Enhancing Soil-Transmitted Helminths Diagnosis Through AI: A Self-supervised Learning Approach with Smartphone-Based Digital Microscopy

Lin Lin[1,2](✉), Daniel Cuadrado[1], Roberto Mancebo-Martín[1], Stella Kepha[3], Paul Gichuki[3], Charles Mwandawiro[3], María Jesús Ledesma-Carbayo[2], Miguel Luengo-Oroz[1], Elena Dacal[1], and David Bermejo-Peláez[1]

[1] Spotlab, R&D, Madrid, Spain
lin@spotlab.ai
[2] ETSI Telecomunicación, Universidad Politécnica de Madrid & CIBER-BBN, Madrid, Spain
[3] ESACIPAC, Kenya Medical Research Institution (KEMRI), Nairobi, Kenya

Abstract. Soil-transmitted helminths (STH), including hookworm, *Ascaris lumbricoides*, and *Trichuris trichiura*, impose significant health burdens in low- and middle-income tropical and subtropical regions, infecting over 1.5 billion people globally. Traditional diagnostic methods like the Kato-Katz technique are time consuming. This study introduces an innovative AI-driven system utilizing affordable 3D-printed adapters and smartphones to digitize Kato-Katz microscopy samples, capturing high-resolution images for subsequent analysis. These digitized images can be uploaded to a telemedicine platform for remote diagnosis and expert consultation.

Central to our system is the development of a foundational AI model for parasite detection and classification. The model operates in two stages: First, an object detection algorithm identifies all parasites in the image, achieving a mean average precision (mAP) of 97.90% on the validation set using the YOLOv8 architecture. Second, a classification algorithm categorizes each detected parasite by species. The classification model is initially trained on a large, unannotated dataset of parasite images using a self-supervised learning (SSL) approach to learn domain-specific visual features, which are often missed while using generic pretraining datasets. Subsequently, it is fine-tuned on a labeled dataset, significantly improving performance. The model initialized with SSL on STH images achieved an F1 score of 91.70%, outperforming those initialized with random weights (F1 score of 55%) and those trained on DINO-Imagenet weights (F1 score of 53%).

By integrating AI with low-cost digital imaging, our approach aims to revolutionize STH diagnosis in resource-constrained settings, aligning with the WHO's 2030 Roadmap for the elimination of neglected tropical diseases.

Keywords: STH · NTD · AI · foundation model · object detection · self-supervised learning

U. Anazodo et al. (Eds.): MImA 2024/EMERGE 2024, CCIS 2240, pp. 144–152, 2025.
https://doi.org/10.1007/978-3-031-79103-1_15

1 Introduction

Soil-transmitted helminths (STH), including hookworm, *Ascaris lumbricoides*, and *Trichuris trichiura*, are common in low- and middle-income tropical and subtropical countries. Over 1.5 billion people are infected, suffering from anemia, gastrointestinal distress, and chronic fatigue [1]. The World Health Organization (WHO) estimates STH infections cause over 3 million disability-adjusted life years (DALYs) lost annually. The WHO 2030 Roadmap for NTDs aims to eliminate these parasites through mass drug administration (MDA) with albendazole and mebendazole [2, 3].

The Kato-Katz technique, used to diagnose STH infections, involves preparing stool samples on microscope slides for visual inspection [4]. While simple and cost-effective, its sensitivity drops if samples are not examined within 30 to 60 min due to egg degradation or hatching, especially with hookworms. Subjective visual assessment also leads to variability and errors in diagnosis.

AI integration in medical imaging has transformed fields like radiology and cardiology in high-income countries. However, its use in low- and middle-income countries (LMICs) is limited. Developing accessible and reliable AI solutions for these regions is vital for global healthcare. AI can enhance the diagnosis and management of diseases like STH infections, common in LMICs [5–8].

AI algorithms for medical imaging need abundant labeled data for training. Self-supervised learning (SSL) has emerged as an alternative, learning features from large unlabeled datasets to reduce the need for labeled data [9].

This work proposes a system that digitizes Kato-Katz samples using a 3D-printed adapter and smartphones. This method captures and stores high-resolution images of stool samples shortly after preparation, preserving their quality for later analysis. These images can be uploaded to a telemedicine platform for remote diagnosis and expert opinions. We also developed a foundation model for stool parasites using a SSL approach, allowing visual representation acquisition without labeled images. This promising foundation AI algorithm is capable of analyzing a single stool sample to detect and classify multiple types of parasites simultaneously.

2 Material and Methods

2.1 Dataset

We collected an extensive dataset composed of 1,380 stool samples from children aged between 5–15 years in Kwale, Kenya. Each stool sample was prepared using the Kato-Katz thick smear method and visually analyzed by conventional microscopy. In parallel, each stool sample was digitized by taking pictures of the field of view (FoV) and the images were transferred to a telemedicine platform. Both processes were made at 100x magnification (~0.08 μm/pixel).

The proposed digitization system is based on a 3D-printed adapter that allows coupling a smartphone to a conventional microscope by aligning the smartphone camera with the objective of the microscope to acquire the images. This adapter can convert any conventional microscope into a digital one and enables the digitization of microscope samples without the need for expensive scanners. Additionally, the system has been designed to be universal, working with any microscope model and any smartphone model.

From this dataset, a total of 163 stool samples (3,075 FoV images) were further analyzed, with all visible parasites labeled by species, including *Ascaris lumbricoides*, *Trichuris trichiura* and hookworm. This annotated dataset was split at patient level into training (65%), validating (20%) and testing (15%). This split was used for both the parasite detection algorithm and the species classification. All images from the remaining unannotated stool samples (1,217 stool samples with 14,680 FoV images) were used for training the self-supervised phase of the pipeline.

For detection, the entire FoV is used, while classification relies on patches cropped from areas with identified parasites. Instead of random cropping for the unsupervised dataset, we used a trained object detector to identify possible parasites, significantly increasing the amount of relevant data. Table 1 illustrates the dataset distribution, which is maintained consistently for both the detection and classification tasks.

Table 1. Data Distribution: Breakdown of unannotated (SSL) and annotated (Train; Validation; Test) data sets. Expert labels include *Ascaris*, *Trichuris*, and hookworm classes, along with artifacts (false positives generated by the detection algorithm). "Images" represents the number of FoV images, while "Total" indicates the number of parasites, which corresponds to the number of cropped patches.

Data set	#Patients	#Images	*Ascaris*	*Trichuris*	Hookworm	Artifact	Total
SSL	1,217	14,680	-	–	–	–	104,885
Train	84	1,637	6,294	1,283	586	2,253	10,416
Validation	43	817	1,351	698	325	937	3,311
Test	36	621	899	811	201	932	2,843

2.2 Overview of the Proposed Foundational Method

The proposed foundational model for parasite detection and classification operates as follows: First, an object detection algorithm identifies all parasites present in an image, regardless of their species. Then, a classification algorithm categorizes each parasite into its specific species. This classification algorithm is based on a foundational model for species differentiation. Initially, it is trained on a large unannotated dataset of parasite images to learn visual representations of stool parasites using a self-supervised approach. After learning these domain specific image-based features, the algorithm is further fine-tuned on a labeled dataset to accurately discern the species of each parasite. Figure 1 illustrates the proposed approach.

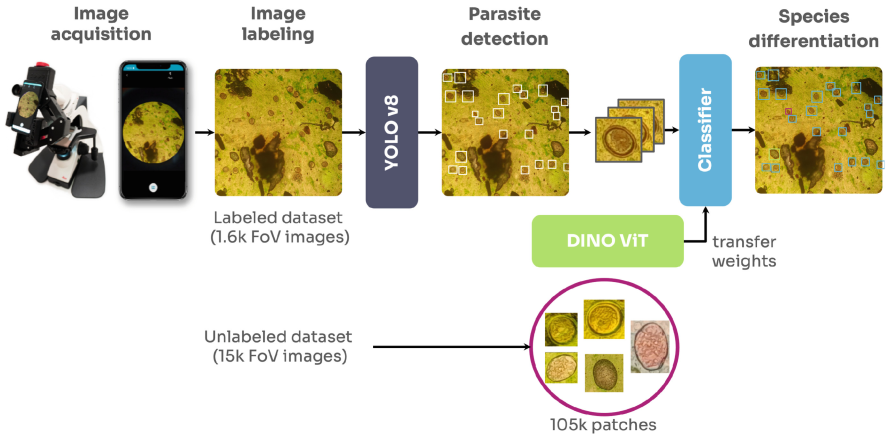

Fig. 1. Overview of the proposed pipeline comprising: (1) image acquisition using a 3D-printed adapter; (2) Image labeling; (3) parasite detection using YOLOv8; (4) self-supervised pre-training with DINO; (5) species differentiation with our DINO pretrained ViT classifier.

2.3 Slide-Level Parasite Detection

Each field of view image is processed through an object detection algorithm to detect all possible parasites regardless of the species. The proposed algorithm for parasite detection in stool sample images was based on a YOLO architecture (YOLOv8) [10], which is a single-stage object detection algorithm that uses a convolutional neural network (CNN) as backbone. Unlike two-stage algorithms, single-stage detection models like YOLO offer enhanced processing speed, making them highly suitable for mobile deployment. For each detected parasite, an image patch is then extracted and further processed by the patch-level parasite classification to determine the specie.

2.4 Patch-Level Parasite Classification

In this study, we aimed to enhance stool parasite differentiation by utilizing SSL to train a feature extractor (backbone) for better data generalization using unlabeled datasets. Specifically, we propose the use of a vision transformer (ViT) [11] (ViT-S/16) as the backbone, which has achieved significant success in various domains and whose application in the medical field is rapidly expanding. During the pre-training (self-supervised phase), we adopted the DINO technique [12], a knowledge distillation method that does not require labeled data.

ViT operates by dividing an image of fixed-size patches (NxN), each patch is passed to a linear operator to obtain patch embedding. To preserve spatial information of each patch, their position is encoded and is added later to each patch embedding. The resulting sequence is fed to the transformer encoder. In order to perform classification, an extra learnable classification token (CLS) is added to the patch embedding, and the result is passed to a linear classification head to classify the image. This classification token and head is not required for SSL.

The general concept of DINO is summarized in Fig. 2. It has two models that follow the same architecture, teacher and student, parameterized by t and s. For an input image I, patches of different sizes were generated: large patches (global crops) and small patches (local crops). All crops are followed by extensive augmentation. For a crop x, pair of view (x1, x2) were generated with random augmentation, both models produce output probabilities Pt and Ps, obtained by normalizing the output of the model with a softmax function. The student model learns to match distribution by minimizing the cross-entropy loss mins CE(Pt(x), Ps(x)). The parameters of the student s are updated by stochastic gradient descent, and the parameters of the teacher t are updated using the exponential moving average (EMA) of the students.

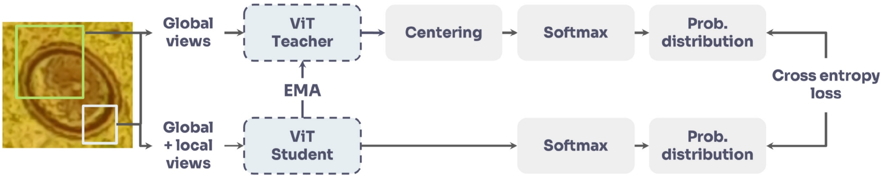

Fig. 2. Overview of the proposed pipeline for training a ViT architecture with DINO SSL strategy.

After pre-training, we conducted supervised training. In this phase, we add a multilayer perceptron (MLP) layer on top of the ViT encoder to perform classification. To ensure the need for SSL pretraining we tried two different approaches: first, by freezing all weights of the architecture except those from the classification layer (linear probing), and second, by enabling fine-tuning of all weights in the architecture, including the ViT encoder.

2.5 Experimental Setup

For comparative purposes, we compared different versions of the YOLOv8 architecture for parasite detection. This comparison aimed primarily to identify the optimal architecture for deployment on edge devices, such as smartphones, enabling real-time detections. Testing multiple YOLOv8 variants allowed us to determine their performance and suitability for this specific application.

Additionally, we trained the ViT model without pre-training to assess the improvement conferred by SSL. The comparison of these models with and without pre-training provided insights into the benefits of SSL in improving model performance.

All comparisons, including those for parasite detection and classification, were conducted on a validation set to ensure consistency and reliability in the performance assessments. The optimal configurations, determined through these validation tests, were then evaluated on an independent test set to confirm their efficacy and generalizability. This comprehensive approach allowed us to identify the best-performing models for both tasks and ensure their readiness for practical deployment.

The metrics used for evaluating the object detection algorithm included mean average precision (mAP), the precision and recall. The performance of the classification algorithm was assessed by measuring the balanced accuracy (BACC) and F1-score to account for data imbalance and to provide a more comprehensive evaluation of the classifier's effectiveness across all classes.

3 Experiments and Results

3.1 Slide-Level Parasite Detection

We conducted a set of experiments to evaluate the performance of various YOLOv8 models: YOLOv8-n (nano), YOLOv8-s (small), YOLOv8-m (medium), and YOLOv8-l (large). Instead of training the model from scratch, we used the weights pretrained with COCO image dataset and fine-tuned on our dataset. Our experiments utilized an input size of 640 × 640 pixels, a learning rate set to 0.01, and we trained each model for 100 epochs with early stopping implemented after 20 epochs without improvement by monitoring the loss in the validation set. We evaluated the mAP, precision, recall, and inference time of each one.

Table 2 illustrates the performance of each YOLOv8 model. Our results indicate that all four models achieved a comparable mAP of 97%. However, YOLOv8-n demonstrated significantly faster inference times, being approximately three times quicker than YOLOv8-l. This makes YOLOv8-n particularly well-suited for deployment on edge devices where computational resources are limited.

Table 2. Parasite detection performance on the validation set. Note: the inference time was calculated on an Intel i5 CPU processor.

Architecture	mAP	Precision	Recall	Inference time/image (ms)
YOLOv8-n	97.07	**91.27**	87.19	**372**
YOLOv8-s	**97.79**	90.72	91.36	465
YOLOv8-m	97.47	89.18	92.33	696
YOLOv8-l	97.04	87.02	**94.36**	1060

3.2 Patch-Level Parasite Classification

During the self-supervised training phase, we employed the following setup: a ViT-S/16 model with a patch size of 16. The model underwent training for 200 epochs with a batch size of 96 and an input size of 224 × 224, utilizing a cosine learning rate scheduler (initial LR: $5e^{-4}$, minimum LR: $1e^{-6}$). The backbone was trained using patches generated by the object detection algorithms (N = 104,885). Figure 3 illustrates the efficacy of the pretrained model in our domain-specific dataset, in extracting relevant features from stool parasites, demonstrating that features belonging to the same class are clustered together while those from different classes are distinctly separated.

To assess the benefits of using SSL on a domain-specific dataset, we conducted an experiment comparing three different pre-training approaches for the ViT architecture: SSL on a domain-specific dataset (DINO-STH), SSL on a domain-agnostic dataset

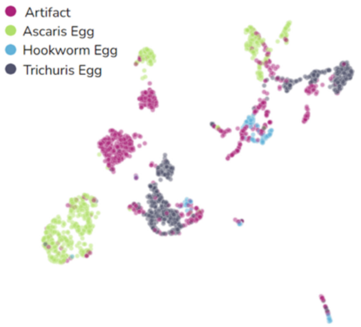

Fig. 3. Uniform manifold approximation and projection (UMAP) visualization of the features extracted from our pre-trained ViT architecture.

(DINO-ImageNet) as it demonstrated strong performance on various computer vision tasks, and no pre-training. In all three cases, we utilized the same ViT-S/16 architecture to ensure a fair comparison. This experiment was designed to determine whether our model, trained on domain-specific data, extracts more relevant features compared to a model trained on domain-agnostic data.

All experiments were conducted under the same hyperparameters: a batch size of 128, learning rate of 0.001, training for 100 epochs with early stopping patience of 10 epochs, and images resized to 224 × 224 pixels. During the supervised training phase, we executed two types of training strategies: fine-tuning, where both the backbone and the linear classifier (MLP) were trained together, and linear probing, where only the linear classifier was trained while keeping the backbone fixed. All these experiments used all available data of the training set (N = 10,416). Table 3 presents the performance of all models on the validation set.

Table 3. Parasite classification performance on the validation set, obtained from full fine-tuning and linear probing.

Pre-training	Fine tuning		Linear probing	
	BACC	F1-score	BACC	F1-score
None	57.48	55.13	47.90	46.61
DINO-ImageNet	55.45	53.48	85.10	85.15
DINO-STH	**91.48**	**91.70**	90.51	90.86

In addition, and for comparison purposes, we performed supervised training with only 200 images per parasite class, instead of using all available dataset, to assess the models' capacity when only a limited labeled dataset is available. When evaluated on the validation set, the model pre-trained on our domain-specific dataset (DINO-STH) achieved a BACC of 86.57% and a F1-Score of 86.24%, whereas when it was pre-trained on the domain-agnostic dataset (DINO-ImageNet) achieved 74.76% and 74.05% of BACC and F1-Score respectively.

3.3 Evaluation of the Proposed System on an Independent Test Set

To evaluate the generalization capabilities of our proposed system, we set aside a subset of labeled images, independent from those used for training and validation. After determining the best configurations from the validation set, we applied these configurations to the test set.

The evaluation revealed that the optimal configuration for parasite detection was YOLOv8-n, which achieved a mean Average Precision (mAP) of 94.93%, precision of 89.63%, and recall of 88.50% on the test set. For parasite species classification, the best performance was achieved using the ViT-S/16 architecture pre-trained on our domain-specific unlabeled dataset through SSL. All detected boxes (those with a probability greater than 0.05) were then processed by the classification algorithm, and the predictions were compared to the annotations made by experts. The performance of the parasite classification algorithm was 80.32% BACC) and 80.17% F1-score. The whole system achieved a mAP of 82.5%, precision of 89.63% and recall of 82.14%.

This assessment on an independent test set underscores the robustness and generalizability of our approach, demonstrating its potential for accurate parasite detection and classification in practical applications.

4 Conclusions

In this work, we presented a comprehensive approach for the detection and classification of soil-transmitted helminth (STH) parasites using YOLOv8 and Vision Transformers (ViT). For the classification task, we created a foundation model based on SSL techniques to leverage a large amount of unannotated data, enabling the model to learn meaningful features. The proposed classification system trained on our domain-specific data achieved better results compared to the ViT backbone pre-trained on domain-agnostic dataset (ImageNet). This improvement is particularly noticeable when the available annotated training set is small. With only 200 training images per parasite class, the performance improved by 12% when comparing our SSL-pretrained model to the one trained on a domain-agnostic dataset. This work is highly relevant because, in the field of medical imaging, the availability of labeled data is often limited. By incorporating SSL and leveraging unannotated data, we have shown that it is possible to enhance model performance, especially in data-scarce environments. This approach, which also leverages 3D-printing technologies and smartphones to enable data digitization without the need for expensive hardware, holds great promise for improving diagnostic accuracy and accessibility in low-resource settings, ultimately aiding in the fight against parasitic diseases. This marks a step toward meeting the World Health Organization's performance benchmarks for in-vitro diagnostic devices, leveraging AI to combat parasitic diseases.

References

1. Soil-transmitted helminth fact sheet, https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections, Accessed 25 May 2023

2. Pullan, R.L., Freeman, M.C., Gething, P.W., Brooker, S.J.: Geographical inequalities in use of improved drinking water supply and sanitation across -Saharan Africa: mapping and spatial analysis of cross-sectional survey data. PLoS Med. **11**, e1001626 (2014). https://doi.org/10.1371/journal.pmed.1001626
3. World Health Organization ed: Ending the neglect to attain the Sustainable Development Goals: a road map for neglected tropical diseases 2021–2030. World Health Organization (2021)
4. Katz, N., Chaves, A., Pellegrino, J.: A simple device for quantitative stool thick-smear technique in Schistosomiasis mansoni. Rev. Inst. Med. Trop. Sao Paulo **14**, 397–400 (1972)
5. Ward, P., et al.: Affordable artificial intelligence-based digital pathology for neglected tropical diseases: a proof-of-concept for the detection of soil-transmitted helminths and Schistosoma mansoni eggs in Kato-Katz stool thick smears. PLoS Negl. Trop. Dis. **16**, e0010500 (2022). https://doi.org/10.1371/journal.pntd.0010500
6. Meulah, B., et al.: A review on innovative optical devices for the diagnosis of human soil-transmitted helminthiasis and schistosomiasis: from research and development to commercialization. Parasitology **150**, 137–149 (2023). https://doi.org/10.1017/S0031182022001664
7. Yang, A., et al.: Kankanet: an artificial neural network-based object detection smartphone application and mobile microscope as a point-of-care diagnostic aid for soil-transmitted helminthiases. PLoS Negl. Trop. Dis. **13**, e0007577 (2019). https://doi.org/10.1371/journal.pntd.0007577
8. Li, Q., et al.: FecalNet: automated detection of visible components in human feces using deep learning. Med. Phys. **47**, 4212–4222 (2020). https://doi.org/10.1002/mp.14352
9. Huang, S.-C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S., Chaudhari, A.S.: Self-supervised learning for medical image classification: a systematic review and implementation guidelines. npj Digital Med. **6**, 74 (2023). https://doi.org/10.1038/s41746-023-00811-0
10. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8 (2023)
11. Dosovitskiy, A., et al.: An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. arXiv. (2020). https://doi.org/10.48550/arxiv.2010.11929
12. Caron, M., et al.: Emerging properties in self-supervised vision transformers. arXiv. (2021)