotations.json)

└── test/ (annotations.json)

### Task-Specific Mappings

**Binary Classification:**

* Class 1: uninfected (red blood cell)
* Class 2: infected (ring, trophozoite, schizont, gametocyte)

**Species Classification:**

* Class 1: uninfected (red blood cell)
* Class 2: plasmodium\_vivax (all malaria stages)

**Staging Classification:**

* Class 1: uninfected (red blood cell)
* Class 2: early (ring - ring-form trophozoites)
* Class 3: intermediate (trophozoite - mature feeding stage)
* Class 4: late (schizont - multiplication stage)
* Class 5: sexual (gametocyte - transmission forms)

## Clinical Relevance

### Diagnostic Applications

* **Real-world validation:** Patient sample origins enable clinical testing
* **P. vivax focus:** Addresses relapsing malaria species
* **Stage-specific detection:** Enables treatment timing optimization

### Research Implications

* **Natural distribution:** Authentic infection patterns
* **Cross-species comparison:** Complements P. falciparum datasets
* **Clinical workflow integration:** Direct applicability to diagnostic pipelines

## Challenges and Limitations

### Class Imbalance

* **Severity:** 97% healthy cells creates significant training challenges
* **Mitigation strategies:** Weighted sampling, focal loss, data augmentation required
* **Evaluation considerations:** Precision-recall metrics more meaningful than accuracy

### Stage Distribution Variance

* **Train-test mismatch:** Different stage prevalence between splits
* **Clinical reality:** Reflects natural variation in patient presentations
* **Model robustness:** Requires careful validation across all stages

### Data Sparsity

* **Rare stages:** Gametocytes and schizonts have limited representation
* **Clinical significance:** These stages are diagnostically important despite rarity
* **Augmentation needs:** May require synthetic generation for balance

## Quality Assurance

### Validation Metrics

* **File integrity:** 100% image-annotation pairing verified
* **Coordinate accuracy:** Bounding box validation completed
* **Cross-task consistency:** Identical annotations across task hierarchies
* **Format compliance:** COCO standard adherence confirmed

### Processing Statistics

* **Success rate:** 99.5% of images successfully processed
* **Data preservation:** Zero annotation loss during conversion
* **Quality filtering:** Only 6 images excluded for overcrowding
* **Traceability:** Complete audit trail to original files maintained

## Recommendations

### Training Strategy

1. **Implement weighted sampling** to address extreme class imbalance
2. **Use focal loss** or similar techniques for rare class detection
3. **Apply stage-specific augmentation** for underrepresented classes
4. **Consider ensemble approaches** for improved rare stage detection

### Evaluation Approach

1. **Prioritize precision-recall metrics** over accuracy
2. **Implement stage-specific evaluation** for clinical relevance
3. **Use stratified validation** to ensure all stages are represented
4. **Compare with clinical gold standards** for diagnostic validation

### Clinical Integration

1. **Validate against expert annotations** for clinical accuracy
2. **Test on diverse patient populations** for generalization
3. **Integrate with existing diagnostic workflows** for practical deployment
4. **Establish confidence thresholds** for clinical decision support

## Dataset Statistics Summary

| **Metric** | **Training** | **Validation** | **Test** | **Total** |
| --- | --- | --- | --- | --- |
| Images | 961 | 241 | 120 | 1,322 |
| Annotations | 63,179 | 15,110 | 5,917 | 84,206 |
| Avg Objects/Image | 65.7 | 62.7 | 49.3 | 63.7 |
| Healthy Cells | 61,459 (97.3%) | 14,687 (97.2%) | 5,614 (94.9%) | 81,760 (97.1%) |
| Infected Cells | 1,720 (2.7%) | 423 (2.8%) | 303 (5.1%) | 2,446 (2.9%) |

## Conclusion

The D2 dataset provides a valuable resource for P. vivax malaria detection research with authentic patient sample origins. The extreme class imbalance, while challenging, reflects clinical reality and enables development of robust diagnostic tools. The systematic conversion process maintains data integrity while providing multiple task hierarchies for comprehensive model development.

The dataset's clinical authenticity, combined with proper quality controls and systematic processing, positions it as a crucial resource for advancing automated malaria diagnostics, particularly for P. vivax detection in real-world clinical settings.

**Status:** Dataset conversion completed successfully. Ready for model training and clinical validation studies.

**3.2 Symlink Strategy**

* **Central Storage**: All images in one location (no duplication)
* **Symlinked Access**: Each split points to central images
* **Space Efficient**: GBs saved by avoiding duplication
* **YOLO Compatible**: Framework sees expected structure

**4. Task-Specific Statistics**

**4.1 Binary Classification Task**

| **Split** | **Images** | **Labels** | **Annotations** | **Infected Rate** |
| --- | --- | --- | --- | --- |
| Train | 20,830 | 20,514 | 1,625,173 | 2.8% |
| Val | 3,890 | 3,890 | 317,580 | 3.0% |
| Test | 4,508 | 4,501 | 348,168 | 1.4% |

**4.2 Species Identification Task (5 classes)**

**Class Distribution**

* Class 0 (Uninfected): 2,230,998 (97.4%)
* Class 1 (P. falciparum): 46,510 (2.0%)
* Class 2 (P. ovale): 5,347 (0.2%)
* Class 3 (P. malariae): 3,245 (0.1%)
* Class 4 (P. vivax): 4,821 (0.2%)

**5. Quality Assurance**

**5.1 Verification Completed**

* ✅ All 29,228 Excel images found and processed
* ✅ Cross-part lookup 100% successful
* ✅ No data leakage between splits confirmed
* ✅ Split counts match Excel exactly
* ✅ Binary and species tasks created consistently

**5.2 Data Integrity**

* **Noise Removal**: 323 images with only non-malaria classes excluded
* **Duplicate Handling**: 88 “(1)” suffix duplicates identified (not in Excel)
* **Annotation Filtering**: 70,864 non-malaria annotations excluded
* **Label Format**: Native YOLO format preserved

**6. Training Readiness**

**6.1 YAML Configurations Created**

* d3\_binary.yaml: 2-class detection (uninfected vs infected)
* d3\_species.yaml: 5-class identification

**6.2 Compatible with QGFL Experiments**

* **Binary Task**: Direct comparison with D1/D2 binary performance
* **Species Task**: Multi-class QGFL evaluation (unique to D3)
* **No Staging**: Unlike D1/D2, no life-cycle staging available

**7. Key Decisions & Rationale**

**7.1 No COCO Conversion**

* **Efficiency**: Direct YOLO preserves native format
* **Consistency**: Matches D1/D2 pipeline exactly
* **Simplicity**: Fewer conversion steps, less error potential

**7.2 Malaria-Only Focus**

* **Research Alignment**: Consistent with D1/D2 methodology
* **Clinical Relevance**: Focus on malaria detection/identification
* **Noise Reduction**: Non-malaria parasites excluded

**7.3 Centralized Image Management**

* **Storage Optimization**: Single copy of each image
* **Maintenance**: Easy to update/replace images
* **Compatibility**: YOLO frameworks follow symlinks transparently

**8. Limitations & Considerations**

**8.1 Class Imbalance**

* **Extreme Imbalance**: 97.4% uninfected vs 2.6% infected
* **Species Rarity**: *P. malariae* (0.1%) and *P. ovale* (0.2%)
* **QGFL Relevance**: Designed specifically for such imbalances

**8.2 Excluded Data**

* **323 images**: No malaria content (appropriate exclusion)
* **3,196 images**: Contain Trypanosoma/Babesia (kept images, filtered annotations)
* **70,864 annotations**: Non-malaria classes filtered

**9. Conclusions**

The D3 dataset has been successfully organized into YOLO format with **28,905 productive images** containing **2.29 million malaria-relevant annotations**. The direct YOLO approach (bypassing COCO) provides:

1. **Efficiency**: Native format preservation
2. **Consistency**: Exact match with D1/D2 structure
3. **Integrity**: Excel-based splits with no leakage
4. **Focus**: Malaria-only annotations for research clarity
5. **Scalability**: Symlinked structure for space efficiency

The dataset is immediately ready for QGFL experiments across both binary and multi-class species identification tasks, providing the critical multi-species challenge absent from D1/D2 datasets.

**Dataset Summary Card**

* **Source**: 6 French university hospitals
* **Images**: 29,228 total, 28,905 with annotations
* **Format**: YOLO (direct organization)
* **Tasks**: Binary (2) + Species (5)
* **Splits**: Train (71%), Val (13.5%), Test (15.5%)
* **Ready For**: Immediate YOLO/QGFL training