AI-Based Detection Methods for Fecal Parasite Diagnostics in Low-Resource Settings

Nathaly Jose-Maria

Georgia Institute of Technology

Global Solutions VIP

December 1, 2023

ADVISORS:

Kelsey Kubelick, PhD

Introduction

STH diagnostic research faces many constraints with the most crucial being few image datasets available. Current microscopic imaging capabilities are tedious and costly, and STH affects mostly impoverished countries. Thus, there are less images available for study. This is why researchers are aiming to develop highly accurate yet low-cost solutions incorporating machine learning (ML). ML minimizes both cost and manual involvement; however, replacing a set of trained pathologist eyes is not easy. Stool samples can carry various types of debris that share many characteristics with STH eggs, and STH eggs have highly specific characteristics across species which is what makes them so difficult to distinguish. As such, ML needs to be trained to know how to detect these eggs while skipping over any distractions. There are several approaches to doing so involving both dataset creation and architecture type, which will be discussed further on. For now, there are basic principles of ML we must know to understand what innovations are being made in artificial intelligence-based (AI) STH diagnostics.

Background Principles

Neural Networks [1]

Neural networks are models in machine learning focused on distinguishing distinctive features within data sets. Using an input dataset, the model trains itself on this data to know what features look like and how to differentiate them from others. As the data trains itself, it becomes more accurate with each iteration. The goal is to minimize the model's loss to ensure accuracy. Neural networks ultimately allowed for rapid and efficient classification and data clustering because of this.

Components

Nodes: An individual node is its own regression model composed of input data, weight, a threshold, and an output. If the output of a node surpasses its threshold, it is activated and subsequently informs the next layer of the network. Otherwise, the node does not pass any data forward. It is important that these parameters are properly initialized as they can lead to problems during training – particularly the problem of an exploding or vanishing gradient [8]. An exploding gradient occurs when weights are initialized too large and therefore causes oscillating around the minimum value without a conclusive answer [8]. A vanishing gradient occurs when weights are initialized too small and therefore leads to a loss convergence before its legitimate minimum value [8].

Layers: A layer is a collection of connected and independent nodes. A network hosts several layers with the essentials being an input layer, a hidden layer, and an output layer. The hidden layer is where calculations are made, and there can be several of these layers.

Activation Function: This function is the calculation that occurs within hidden layers. This determines the output of a node and therefore its activation toward a subsequent layer.

Direction Flow: Data can flow in two ways: feedforward, where data flows in one direction only from input to output, and backpropagation, where data moves from output to input [6]. Backpropagation is useful for calculating the error associated with each neuron and further optimize the model with parameter fitting [5,6].

Convolutional Neural Networks (CNNs)

Convolutional neural networks are a type of neural network likewise involving nodes, layers, and an activation function. This network is better suited for classification task as it takes advantage of an input's grid representation to denote features based on the properties of neighboring pixel clusters. Convolutional neural networks are comprised of three main layers: the convolutional layer, the pooling layer, and the fully-connected layer. A model may have several convolutional or pooling layers, but it has only one, final fully-connected layer. Earlier layers focus on simpler features while later layers recognize larger elements of an object because of the model's increasing complexity toward total object identification [2].

Components [3]

Convolutional Layer: This layer is where most computation occurs. It requires input data, a filter, and a feature map. The filter will work as a feature detector to identify if a specific feature is present in the input data. The filter moves across different areas of an input image. It can have various dimensions, such as 2x2 or 3x3, and the dot product of the filter against the input image is added to the feature map, or output. Thus, this output is a series of dot products. Weights are constant as a filter moves across an image; however, other properties, or hyperparameters, can be manipulated to affect the output [7]. These include the number of filters being run on an image, the stride, or distance, a filter moves across an input image, and padding, which changes the border dimensions of the output to accommodate the sizing interactions between the filter and the image [7]. Consistent with general neural networks, adding convolutional layers allows the model to start identifying simple features with increasing complexity as the model progresses.

Pooling Layer: This layer conducts dimensionality reduction to reduce the number of parameters in the input and minimize the complexity of the model. Like a filter, the layer moves across an image, but the layer applies a function to the image as opposed to utilizing consistent weights. A lot of information can be lost in this layer, but it significantly improves efficiency, reduces complexity, and limits the risk of overfitting, a common issue in machine learning where the output becomes too specific to model predictions such that it does not result as it should on new different data because of this bias [4].

Fully-Connected Layer: This layer ensures that the input image is fully connected to the output of the previous layer. This full connection performs the final classification task using the features extracted in previous layers.

Architectures: Many different architectures have been born from the creation of CNN, and many of them are similar aside from how much more efficient one is compared to another depending on how parameters are optimized. Some of these architectures are used through the studies mentioned in this paper, including the Inception series [22], U-Next Structure [24], the VGG series [23], the ResNet series [25], the MobileNet series [27], and Single-Shot Detection [30]. I would encourage further researching these concurrently with this paper to better understand how the upcoming experiments function.

Evaluation Methods

Validity [19]: To ensure the validity of a test, it is generally evaluated using a 2x2 table. The tables looks at the results of a test versus the true answer of a finding. When a test diagnoses

someone as diseased and the patient is truly diseased, this is a `True Positive (TP)`. When a test diagnoses someone as diseased and the patient is truly healthy, this is a `False Positive (FP)`. When a test diagnoses someone as healthy and the patient is truly diseased, this is a `False Negative (FN)`. When a test diagnoses someone as healthy and the patient is truly healthy, this is a `True Negative (TN)`.

Precision [20]: Precision determines what proportion of positive identifications are correct. The formula for calculating this is `TP / (TP + FP)`.

Recall [20]: Recall determines what proportion of actual positives were identified correctly. The formula for calculating this is `TP / (TP + FN)`.

F1 Score [21]: An F1 score is a measure of a model's accuracy on a dataset. This is the harmonic mean of precision and recall, or `(2 * Precision * Recall) / (Precision + Recall)`, which is equivalent to `TP / (TP + (FP + FN)/ 2)`.

Sensitivity [19]: Sensitivity is the ability of a test to correctly classify an individual as diseased. The formula for calculating this is `TP / (TP + FN)`.

Specificity [19]: Specificity is the ability of a test to correctly classify an individual as disease-free. The formula for calculating this is `TN / (TN + FP)`.

Positive Predictive Value (PPV) [19]: PPV tells us how many of test positive are true positive. The higher this number is, the better the test is at meeting the gold standard. The formula for calculating this is `TP / (TP+ FP)`.

Evidently, Precision and PPV are equivalent, and Recall and Sensitivity are equivalent [31].

Current Field Work

Regarding the context of the issue at hand, AI-based developments have leveraged various ML models to accurately classify and distinguish STH eggs in stool samples. Given the limited sample of imaging available, one of the largest considerations when developing an ML model is whether to build an independent learning system. This is a significant consideration as it determines how the model will identify image features – using its own raw data, or incorporating knowledge from other data. This question is why we will look into two trains of thought on AI-based detection models for STH diagnostics: the first is custom learning, or training a model using solely a custom image database; the second is transfer learning, or training a model using both a custom and online image database.

Custom Learning

For lack of a better term, we will call models using their own custom image databases as ones that utilize 'custom learning.' Despite many of these studies using pre-built architectures, these studies cannot be considered transfer learning as they do not incorporate learnings from other tasks. Please read ahead to the section of transfer learning for more information.

CL1) Kankanet: An artificial neural network-based object detection smartphone application and mobile microscope as a point-of-care diagnostic aid for soil-transmitted helminthiases [11].

Context: This study pilots the use of smartphone microscopy and a custom-built artificial neural network-based (ANN) object detection application named Kankanet to address these two needs.

Methods: Model: This study utilized the TensorFlow repository to create Kankanet, an ANN object detection system built upon a Single Shot Detection meta-architecture and a MobileNet feature extractor (a CNN developed for mobile vision applications). *Training:* The study trained two models – one trained solely with microscope images, and a second trained with both microscope and USB Video Class images. The models were tested using randomly selected images from the evaluation image set – images not included in the training set. The models worked to analyze images in real time by projecting a bounding-box over each detected object, displaying the name of the object detected, and a confidence rating. Ground truths were established by trained parasitologists. *Evaluation:* The study was evaluated on sensitivity, specificity, PPV, and NPV.

Results: The first model evaluating solely microscope images found generally high sensitivity and specificity for T. trichiura (100% & 91%) and low sensitivity and specificity for both A. lumbricoides (57.1% & 50%) and hookworm (0% & 80%). Moreover, T. trichiura had a PPV and NPV of 80% and 100% while A. lumbricoides had 44.4% and 62.5% respectively, and hookworm had 0% and 61.5% respectively. The second model evaluating both microscope images and UVC field images found higher sensitivity and specific for both A. lumbricoides (69.6% & 61.1%) and hookworm (71.4% & 100%), but lower sensitivity and specificity for T. trichiura (15.4% & 97.8%). T. trichiura had a PPV and NPV of 66.7% and 80% while A. lumbricoides had 92% and 23.9% respectively, and hookworm had 100% and 96.2% respectively. In practicality, just one egg per fecal sample slide needs to be detected to encourage patient treatment with medication. Despite having lower sensitivities and specificities relative to the golden Kato-Katz (KK) standard, the study's model still does this effectively. Moreover, the application of this method is cheaper per person relative to standard microscopy cost, and it just needs a more robust training set to provide better statistics. This is dependent on image capture quality. However, if field usage of this model is still unacceptable, it can also serve as a valuable training aid for identifying the presented helminths.

CL2) FecalNet: Automated detection of visible components in human feces using deep learning [14].

Context: This study sought to create a method to automate the detection and identification of feces components for early gastrointestinal disease diagnosis using multiple deep neural networks.

Methods: Model: ResNet152 is used as a basic network to extra and learn the characteristics of fecal components. This network works with the feature pyramid network (FPN) to enhance multiscale features and obtain a feature map. The FPN consists of a backpropagating layer, forward propagating layer, and a feature map of convolutional modules containing convolutional layers. The combined network then uses a full convolutional network (FCN) in a regressive and classification subnetwork. The classification subnetwork iteratively classifies and localizes fecal components and improves its loss function to optimize subsequent classification results. The network does so by utilizing an feature map input layer, four conversion layers each

containing 256 filters and activated by ReLu. The last layer then activates the sigmoid activation function as classification predictions are output. The regressive subnetwork mimics the classification subnetwork aside from its layer and output specifications. As opposed to other models, images are not segmented, and the model still boasts precise identification. *Training:* The model uses a custom image database consisting of six fecal components. *Evaluation:* The model is evaluated on its precision, recall, and calculation time.

Results: The model boasts an average precision of 92.16%, a recall of 93.56%, and a computation time of 1.02s. These metrics are superior compared to various other proposed conventional methods and deep learning methods (Zhang, Liu, Wang, Wang, Tchinda, Lin, Joseph, Li, Du). When looking at the species tested, erythrocytes had a precision and recall of 92.82% & 93.38% respectively, leukocytes 93.99% & 96.11%, intestinal mucosal epithelial cells 90.71% & 92.41%, hookworms 89.95% & 93.88%, ascarid 96.90% & 91.21%, and whipworms 88.961% and 94.37%.

CL3) Combining collective and artificial intelligence for global health diseases diagnosis using crowdsourced annotated medical images [15].

Context: This study proposes training deep learning algorithms using crowdsourced annotations from non-experts through playing a video game.

Methods: Model & Training: A CNN algorithm was used to solve the classification task of differentiating egg presence and species. The study leveraged the MobileNetV2 model by pretraining it on a larger dataset from ImageNet and adjusting the parameters to fit the context. Moreover, because crowdsourced annotations contain incorrect labels, the model utilized soft bootstrapping cross entropy loss function to minimize the value of incorrect labels during analysis. Data was augmented with transformations to generate more training data. *Evaluation:* The model was evaluated using accuracy and area under the receiver operating characteristic curve.

Results: Despite school-age children being less accurate with their labelling than adults, both groups were accurate enough with the former group averaging at an accuracy above 94%. Predictions across experts, adults, and children were generally comparable, demonstrating the effectiveness of crowdsourcing.

CL4) Point-of-care mobile digital microscopy and deep learning for the detection of soil-transmitted helminths and Schistosoma haematobium [17].

Context: This study sought to evaluate the imaging performance of a particular miniature digital microscopy scanner for diagnosing STHs. The study also sought to train a deep learning-based image analysis algorithm to automatically detect STHs from self-capture images.

Methods: Model: The study used a commercially available image analysis software platform WebMicroscope, which utilizes deep-learning based machine learning algorithms to create software models or computer vision applications. Its contextual usage in this study was for analyzing and classifying image features. The application utilizes annotated images to location candidate areas of interest, which in this case are STH eggs, and organized them according to species. The application works with two algorithms: the first analyzes the whole image for egg candidates by comparing image features with learned objects of interest; the second algorithm was passed in the information of the first for species classification and differentiation. *Training:* A set of images were labeled on an object level and annotated to indicate a center point objects of interest. The training set consisted of 205 randomly selected images of stool samples with single

fields of view capture with the mobile microscope and manually confirmed to contain visible eggs. *Evaluation:* The study was evaluated using positive predictive value (PPV).

Results: Of the 7385 images captured by the mobile microscope, 410 were manually confirmed to contain STH eggs containing 434 total eggs. The training set contained 217 eggs while the testing set contained 195 eggs. All testing set eggs were correctly detected by the model. A. lumbricoides and T. trichiura boasted relatively high PPVs of 93.7% and 100% respectively, but hookworm demonstrates a PPV of 69.6%

Transfer Learning

A common practice used in machine learning is the idea called transfer learning, or "utilizing knowledge acquired for one task to solve related ones" [18]. For example, say one task (T1) needs to identify objects in a restaurant setting, and another needs to identify objects in a café or park setting (T2). If more data exists for T1 than T2, then generalized information from T1, such as edges, shapes, corners, and intensity, can be shared such that T2 gets the feature distinctions it needs without rebuilding a new network. Moreover, T2 is able to start work on more unique features, like espresso machines or lamp posts, sooner and faster [18].

For this reason, we will look at several examples of transfer learning where many studies utilize similar base models that share simple features, such as TensorFlow [28], ResNet [25], and MobileNet [27], databases that contribute to the knowledge of feature formation, such as COCO [29] and ImageNet [26], and how these are then adapted to fit the need of each use case.

***TL1) Automatic classification of cells in microscopic fecal images using convolutional neural networks** [9].*

Context: This study seeks to identify the visible image of fecal composition based on deep learning, particularly region proposal and candidate recognition. The authors do this using a CNN architecture based on Inception-v3 and principle component analysis (PCA), and this method achieves higher-than-average precision of 90.7% relative to other mainstream CNN models.

Methods: Model: Images are fed into the Inception-v3 model, which results in a feature map that gets fed into PCA for dimension reduction to create a classification network for type recognition and a regression network for location correction. The last pooling player of the inception network was used as a feature extraction layer. *Training:* Training was conducted across two modules. Module 1 used the Inception with training by an ImageNet dataset. Module 2 passes different size inputs into the network trained by Module 1 to extract feature information for the feature-map layer. PCA then conducts dimension reduction on the feature information. *Evaluation:* The model is evaluated on its precision, recall, and F1 score.

Results: Average precision was 90.7%, recall was 92.5% , and the F1 score was 91.6%. These statistics were all superior compared to other CNN models (VGG-19, Inception-v3, Inception-v4, Inception-Resnet-v2). Moreover, the model boasted a low time consumption of 1200ms. The model detected red blood cells with a precision and recall of 92.9% & 95.7% respectively, white blood cells 88.8% & 88.5%, and molds 93.2% & 91.7% respectively.

***TL2) Mobile microscopy and telemedicine platform assisted by deep learning for the quantification of Trichuris trichiura infection** [10].*

Context: This study created a deep learning algorithm for automatic assessment and quantification of parasitological infections by STH.

Methods: *Model:* Single-Shot Detection architecture, MobileNet Network, and COCO image database. *Training:* Patches were extracted from images nearby ground-truth labels. Patch sizes were selected to be the approximate size of a ground truth. To augment the training batch size, sub-patches were selected from areas within originally selected patches. This was done so that the same object could appear in different locations within the image while still being the same object. From 51 KK slides that were obtained, 1507 images were produced with 797 containing eggs and 711 containing none. *Evaluation:* The model is evaluated on its precision, recall, and F1 score.

Results: For Trichuris spp. detection, the model boasted a precision of 98.4%, a recall of 80.9%, and an F1 score of 88.5%. Only 99 eggs were incorrectly detected among 20090 images containing none, which resulted in a specificity greater than 99% per image patch. When the algorithm was tested for its generalizability across multiple STH species, it found that Trichuris spp. had a new precision, recall, and F score of 95.31%, 89.71%, & 92.43% respectively while Ascaris spp.has 93.41%, 96.45%, & 94.91%. The mean of these two species was 94.36%, 93.08%, & 93.97% respectively.

***TL3) A low-cost, automated parasite diagnostic system via a portable, robotic microscope and deep learning** [13].*

Context: This study seeks to create a cost-effective, automated parasite diagnostic system that doesn't require special sample preparation or a trained user to develop. Images are recorded, automatically segmented, then analyzed using a trained CNN that distinguishes eggs from debris.

Methods: *Model:* The CNN uses a U-Net structure with 19 convolutional layers, a forward path, and a backpropagating path. The first max pooling layer is a deeper layer within the network to make learned feature more invariant to small translations. *Training:* The inception network was trained by an ImageNet dataset. Weights were initialized randomly pulling from a Gaussian, or Normal, Distribution. The data boasted 951 images, 643 of which were labelled and contained eggs of various parasitic species, and 308 of which contained no eggs. The model was trained with 873 images, 564 of which were labelled with eggs, and 308 of which contained no eggs. The model was tested with 79 images, all of which were labelled with eggs. *Evaluation:* The model is evaluated on its sample error, sensitivity, specificity, and accuracy.

Results: The model's sensitivity, specificity, and accuracy increased with analysis of more grids for the Eimeria species with single grid results boasting 98% sensitivity, 72% specificity, and 92% accuracy while four grid results boasted 100% sensitivity, 80% specificity, and 96% accuracy. All statistics were equivalent to 1 for nematode species.

***TL4) 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** [16].*

Context: This study sought to create an AI-based digital pathology device to explore automated scanning and detection of helminth eggs in stool prepared with the KK technique.

Methods: *Model:* R-FCN ResNet101 COCO model by TensorFlow 1 Detection Model Zoo with updated object classes and label mapping. *Training:* Field-of-view images taken were randomly shuffled and split into three for training, validation, and testing. A 70:20:10 ration was

applied as closely as possible for each present parasitic species. *Evaluation:* The model is evaluated on its precision, recall, and F1 score.

Results: The model boasted an average precision of 94.9%, a recall of 96.1%, and an F1-csore of 95.4%. Moreover, the model operates in 1.58s per image. When looking at the species tested, Ascaris had a precision, recall, and F1-score of 95.4%, 95.9%, & 95.6% while Trichuris had 94.2%, 96.4%, 95.3%, Hookworm had 95.0%, 97.7%, & 96.3%, and Schistosoma had 91.8%, 86.2%, & 88.9%.

Discussion

These diverse studies evidently had various methods of evaluating their performance. Moreover, a number of this studies were not directly related to STH egg detection but rather image object detection in fecal matter. Unfortunately, this makes it difficult to compare results across studies. However, we can still develop approximate relations and valuable outcomes regarding the most applicable techniques on STH egg detection. Moreover, we can utilize existing sensitivity and specificity statistics for KK manual examination as a baseline for those sharing similar context and metrics of evaluation regarding STH egg detection. Tarafder et al. showed that A. lumbricoides is examined at a sensitivity of 97% and a specificity of 96% respectively while hookworm is identified at 65% and 94%, and T. trichiuras at 91% and 94% [12]. Because these metrics are species-specific, we will have to examine statistical results both generally and across species.

Comparable STH Egg Detection Studies (CL1, CL2, CL4, TL3, TL4)

Despite utilizing different evaluation methods, the various STH egg detection-relevant studies can still be compared to one another given the similarity of their statistics. Please see the passage about Evaluation Methods in the Background Principles section for more information. As such, included are various tables consolidating the various evaluation methods and resulting statistics of these studies for easier comparison. It was decided that A lumbricoides, T. trichuria, and hookworm were the best species to use for comparing between models as they are the most present and therefore better comparable in the various analyzed studies on soil-transmitted helminths.

Summary of Studies and their Architectures, Eval Methods, Species, & Avg Statistics

	Standard	Study				
	KK	CL1	CL2	CL4	TL2	TL4
Architecture	Manual	MobileNet	ResNet152	WebMicroscope	Single-Shot Detection, MobileNet, COCO	R-FCN ResNet101, COCO, TensorFlow 1
Evaluations	Sensitivity, Specificity	Sensitivity, Specificity, PPV	Precision, Recall	PPV	Precision, Recall, F1	Precision, Recall, F1
Relevant Species	A. lumbricoides, T. trichiura, Hookworm	A. lumbricoides, T. trichiura, Hookworm	A. lumbricoides, Hookworm	A. lumbricoides, T. trichiura, Hookworm	A. lumbricoides, T. trichiura	A. lumbricoides, T. trichiura, Hookworm
Avg Sensitivity (Recall) (%)	84.33	52.3667 | 80.2	92.545		88.675	96.667
Avg Specificity (%)	94.67	73.667 | 87.033				
Avg Precision (PPV) (%)		41.4667 | 86.233	93.43	87.7667	95.91	94.87
Avg F1-Score (%)		69.422 | 62.528	92.921		91.71	95.733

Table 1. Summary of Studies & their Architectures, Eval Methods, Species, & Avg Statistics.
This summary table was made to provide a better single-view comparison between the various STH-based studies and their key information – particularly architecture, evaluation methods, and average statistics. Not all studies shared the same evaluation methods; however, given the comparable nature of these statistics, sensitivity and recall statistics were paired as an identical identifier while precision and PPV were likewise paired. The average statistics were personally computed by averaging the statistics of a particular study across the three presented species irrelevant of the other species studied in a paper's individual research. F1 score cells highlighted in yellow were personally computed with the formula specified in the Background Principles section []. Moreover, there are two statistics for sensitivity and specificity for study CL1 as it represents both Model 1 and Model 2. Model 1 was more successful for T. trichiura while Model 2 was more successful for both A. lumbricoides and hookworm.

General Relationships: When looking at the summation table, we can see some clear attributes both shared among studies and distinguishable from the others. First, it is evident that most statistics are generally around the same range from one another. This indicates that these algorithms all perform close to the target range of the KK golden standard, which is ideal as it establishes a general baseline of performance among these algorithms. This makes it such that developers only need to focus on finding the write parameters for optimizing their algorithms such that they can reach greater efficiencies and accuracies.

Custom vs Transfer Learning: When continuing to look at more general relationship, it is surprising to see that the custom learning models (CL1, CL2, & CL4) were generally less accurate compared to those that utilized transfer learning (TL2 & TL4) when looking at F1 scores as a representative of accuracy but also comparing the deviations between the models' sensitivities, and precisions. This is interesting because custom learning models create particular datasets and training models specific to the information being investigated; however, it seems that transfer learning's ability to share more common, less detailed features helped the other models in getting not only precise egg distinctions but consistent ones at that. Contextually, despite CL2 operating on a custom learning model, it works with many more parasite species relative to CL1 and CL4, yet it seems to perform better than the two despite the circumstances. Once again, I wonder if the presence of more diverse species better trains the algorithm to better detect the distinction between the species.

Moreover, there seems to be a stronger preference for sensitivity, specificity, and PPV evaluation in custom learning cases as opposed to transfer learning cases that seem to prefer precision, recall, and F1-score evaluation. This is likely because of the nature of transfer learning it that it needs to be compatible across projects. As such, to test its efficiency across these several projects, it needs a consistent, comparable metric. Sensitivity and specificity seem to be more specific to biological statistical cases, hence why the custom learning methods seem to utilize it instead of the more computer-science based transfer learning technique.

CL1: Both models of CL1 evidently underperform compared to the KK technique; however, as the study mentioned, given its closeness to the target goal, this model can still be very valuable in STH egg detection as it may not completely replace manual detection but at least ease the process by indicating the presence of eggs as opposed to an accurate count. Many times just the presence if needed to provide medication, so CL1, among the other noted algorithms, are capable of doing just that.

Continuing to look at CL1, the reason for it having a low average PPV is likely because both models work entirely different when applied to the three tested species, therefore causing a disparity when looking at averaged statistics. Model 1 works much better for identifying T. trichiura at the cost of misidentifying the other two species while Model 2 works much better for identifying A. lumbricoides and hookworm at the cost of misidentifying the last remaining species. This issue arises because of the nature of STH eggs in being hard to identify among debris but also containing very specific features that can be hard to isolate through machine learning.

The contrasting behavior of the models raises an interesting idea of taking advantage of multiple models for several species as it may be impossible to get perfect results on all species with a single generalized model. While that would be ideal for efficiency and ease, it may not be the most practical approach to start with. While other algorithms seem to be able to maintain high averages despite the differences between the species, this scenario is a good example to posit alternative thinking.

Studies & their A. lumbricoides Statistics

	Standard	Study				
	KK	CL1	CL2	CL4	TL2	TL4
Sensitivity (Recall) (%)	97	57.1 | 69.6	91.21		96.45	95.9
Specificity (%)	96	50 | 61.1				
Precision (PPV) (%)		44.4 | 92	96.9	93.7	93.41	95.4
F1-Score (%)		49.955 | 79.248	93.969		94.91	95.6

Table 2. Studies & their A. lumbricoides Statistics.
This table was made to provide a better single-view comparison between the various STH-based studies and their results regarding A. lumbricoides.

A. lumbricoides: Moving onto investigating the statistics of individual species, it appears analysis on A. lumbricoides is quite promising across all avenues. The species has generally high sensitivity and specificity, especially with the seen transfer learning models. Once again, it is interested to see lower statistics with the custom learning models given their advantage in personalizing the training to the data's needs.

Studies & their T. trichiura Statistics

	Standard	Study			
	KK	CL1	CL4	TL2	TL4
Sensitivity (Recall) (%)	91	100 | 15.4		80.9	96.4
Specificity (%)	94	91 | 97.8			
Precision (PPV) (%)		80 | 66.7	100	98.4	94.2
F1-Score (%)		88.889 | 25.023		88.5	95.3

Table 3. Studies & their T. trichiura Statistics.
This table was made to provide a better single-view comparison between the various STH_based studies and their results regarding T. trichiura.

T. trichiura: Unlike with A. lumbricoides, it appears there is more inconsistency across the models when identifying T. trichiura. The context of CL1's contrasting models explains its wide difference, but it is interesting to see a higher deviation in results, particularly with TL2 having a much lower recall than expected yet such high precision.

Studies & their Hookworm Statistics

	Standard	Study			
	KK	CL1	CL2	CL4	TL4
Sensitivity (Recall) (%)	65	0 | 71.4	93.88		97.7
Specificity (%)	94	80 | 100			
Precision (PPV) (%)		0 | 100	89.95	69.6	95
F1-Score (%)		#DIV/0! | 83.314	91.873		96.3

Table 4. Studies & their Hookworm Statistics.
This table was made to provie a better single-view comparison between the various STH-based studies and their results regarding hookworm.

Hookworm: Finally, the results on hookworm are generally unsurprising. Across my research, hookworm was found to be one of the most inconsistent STHs to image and correctly identify. If it the species is difficult to classify manually, evident with the 65% sensitivity rate, then it is reasonable that creating the rules for how to classify a hookworm egg comes with its challenging. Regardless, I am surprised at the encouraging results of TL4 given its high precision, recall, and F1-score. The distinguishing factor for this algorithm could be a multitude of things. However, given that it is the only transfer learning model of those to analyze hookworm, that once again raises my suspicions in that utilizing transfer learning is better teaching the model despite the ambiguous criteria for identification. I would want to find further research about utilizing transfer learning for hookworm identification to truly confirm my suspicion.

Related Fecal Object Detection Studies (CL3, TL1, TL3)

CL3: This is one experiment that looks to improve STH egg detection outside of traditional methods of algorithms for field work. It is a novel, unique take on the issue, and one that seems very promising. Despite receiving image annotations from non-experts that would be seemingly useless and incorrect, this study has proven otherwise how crowdsourcing and taking advantage of a greater focus group aside from trained pathologists and parasitologists can help rapidly improve STH egg detection by still taking advantage of manual review. The gamification made out of egg identification makes this an appealing alternative as it completely flips the script on the traditional ideas of egg identification being tedious and unpleasant because of exposure to noxious smells and long work hours. While the implementation is still not perfect, the model is a seemingly effective method of gathering additional manual verification on any images passed through the initial model.

TL1: This experiment likewise tackles the problem of egg identification through an alternate route in focusing on model optimization rather than image improvement or object detection. As opposed to the other architectures utilized in the comparable STH egg identification studies, this model uses a joint PCA-Inception network. This is a network that takes advantage of two models in one, and it was verified to have superior functionality compared to other single-unit architectures like the variations of the Inception series and VGG. Despite not testing directly with STH eggs, the model was still valuable in classifying objects in fecal matter at a relatively high rate with precision and recall rates for red blood cells, white blood cells, and molds all resulting above 88%.

TL3: This experiment is quite like the comparable STH egg detection studies, except in that the study focuses on identifying animal STH infections, including ascarid Toxocara spp. in dogs, strongyles in sheep, Trichuris spp. in monkeys, and Coccidian parasites in dogs, sheep, and cows. The species are successfully distinguished from each other with basic parameters of object area and minor axis length. Given that treatment for nematodes and Coccidian parasites is different from each other yet similar within each group, the study mainly focused on creating this distinction as opposed to inner-group species distinctions. As such, the study boasts a high sensitivity, specificity, and accuracy or Eimeria at the single grid level that only improves at the 4 grid level. Moreover, the study has a sensitivity, specificity, and accuracy of 100 across both grid sizes for nematodes that make them clearly distinguishable.

Conclusion

This expansive investigation into modern AI-based techniques for fecal parasite diagnostics has concluded in surprising evidence towards the superiority of transfer learning. All working algorithms on parasite diagnostics are generally similar in performance aside from specific architectures or models; however, it seems that transfer learning results in not only more accurate egg identification but also more consistent egg species distinction. In retrospect, this conclusion is unsurprising as the use of more diverse imaging toward complex feature extraction is monumentally valuable in saving resources like time and computer power; however, I did not expect to find that models made custom to work as detailed at STH egg identification cannot compete nearly as close with other models taking advantage of network knowledge.

As such, I would personally recommend for research in this field to continue utilizing transfer learning for more optimal results. However, I would not completely disregard novel ideas or considerations. The use of egg detection solely for treatment distribution as opposed to accurate counting can be monumental in providing medication without too much latency between physician notification and treatment distribution. Moreover, the use of multiple models

towards the identification of several species instead of reliance on a single, generalized model can be valuable if areas see some species of STH more prevalently than others. Considering utilization of public knowledge can also be extremely potent in teaching models faster and with more accuracy without overly taxing pathologists and parasitologists better trained for the role. Next, considering different types of machine learning architectures with trained experts can be extremely valuable in navigating egg identification in a way biologists untrained in machine learning cannot yet understanding. Lastly, looking to animal patient studies can also bear fruitful techniques for application in human medicine. There are many ways this research can continue to go, and it is exciting to see all the possibilities come to fruition. However, much more work needs to be done in this field. This includes hardware improvements such that images are being taken at the highest quality possible in low-resource fields such that the data analysis through machine learning can work much quicker and better with less manual cooperation. Overall, research on STH egg identification is becoming quite promising and will hopefully rid low-resource areas of infections at a rate satisfactory to WHO expectations soon.

References

[1] IBM (n.d.). *What are neural networks?* IBM. https://www.ibm.com/topics/neural-networks.

[2] IBM (n.d.). *What are convolutional neural networks?* IBM. https://www.ibm.com/topics/convolutional-neural-networks.

[3] Mishra, M. (2020, Aug. 26). *Convolutional Neural Networks, Explained*. Medium. https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939.

[4] Amazon Web Services. (n.d.). What is Overfitting? AWS. https://aws.amazon.com/what-is/overfitting/#:~:text=Overfitting%20is%20an%20undesirable%20machine,on%20a%20known%20data%20set.

[5] HMKCode (2019, Nov. 03). *Backpropagation Step by Step.* hmkcode.com. https://hmkcode.com/ai/backpropagation-step-by-step/.

[6] Jefkine (2016, Sept. 05). Backpropagation in Convolutional Neural Networks. jefkinee.com. https://www.jefkine.com/general/2016/09/05/backpropagation-in-convolutional-neural-networks/

[7] Ramesh, S. (2018, May 07). *A guide to an efficient way to build neural network architectures – Part II: Hyper-parameter selection and tuning for Convolutional Neural Networks using Hyperas on Fashion-MNIST.* Medium. https://towardsdatascience.com/a-guide-to-an-efficient-way-to-build-neural-network-architectures-part-ii-hyper-parameter-42efca01e5d7.

[8] Katanforoosh & Kunin. (2019). *Initializing neural networks*. deeplearning.ai. https://www.deeplearning.ai/ai-notes/initialization/index.html.

[9] Du, K., Liu, L., Wang, X., Ni, G., Zhang, J., Hao, R., Liu, J., & Liu, Y. (2019, Apr. 05). Automatic classification of cells in microscopic fecal images using convolutional neural networks. Biosci Rep 39(4). https://doi.org/10.1042/BSR20182100.

[10] Dacal, E., Bermejo-Pelaez, D., Lin, L., Alamo, E., Cuadrado, D., Martinez, A., Mousa, A., Postigo, M., Soto, A., Sukosd, E., Vladimirov, A., Mwandawiro, C., Gichuki, P., Aba Williams, N., Munoz, J., Kepha, S., & Luengo-Oroz, M. (2021, Sept. 07). Mobile microscopy and telemedicine platform assisted by deep learning for the quantification of Trichuris trichiura infection. *PLOS Neglected Tropical Diseases*. https://doi.org/10.1371/journal.pntd.0009677.

[11] Yang, A., Bakhtari, N., Langdon-Embry, L., Redwood, E., Grandjean Lapierre, S., Rakotomanga, P., Rafalimanantsoa, A., De Dios Santos, J., Vigan-Womas, I., Knoblauch, A., & Marcos, L.A. (2019, Aug. 05). 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 Neglected Tropical Diseases*. https://doi.org/10.1371/journal.pntd.0007577.

[12] Tarafder, M.R., Carabin, H., Joseph, L., Balolong Jr., E., Olveda, R., & McGarvey, S.T. (2010, Mar. 10). Estimating the sensitivity and specificity of Kato-Katz stool examination technique for detection of hookworms, Ascaris lumbricoides, and Trichuris trichiura infections in humans in the absence of a 'gold standard.' International Journal for Parasitology, 40(4): 399-404. https://doi.org/10.1016/j.ijpara.2009.09.003.

[13] Li, Y., Zheng, R., Chu, K., Xu, Q., Sun, M., & Smith, Z.J. (2019, May 13). A low-cost, automated parasite diagnostic system via a portable, robotic microscope and deep learning. *Journal of Biophotonics*, 12(9). https://doi.org/10.1002/jbio.201800410.

[14] Li, Q., Li, S., Liu, X., He, Z., Wang, T., Xu, Y., Guan, H., Chen, R., Qi, S., & Wang, F. (2020, June 24). FecalNet: Automated detection of visible components in human feces

using deep learning. *Medical Physics*, 47(9):4212-4222. https://doi.org/10.1002/mp.14352.

[15] Lin, L., Bermejo-Perez, D., Capellan-Martin, D., Cuadrado, D., Rodriguez, C., Garcia, L., Tome, R., Postigo, M., Jesus Ledesma-Carbayo, M., & Luengo-Oroz, M. (2021). "Combining collective and artificial intelligence for global health diseases diagnosis using crowdsourced annotated medical images," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 3344-3348. https://doi.org/10.1109/EMBC46164.2021.9630868.

[16] Ward, P., Dahlberg, P., Lagatie, O., Larsson, J., Tynong, A., Vlaminck, J., Zumpe, M., Ame, S., Ayana, M., Khieu, V., Mehonnen, Z., Odiere, M., Yogannas, T., Van Hoecke, S., Levecke, B., & Stuyver, J. (2022). Affordable artificial intelligence-based digital pathology for neglected tropical diseases: A proof-of-concept for the detectin of soil-transmitted helminths and Schistosoma mansoni eggs in Kato-Katz stool thick smears. *PLOS Neglected Tropical Diseases*. https://doi.org/10.1371/journal.pntd.0010500.

[17] Halmstrom, O., Linder, N., Ngsala, B., Martensson, A., Linder, E., Lundin, M., Moilanen, H., Suutala, A., Diwan, V., & Lundin, J. (2017, Aug. 15). Point-of-care mobile digital microscopy and deep learning for the detection of soil-transmitted helminths and Schistosoma haematobium. *Global Health Action*, 10(3). https://doi.org/10.1080/16549716.2017.1337325.

[18] Sarkar, D. (2018, Nov. 14). *A Comprehensive Hands-on Guide to Transfer Learning with Real-World Application in Deep Learning*. Medium. https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a.

[19] Parikh, R., Mathai, A., Parikh, S., Sekhar, G.C., & Thomas, R. (2008). Understanding and using sensitivity, specificity, and predictive values. Indian J Opthalmol, 56(1): 45-50. https://doi.org/10.4103/0301-4738.37595.

[20] Google. *Classification: Precision and Recall*. Google Machine Learning Education. https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

[21] Kundu, R. (2022, Dec. 16). *F1 Score in Machine Learning: Intro & Calculation.* V7labs.com. https://www.v7labs.com/blog/f1-score-guide.

[22] Raj, B. (2018, May 29). *A Simple Guide to the Versions of the Inception Network.* Medium. https://towardsdatascience.com/a-simple-guide-to-the-versions-of-the-inception-network-7fc52b863202

[23] Boesch, G. (n.d.). *VGG Very Deep Convolutional Networks (VGGNet) – What you need to know.* Viso.ai. https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/

[24] University of Freiburg. (n.d). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Uni Freiburg. https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/

[25] He, K., Zhang, X., Ren, S., & Sun, J. (n.d.). *Residual Network*. Papers With Code. https://paperswithcode.com/method/resnet

[26] Stanford Vision Lab. (2020). *About ImageNet*. ImageNet. https://www.image-net.org/about.php

[27] PA, S. (2020, Jun. 10). *An Overview on MobileNet: An Efficient Mobile Vision CNN.* Medium. https://medium.com/@godeep48/an-overview-on-mobilenet-an-efficient-mobile-vision-cnn-f301141db94d

[28] TensorFlow. (n.d.). *Why TensorFlow.* TensorFlow. https://www.tensorflow.org/about
[29] Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., & Dollar P. (2014, May 1). COCO (Microsoft Common Objects in Context). *Microsoft COCO: Common Objects in Context.* Papers With Code. https://paperswithcode.com/dataset/coco
[30] ArcGIS. (n.d.). *How single-shot detector (SSD) works?* ArcGIS Developers. https://developers.arcgis.com/python/guide/how-ssd-works/
[31] Lekhtman, A. (2019, Aug. 5). *Data Science in Medicine – Precision & Recall or Specificity & Sensitivity?* Medium. https://towardsdatascience.com/should-i-look-at-precision-recall-or-specificity-sensitivity-3946158aace1#:~:text=Precision%20is%20also%20called%20PPV%20(Positive%20Predictive%20Value).