2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Combining collective and artificial intelligence for global health diseases diagnosis using crowdsourced annotated medical images

Lin Lin[†,1,2], David Bermejo-Peláez[†,1], Daniel Capellán-Martín[1], Daniel Cuadrado[1]
Cristina Rodríguez[1], Lydia García[1], Nuria Díez[1], Rocío Tomé[1]
María Postigo[1], María Jesús Ledesma-Carbayo[2], Miguel Luengo-Oroz[1]

Abstract— **Visual inspection of microscopic samples is still the gold standard diagnostic methodology for many global health diseases. Soil-transmitted helminth infection affects 1.5 billion people worldwide, and is the most prevalent disease among the Neglected Tropical Diseases. It is diagnosed by manual examination of stool samples by microscopy, which is a time-consuming task and requires trained personnel and high specialization. Artificial intelligence could automate this task making the diagnosis more accessible. Still, it needs a large amount of annotated training data coming from experts.**

In this work, we proposed the use of crowdsourced annotated medical images to train AI models (neural networks) for the detection of soil-transmitted helminthiasis in microscopy images from stool samples leveraging non-expert knowledge collected through playing a video game. We collected annotations made by both school-age children and adults, and we showed that, although the quality of crowdsourced annotations made by school-age children are sightly inferior than the ones made by adults, AI models trained on these crowdsourced annotations perform similarly (AUC of 0.928 and 0.939 respectively), and reach similar performance to the AI model trained on expert annotations (AUC of 0.932). We also showed the impact of the training sample size and continuous training on the performance of the AI models.

In conclusion, the workflow proposed in this work combined collective and artificial intelligence for detecting soil-transmitted helminthiasis. Embedded within a digital health platform can be applied to any other medical image analysis task and contribute to reduce the burden of disease.

I. INTRODUCTION

Achieving Universal Health Coverage by 2030 is one of the Sustainable Development Goals and World Health Organization priorities (WHO) [1]. Half the world's population lacks access to essential health services and diagnosis is a key step to achieve universal healthcare. Many of those diseases are diagnosed by visual inspection, which requires experts in front of the microscope and other medical devices at a certain time, a resource that is not always available. Artificial intelligence (AI) presents an opportunity to support these diagnostic processes. The number of AI-based medical devices for diagnosis is increasing. From 2015 to 2020, 222 AI devices were approved in USA and 240 in Europe [2]. Most of them were developed for radiology and cardiovascular diseases, and none of them is for microscopy applied to microbiology. Recently a few algorithms were trained to detect Malaria's parasites [3], [4] and helminth's eggs in fecal samples [5]–[7], showing the potential of AI algorithms for microscopic images. Notwithstanding the above, more studies are needed to create algorithms approved by regulatory institutions.

Soil-transmitted helminthiasis (STH) is a neglected tropical disease (NTD) that affects the poorest and most deprived communities. According to WHO's report, there are 1.5 billion people affected by Helminths worldwide. WHO established a roadmap to eliminate STH to reduce the global health burden. To accelerate the elimination, innovation and new technologies like AI are needed [8]. The recommended diagnosis method for STH is Kato Katz, which is a laboratory method for preparing stool samples for the later detection and quantification of STH eggs under a microscope [9].

Image annotation to train AI models is a time-consuming labour that poses an important burden into experts. However, in recent years, the use of crowdsourcing has been proposed to overcome this problem by delegating this task on a large group of untrained annotators. Several studies have already demonstrated the validity of the use of crowdsourcing for annotating medical images. In 2012, Luengo-Oroz et. al demonstrated that the combination of the annotations collected using a video game of 22 players achieved a malaria parasite counting accuracy higher than 99% in thick malaria smears [10]. In 2019, Linares et. al demonstrated that combined annotations form 25 players were able to distinguish most Malaria species with an accuracy of 99% [11]. Furthermore, Keshavan et. al combined crowdsourcing and Deep Learning (DL) to predict the quality of Magnetic Resonance Imaging [12].

Within this context, this work proposes a methodology to train DL algorithms for quantifying parasitic infection in microscopy images with the following objectives: 1) to assess the feasibility of training DL algorithms for the differentiation of helminths eggs based on microscopy images with annotations obtained from crowdsourcing using a custom video game, 2) to identify the relationship between the amount of training data and the deep learning model performance using incremental training and 3) to compare the performance of AI models trained with data annotated by both untrained school-age children and adults (general population).

[1] Spotlab SL, Madrid, Spain

[2] Biomedical Image Technologies, ETSI Telecomunicación, Universidad Politécnica de Madrid & CIBER-BBN, Madrid, Spain

[†] LL and DBP contributed equally to this work and are considered co-first authors.

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

II. METHODOLOGY

A. Crowdsourcing image annotation

We developed SpotWarriors (SW), a publicly available[1] set of mini-games that contribute to the diagnosis of diseases while playing, by generating crowdsourced annotated medical images. For the purpose of this project, we focused on a mini-game for the classification of small image patches from digitized stool samples for the identification of different helminths eggs, including *Ascaris* spp., *Trichuris* spp., Hookworms, and images without eggs. Figure 1 shows a screenshot of the game used.

All data used in this study for training, validation and testing of the AI algorithm came from 41 digitized stool samples from 6 different infected patients who were part of a follow-up study. Digitization of samples were made at 10x magnification. Ethical approval was obtained from the Kenya Medical Research Institute (KEMRI) Ethics Review Committee (SERU 3873).

From all digitized samples, we generated a total of 10319 cropped image patches (256x256 pixels) without overlap. Our interpretation is that all image patches can be considered as independent although they come from a limited number of subjects. We introduced 700 randomly selected image patches in the video game which were annotated by at least 20 adults and 20 school-age children (from 11 to 18 years old) players. These annotated images were used for training DL algorithms. For comparative purposes, and to assess the quality of crowdsourced annotations, these training images were also analyzed by experts microscopists.

Annotations from school-age children were obtained by organizing workshops in different schools. The workshops, presented in collaboration with the teachers, included an explanation of the project and concepts related to global health, artificial intelligence and collective intelligence in addition to playing the game. Data from adults was collected anonymously from online players.

Additionally, the remaining images were annotated by experts and were used as validation and test sets (2932 and 6678 images respectively, randomly separated). For crowdsourced annotated images, the ground truth (GT) was generated using the majority voting rule, where the most common response among players was chosen. The final distributions for the training set (those images introduced in the game), validation set, and test set are presented in Table I.

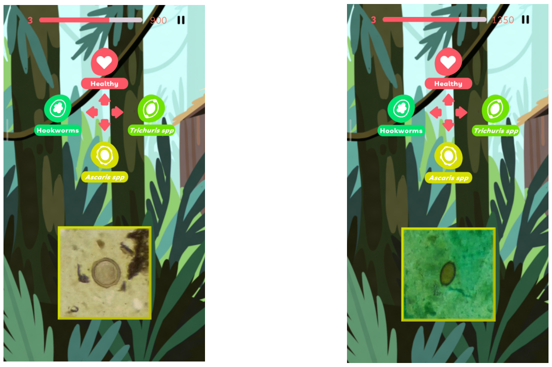

Fig. 1: Screenshot of the mini-game used to collect data with *Ascaris* spp. (left) and *Trichuris* spp. (right)

	Train			**Validation**	**Test**
	E	SAC	A	E	E
Ascaris spp.	179	182	176	1258	2819
Trichuris spp.	212	198	211	241	570
Healthy	309	320	313	1433	3298
Total	700	700	700	2932	6687

TABLE I: Distribution of the training, validation and test sets. Training images are annotated by three groups: experts (E), school-age children (SAC) and adults (A).

B. AI architecture: Deep learning model

In the present work, we used a Convolutional Neural Network (CNN)-based algorithm to solve the classification task for differentiating helminths eggs along with non infected sample images. The algorithm, given an image, returns an output probability distribution along the different classes under study, and the final predicted label is then computed as the one that has the highest probability. Particularly, we used the MobileNet V2 model [13], a light-weighted architecture designed to run on mobile phones in an efficient manner. This particular architecture has three main components including depthwise convolutions that significantly reduce the number of parameters, inverted residual connection blocks which modify residual blocks for efficiency purposes, and linear bottleneck layers without any non-linear activation function in order to preserve information in the low dimensional space. MobileNet V2 architecture is composed by 157 layers and involves only 3.5 million of parameters, compared to the 138,4 million of parameters that are involved in the well known VGG-16 architecture along its 23 layers [14].

To overcome the limitation of having a small training dataset, we used a transfer learning technique by pretraining the MobileNet V2 model on a large dataset (ImageNet [15]) and fine-tuning it in our dataset for the classification of helminths eggs. Fine-tuning was performed by freezing the earlier layers which learn generic features, and retraining later layers, responsible for extracting specific features of the problem under study. Using this technique we can reduce the computational cost and result in better performance than training from scratch, specially when little training data is available.

Because crowd-sourced annotations usually contain some incorrect labels, we used soft bootstrapping cross entropy loss function, which minimize the damage of incorrect labels by dynamically updating the targets of the prediction based on the actual state of the model [16]. The loss function is defined as equation 1, where $\mathbf{q}$ is the prediction, $\mathbf{t}$ is the target, β is the scaling factor between predictions and targets

[1] `https://spotwarriors.org/en/`, also available at Google Play Store and Apple App Store.

and L is the number of classes under study.

$$\mathcal{L}_{\text{soft}}\left(\mathbf{q},\mathbf{t}\right) = \sum_{k=1}^{L}\left[\beta t_k + (1-\beta) q_k\right]\log\left(q_k\right) \tag{1}$$

Furthermore, we used data augmentation including rotation, shift, flip, zoom, and shear transformations to generate more training data to further improve the model performance. Additionally, we also used early stopping technique during the training process with a patience of 10 iterations to avoid overfitting on the training set.

DL models were designed and trained using Keras with Tensorflow, and using a GPU NVIDIA Tesla T4 16GB.

III. Experiments and Results

In this section, we first evaluated the quality of crowdsourced annotations using a majority voting mechanism. We continued with the study of the effect of the training sample size on the performance of the model. And finally we compared the DL model trained with school-age children and adults annotations.

To evaluate the quality of crowdsourced annotations, we used the accuracy metric, which is defined as $ACC = TP + TN/N$ where TP, TN and N stand for true positives, true negatives and total number of samples respectively.

We used the area under the receiver operating characteristic curve (AUC) for evaluating the performance of DL algorithms. AUC measures the performance of the model across all possible probability thresholds. Macro-average AUC along classes is computed in order to avoid bias due to imbalanced class distribution.

In order to obtain a robust metric not affected by training instability, we repeated the training process 5 times, and calculated the mean and standard deviation of the performance metrics. The training of the models was carried out using the training set while the validation set was used for hyperparameter tuning. The incremental training experiment and the final performance evaluation were assessed on the test set.

It should be noted that we did not include Hookworm class in the analysis due to the lack of representativity in our database. No preprocessing was made on the images.

A. Quality of crowdsourced annotations

In order to select the optimal size of the quorum that best performs in comparison with the expert annotations, we used a bootstrap sampling method. For each image, we generated the final annotation using the majority voting rule considering only N randomly selected annotations. To measure the stability of each quorum size (N) we repeated this process 10 times. Table II shows the difference on the quality of crowdsourced annotations using different quorum sizes, by comparing the generated annotations by players and annotations made by experts. As derived from the table, we can observe that annotations based on 20 different player responses obtained the best performance.

Even though adult annotations obtained better accuracy with respect to expert annotations, annotations from 20 school-age children were found to be of enough quality (accuracy $> 94\%$).

Quorum size (N)	School-age children	Adults
5	0.842 (0.007)	0.966 (0.004)
10	0.910 (0.006)	0.988 (0.002)
15	0.931 (0.005)	0.989 (0.003)
20	0.946 (0.004)	0.991 (0.002)

TABLE II: Mean accuracy and the standard deviation of crowdsourced annotations using different quorum sizes compared to the ones made by experts.

B. DL model: hyperparameter tuning

Mobilenet V2 is built on different blocks. To determine the optimal number of layers to be fine-tuned during transfer process, we froze a determined number of blocks and fine-tuned the rest. For this experiment, models were trained with experts annotations.

Furthermore, with the aim of evaluating the effectiveness of soft bootstrapping loss for noisy labels, and to select the best loss function for this particular case study, we trained the DL model using both conventional cross entropy and soft bootstrapping cross entropy ($\beta = 0.95$) loss functions. Models were trained with school-age children and adults annotations (expert annotations do not contain noisy labels). As derived from Table III, the best performance was obtained when the first 46 layers were not fine-tuned and when bootstraping categorical cross entropy was used as the loss function.

Hyperparameter		AUC
Number of frozen layers		
	19 (block 2)	0.873 (0.033)
	46 (block 5)	**0.919 (0.019)**
	73 (block 8)	0.914 (0.014)
	99 (block 11)	0.917 (0.011)
Loss function		
Children	Cross entropy	0.927 (0.007)
	Bootstrapping cross entropy	**0.932 (0.006)**
Adults	Cross entropy	0.911 (0.014)
	Bootstrapping cross entropy	**0.925 (0.016)**

TABLE III: Hyperparameter selection. Mean AUC and standard deviation are shown.

C. Incremental training

To study the effect of training sample size on the models performance we trained the model incrementally using different sample sizes. We performed this incremental training by steps of 100 training images from 100 to 700.

Figure 2 shows the results of the incremental training experiment, revealing the importance of the number of samples used for training and its impact on the model performance.

D. Differences between annotator groups

We compared the performance of the AI model trained with school-age children, adults and expert annotations. Table IV summarizes the results of the three models using all available training samples (N=700). The results show that all

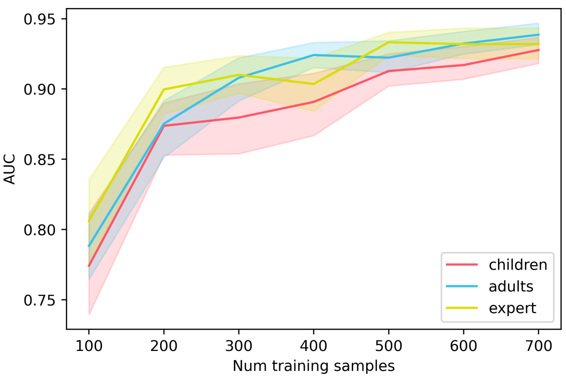

Fig. 2: Evolution of the model performance (AUC) as the training sample size increases. Results from data annotated by children, adults and experts are presented independently.

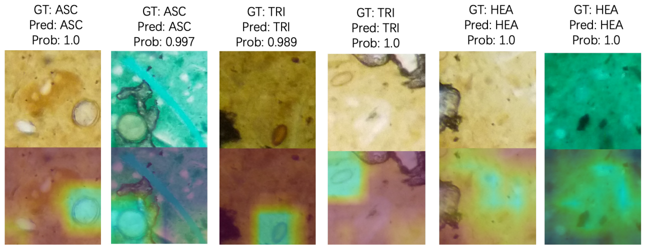

Fig. 3: Examples of image samples representing all classes under study (upper row) along whit its activation maps generated by the DL algorithm (bottom row). Correct label (GT) as well as predictions (Pred) appear above each image. Activation maps for healthy samples (HEA) focused on the entire image with no significant activation, while activation maps for *Ascaris* spp. (ASC) and *Trichuris* spp. (TRI) focused exactly on the egg location.

models perform similarly, highlighting the power of the use of crowdsourced annotations which obtained similar results when compared to the ones obtained by the model trained on expert-based annotations. It should be noted that although the quality of school-age children annotations were lower than the ones collected by adults (difference of 4.5% in the accuracy, see Table II), the AI algorithm is robust to noisy labels and decreases the difference in terms of the model performance (0.9%).

AUC	Experts	Adults	Children
Ascaris spp.	0.960 (0.003)	0.958 (0.007)	0.948 (0.002)
Trichuris spp.	0.912 (0.002)	0.926 (0.015)	0.912 (0.026)
Healthy	0.924 (0.019)	0.932 (0.010)	0.923 (0.008)
Mean	0.932 (0.015)	0.939 (0.01)	0.928 (0.012)

TABLE IV: Detailed performance of the DL algorithm trained with different annotations. Mean AUC and standard deviation is reported.

Figure 3 shows the prediction result of the model on different images from the test set, including the three classes under study (*Ascaris* spp., *Trichuris* spp. and healthy). In addition, we computed the gradient-weighted class activation mapping (Grad-CAM) to visualize the the regions in the image that is important for the model to make the decision [17].

IV. CONCLUSIONS

This work shows promising results on the use of crowdsourced annotation for the development of AI-based diagnosis systems, and validates its use in the medical image field, where manual annotations from experts is a time-consuming labour, and requires high specialization.

In this work, we collected crowdsourced annotations by using a customized video game to classify different species of Helminths eggs, and used these annotations to train a CNN architecture (MobileNet V2). Particularly, we obtained crowdsourced annotations from both untrained school-age children and adults, and the results showed that DL models trained on those annotations performed in a similar manner compared to the ones trained with expert annotations (AUC of 0.928, 0.939 and 0.932 for children, adults and expert annotations respectively). No significant differences in the model performance were found when used children, adults or experts annotations.

On the other hand, we showed the impact of the training sample size on the performance of the AI models. We showed that we obtained better performance as the training sample size increases. In particular, model performances increased approximately by a factor of 20% when trained on all available samples (700) compared to the result obtained when only 100 images were used for training. As derived from this experiment, we can conclude that we could train our DL model in an iterative manner as we obtain more images annotated by players, and thus obtaining a more robust algorithm with a better predictive capacity.

Additionally, this work lays the foundation for the use of video games as data enrichment platforms to automate and scale the medical image labeling process using human collective intelligence, enhancing human relevance in the process of developing AI algorithms.

ENVIRONMENTAL IMPACT

In this study a cumulative of 43 hours of computation was performed on GPU (Tesla T4), which 20 hours contributed to obtain the final results. The total emissions, estimated by MachineLearning Impact calculator presented in [18], was 0.9 kg of CO_2. Virtual machines that host our SpotWarriors game are estimated to emit 0.39 kg CO2eq per month.

ACKNOWLEDGMENTS

This project was granted by Nesta (innovation foundation) and supported by Red.es and FEDER (2018/C003/00010900 ”Impulso al Sector del Videojuego”) to develop SpotWarrios. LL and UPM members were supported by the industrial doctorate program of the Community of Madrid (IND2019/TIC-17167). We thank all schools that allowed us to organize workshops to collect annotations by using SpotWarriors game.

References

[1] United Nations. Transforming our world: the 2030 Agenda for Sustainable Development — Department of Economic and Social Affairs, 2015.

[2] U. J. Muehlematter et al. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *The Lancet Digital Health*, 0(0), jan 2021.

[3] F. Yang et al. Deep Learning for Smartphone-Based Malaria Parasite Detection in Thick Blood Smears. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1427–1438, may 2020.

[4] Vijayalakshmi A and Rajesh Kanna B. Deep learning approach to detect malaria from microscopic images. *Multimedia Tools and Applications*, 79(21-22):15297–15317, jun 2020.

[5] O. Holmström et al. Point-of-care mobile digital microscopy and deep learning for the detection of soil-transmitted helminths and Schistosoma haematobium. *Global Health Action*, 10(3), 2017.

[6] A. Yang 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 Neglected Tropical Diseases*, 13(8):e0007577, 2019.

[7] B. A. Mathison et al. Detection of intestinal protozoa in trichrome-stained stool specimens by use of a deep convolutional neural network. *Journal of Clinical Microbiology*, 58(6):1–13, 2020.

[8] WHO. Ending the neglect to attain the Sustainable Development Goals – A road map for neglected tropical diseases 2021–2030 (Geneva: World Health Organization), pp. 55. Accessed on 7th July 2020. Technical report, 2020.

[9] N. Katz et al. A simple device for quantitative stool thick-smear technique in Schistosomiasis mansoni. *Revista do Instituto de Medicina Tropical de Sao Paulo*, 14(6):397–400, nov 1972.

[10] M. A. Luengo-Oroz et al. Crowdsourcing malaria parasite quantification: An online game for analyzing images of infected thick blood smears. *Journal of Medical Internet Research*, 14(6):1–14, nov 2012.

[11] M. Linares et al. Collaborative intelligence and gamification for online malaria species differentiation. *Malaria Journal*, 18(1):21, dec 2019.

[12] A. Keshavan et al. Combining citizen science and deep learning to amplify expertise in neuroimaging. *Frontiers in Neuroinformatics*, 13:29, may 2019.

[13] M. Sandler et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Technical report, 2018.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. Technical report, 2015.

[15] O. Russakovsky et al. Best of both worlds: Human-machine collaboration for object annotation. Technical report, 2015.

[16] S. E. Reed et al. Training deep neural networks on noisy labels with bootstrapping. *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*, pp. 1–11, 2015.

[17] R. R. Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2016.

[18] A. Lacoste et al. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.