Revolutionizing Parasitic Infection Diagnosis in Northern Nigeria: An Integrated Machine Learning Approach for the Identification of Intestinal Parasites and Associated Risk Factors

Yusha'u El-Sunais * and Nkiru Charity Eberemu

Posted Date: 20 March 2024

doi: 10.20944/preprints202402.1217.v2

Keywords: Intestinal parasitic infections, Artificial Intelligence (AI), Machine Learning, You Only Look Once (YOLO) V8, CIFE, CMIM, DIRS, ICAP

 doi:10.20944/preprints202402.1217.v2

Article

Revolutionizing Parasitic Infection Diagnosis in Northern Nigeria: An Integrated Machine Learning Approach for the Identification of Intestinal Parasites and Associated Risk Factors

Nkiru Charity Eberemu [1] and Yusha'u El-Sunais [2,*]

[1] Federal University Dutsin-ma; neberemu@gmail.com
[2] Mind Harvest Innovators, Nigeria
* Correspondence: yelsunais2014@gmail.com; +234(0)7050700903

Abstract: Intestinal parasitic infections pose a significant public health challenge in Northern Nigeria, with prevalence rates ranging from 20% to 70%. Traditional diagnostic methods, primarily microscopic examination of stool samples, face limitations such as low sensitivity and high costs. This research addresses these challenges by proposing an Artificial Intelligence (AI)-based platform for the identification and counting of intestinal parasites. A total of 651 samples were collected from the 7 Northern State of Nigeria along with questionnaire collecting data on demographic, socio-economic, hygiene habits and environmental factors. Leveraging the You Only Look Once (YOLO) V8 model, trained on a dataset of 360 pre-processed and 467 annotated images, the AI model demonstrated promising performance metrics. Information-Theory-Based approaches including (CIFE, CMIM, DISR and ICAP) were used on various machine learning models (Naïve Bayes, Random Forest, Support Vector Machine and Decision Tree) to analyse the risk factors associated with the parasitic disease. The precision-recall curve, average precision, mean average precision, and F1 score indicated reliable detection and classification across various parasite classes. The accuracy (97%) and AUC (99%) scores shows that CIFE with Random Forest has the best performance indicating the significant risk factors associated with the parasitic diseases. The model exhibited a well-balanced trade-off between precision and recall, showcasing its potential as a cost-effective and accessible tool for improving the diagnosis and treatment of intestinal parasitic infections in resource-limited settings.

Keywords: intestinal parasitic infections; Artificial Intelligence (AI); machine learning; You Only Look Once (YOLO) V8; CIFE; CMIM; DIRS; ICAP

1. Introduction

Intestinal parasitic infections are a major public health problem in Northern Nigeria, with a high prevalence among both urban and rural populations. According to a study conducted by the Federal Ministry of Health in Nigeria (Olowokure et al., 2013), the prevalence of intestinal parasites in the country ranges from 20% to 50%. In Northern Nigeria, the prevalence of intestinal parasitic infections is particularly high, with some studies reporting rates as high as 70% (Adebisi et al., 2017). The most common types of intestinal parasites found in Northern Nigeria include Schistosoma mansoni, Ascaris lumbricoides, Taenia saginata, Entamoeba histolytica, and Giardia intestinalis (Okoh et al., 2011). These parasites can cause a range of symptoms, including diarrhea, abdominal pain, and malnutrition, and can lead to serious complications if left untreated (Okoh et al., 2011).

A study by Adebisi et al. (2017) found that the prevalence of Schistosoma mansoni and Ascaris lumbricoides among school-aged children in Northern Nigeria was 40.5% and 32.9%, respectively. Another study by Okoh et al. (2011) found that the prevalenceof Taenia saginata among adult cattle slaughterers in Northern Nigeria was 31.3%. These studies indicate that intestinal parasitic infections are prevalent among different population groups in Northern Nigeria, highlighting the need for effective diagnostic and treatment methods.

 doi:10.20944/preprints202402.1217.v2

The traditional method of diagnosing parasitic infections in Northern Nigeria is through microscopic examination of stool samples by trained medical personnel. However, this method is often limited by the lack of expertise and shortage of medical staff in laboratory, which can lead to difficulties in terms of identification and counting the parasites. For example, a study by Olowokure et al. (2013) found that the sensitivity of microscopy for identifying intestinal parasites was only 57.1%, indicating that many cases of infection may be missed using this method. In addition to the limitations of microscopy, the cost of laboratory testing can also be a barrier for many individuals, particularly in resource-limited settings (Adebisi et al., 2017).

Given these limitations, there is a clear need for alternative diagnostic methods that are accurate, cost-effective, and accessible, especially in resource-limited settings.

Artificial Intelligence (AI) has the potential to overcome these limitations by providing a more accurate, cost-effective, and accessible method of identifying and counting parasites (Aydin et al., 2018). Recent studies have shown that AI-based approaches have the potential to improve diagnostic accuracy for parasitic infections (Jain et al., 2018). For example, a study by Aydin et al. (2018) demonstrated that an AI-based model was able to accurately identify and classify parasitic eggs in fecal samples with high sensitivity and specificity. Similarly, a study by Jain et al. (2018) showed that an AI-based model was able to accurately identify and classify different types of intestinal parasites from microscopic images.

Also, understanding the risk factors associated with Intestinal Parasitic Infections (IPIs) is imperative for developing effective prevention and intervention strategies. Socioeconomic factors, water and sanitation conditions, hygiene practices, and environmental elements play pivotal roles in the transmission and prevalence of these infections (Abdulhadi, 2017). Recognizing these risk factors enables public health initiatives to be tailored to specific populations and regions, addressing the root causes of IPIs (Sadauki *et al*., 2022). Moreover, an in-depth understanding of risk factors facilitates targeted educational campaigns, fostering awareness and behavioral changes that can contribute to the overall reduction of IPIs (Younes *et al*., 2021).

Historically, the identification of risk factors for intestinal parasitic infections has relied on traditional statistical methods (Zafar *et al*., 2022). While these methods have provided valuable insights, they are not without limitations. Conventional statistical approaches often struggle to handle complex, nonlinear relationships and interactions among variables (Kattula *et al*., 2014). Moreover, these methods may face challenges in dealing with high-dimensional data, especially when the number of potential risk factors is substantial (Ranganathan *et al*., 2017). The need for expert input in selecting variables and potential biases in data interpretation are additional drawbacks (Zafar *et al*., 2022). As highlighted, traditional statistical approaches may not fully capture the complexities of risk factor identification, necessitating more advanced and sophisticated methodologies.

To overcome the limitations of traditional statistical methods, machine learning (ML) approaches, specifically feature selection techniques, offer a promising avenue (Oh *et al*., 2022) for identifying and understanding risk factors associated with Intestinal Parasitic Infections (IPIs) (Ranganathan *et al*., 2017). Advanced ML algorithms, such as Random Forest, Decision Tree, Naive Bayes, and Support Vector Machine, combined with feature selection methods like Conditional Infomax Feature Extraction (CIFE) by Ling & Zang (2006), Conditional Mutual Information Maximization (CMIM) by Fleuret (2004), Double Input Symmetric Relevance (DISR) as proposed by Meyer *et al*. (2008), and Interaction Causality and Prediction (ICAP) by Jakulin (2005), provide a robust framework for extracting relevant variables and relationships from complex datasets. These techniques enhance the accuracy and efficiency of risk factor identification, allowing for a more nuanced understanding of the multifaceted nature of IPIs. The integration of machine learning in risk factor analysis holds significant promise in advancing our comprehension of the determinants of intestinal parasitic infections in Northern Nigeria, contributing to more effective diagnosis and preventive strategies.

2. Methodology

 doi:10.20944/preprints202402.1217.v2

2. Sample Collections

A total of 651 fecal samples were collected and preserved from both government and private schools across Northern Nigeria, following the guidelines recommended by the World Health Organization (WHO) in 2013. The sampling was conducted in seven states: Jigawa, Kaduna, Kano, Katsina, Kebbi, Sokoto, and Zamfara, with each state further divided into three clusters corresponding to their Senatorial Zones. In total, 31 samples were collected from each Senatorial Zone, and 93 samples from each state. The collection process included the administration of a questionnaire to gather information from each participant.

The Formalin-Ether-Sedimentation technique was employed, involving the addition of a small quantity of feces to formalin and ether to preserve and concentrate the parasites within the samples. Research assistants, recruited for the study, facilitated the collection of these samples. Subsequently, all fecal samples were transported to the Biology Department Laboratory at Federal University Dutsin-ma for further investigation.

2.2. Data Collection

A comprehensive set of questionnaires was administered to gather data on various aspects, including demographic information, socioeconomic status, hygiene habits, and environmental factors, as presented in Table 1. The data provided information on the associated risk factors of Intestinal Parasites in Northern Nigeria.

Table 1. Risk Factors for Intestinal Parasitic Infection in Northern Nigeria.

Demography	Socioeconomic	Hygiene Habits	Environmental
State	Type of Toilet	Hand washing after Toilet Use	Taking-off Shoes while Playing
Age	Water Source	Using Soap for Hand washing after Toilet	Playing with Soil
Father Education	Presence of Pet at home	Hand Washing before Eating	Eating while Playing
Mother Education		Fingernails Cleanliness	Open Defecation
Farther Occupation		Eating of Raw Vegetables	
		Sucking Fingernails	

2.3. Parasitic Examination

For helminthes and eggs examination were done according to the technique proposed by Amin et al. (2020). The stool samples were examine using the iodine concentration method for parasitic eggs. The samples were then prepared and suspended in a formalin solution, then filtered and combined with ethyl acetate before centrifugation. Following centrifugation, the remaining sediment was examined microscopically to identify parasitic eggs. For the helminths examination, a swabbed sample was combined with saline and placed on a slide, ensuring no air bubbles were present. Direct microscopic examination was performed to detect helminth ova.

For intestinal protozoa, the technique of Fasipe et al. (2020) was employed. Stool samples were examined for the presence of parasitic eggs using formal ether concentration techniques. On gram of feces was suspended in 5ml of 10% SAF solution and mixed thoroughly. The sample mixture was decanted and a drop of the precipitate was picked using a pipette and then placed on a clean microscope slide and microscopic examination of stool samples for the presence of intestinal protozoan cysts or trophozoites was done by direct saline-Logol'siodine wet mount method (Cheesbrough, 2006).

2.4. Parasitic Image Capture

To capture images, a high-resolution microscope camera was employed. The Olympus DP74 microscope camera, with a resolution of 16 megapixels and the capability to capture high-quality

 doi:10.20944/preprints202402.1217.v2

images at different magnification levels (Olympus, 2021), was utilized. The camera was directly connected to the microscope, facilitating the straightforward capture and digitization of image. After each image, ATLAS pictorial guide for Intestinal Identification was employed to identify each image and categorized based on the parasitic eggs and or parasites. These images were subsequently employed for the training and evaluation of the AI model.

2.5. Image Data Pre-Processing and Annotations

The objective of image pre-processing was to enhance image quality, eliminate unwanted noise that could adversely affect the AI model's performance, and provide the model with labeled data for learning to identify various parasite types in the images.

Following the guidelines of Kuzborskij et al. (2020), the image pre-processing was done. This involved drawing image bounding boxes on each parasites, focusing specifically on the area of interest, which in this context was the different species parasites and the parasitic eggs. Techniques such as histogram equalization and contrast stretching were applied to enhance image visibility and emphasize the distinct features of the parasites. Additionally, color normalization was employed to ensure consistent color representation of parasites across all images.

For image annotations, the methodology proposed by Kuzborskij et al. (2020) was adopted. Specifically, the tool LabelImg was utilized to draw bounding boxes around parasites and classify them into different categories as presented in (Figure 1). The annotation process was carried out with the classified images initially identified using ATLAS pictorial guide for Intestinal Identification. This ensured the accuracy and consistency of annotations across all images.

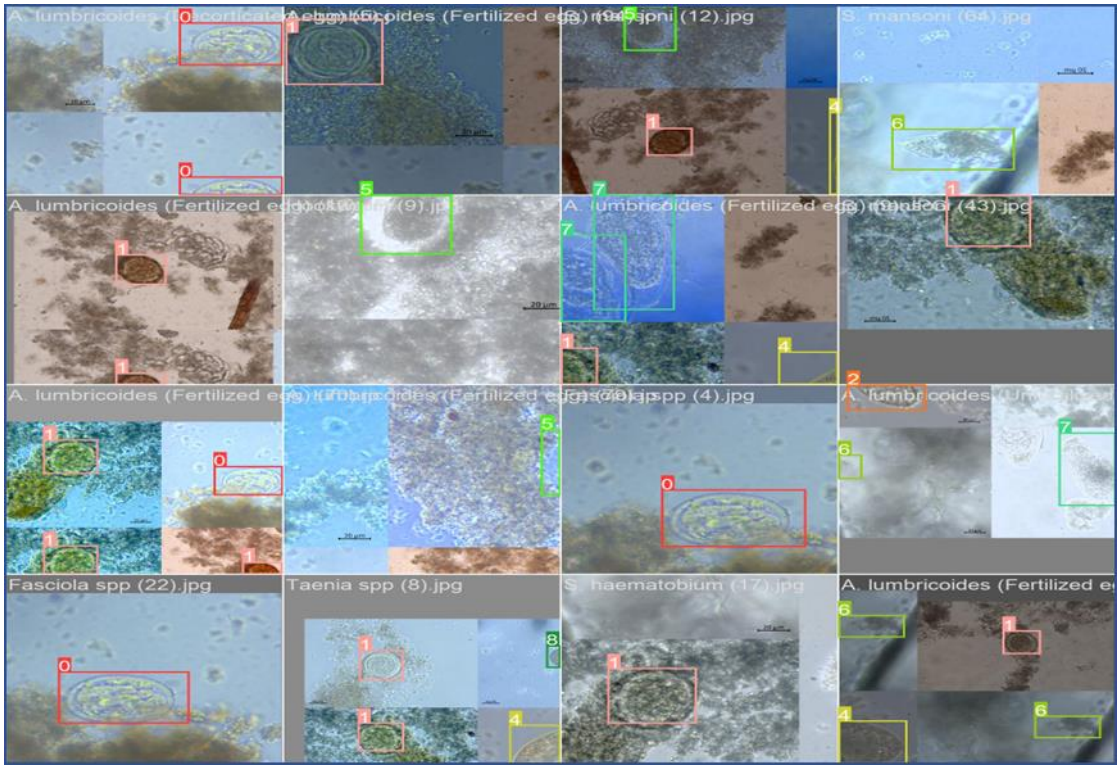


Figure 1. Pre-Processed Image with Bounding Boxes Annotation.

2.6. Risk Factors Assessment

To assess risk factors, the study employed four information-theoretical-based approaches. These approaches include Conditional Infomax Feature Extraction (CIFE), Conditional Mutual Information Maximization (CMIM), Double Input Symmetric Relevance (DISR), and Interaction Causality and Prediction (ICAP) respectively.

 doi:10.20944/preprints202402.1217.v2

The risk assessment was conducted using 10 thresholds to identify the 10 most significant risks determined by each method. The selected factors were then utilized for machine learning classifications

2.7. Building AI Model

2.7.1. Parasitic Detection

The AI model was implemented using You Only Look Once (YOLO) V8. YOLO, as described by Redmon et al. (2016), is a real-time object detection algorithm designed to identify and localize objects in an image.

The architecture of YOLO is founded on a single convolutional neural network (CNN) trained end-to-end to predict the class probability and bounding box coordinates for each object in an image. The algorithm partitions an image into a grid of cells, with each cell responsible for predicting the object present in it (Redmon et al., 2016).

In the context of identifying intestinal parasites, the CNN was trained on pre-processed and annotated images of parasites. The model was train to detect and classify various types of parasites present in the images. Once the model was trained, it became capable of detecting and classifying new images of parasites not included in the training dataset, as outlined by Jain et al. (2018).

2.7.2. IPI Disease Prediction

The model was trained by incorporating each of the ten significant factors derived from CIFE, CMIM, DIRS, and ICAP. The assessment of children's intestinal parasitic infection status employed Random Forest Classification, Support Vector Machine, Gaussian Naïve Bayes, and Decision Tree algorithms. This choice was made due to the utilization of advanced machine learning approaches, which operate under the premise that computers can discern intricate patterns and interactions within datasets using mathematical rules and statistical assumptions (Ranganathan *et al*., 2017). Unlike an epidemiological or statistical approach, machine learning does not depend on strong assumptions about data linearity or the mutual dependence of predictor variables. Instead, it relies on iterative computing techniques to extract insights from extensive datasets. Recent studies have employed diverse machine learning methods to accurately predict and identify relevant risk factors for disease outcomes.

2.7.3. Indicators for Model Performance Evaluation

In assessing the detection of malaria parasites in this study, metrics such as the precision-recall (P-R) curve, average precision (AP), and mean average precision (mAP) were utilized. Precision, a measure of accuracy in information retrieval contexts where precision and recall are often considered together, quantifies the ratio of relevant targets accurately identified among the returned results to the total number of targets returned for a specific query. The evaluation includes terms like true positive (TP), true negative (TN), false positive (FP), and false negative (FN) to describe classification outcomes. TP signifies the correct prediction of positive instances as positive, TN denotes the correct prediction of negative instances as negative, FP indicates negative instances incorrectly predicted as positive (false positives), and FN represents positive instances incorrectly predicted as negative (false negatives). The precision formula is expressed as follows:

$$Precision = \frac{TP}{TP + FP}$$

 doi:10.20944/preprints202402.1217.v2

Additionally, the recall rate, measuring the proportion of relevant targets among all relevant targets, is defined as:

$$Recall = \frac{TP}{TP+FN}$$

In certain cases, specific values offer a clearer representation of the test model's performance than a graphical representation. Average precision (AP) is commonly used for this purpose, calculated using the formula:

$$AP = \int_0^1 p(r)d(r)$$

In this formula, 'p' represents precision, 'r' represents recall, and precision is a function of recall. Therefore, the average precision corresponds to the area under the precision-recall (P-R) curve, and mAP (mean average precision) is the average of the average precision values across all categories.

3.0. Results and Discussion

3.1. Intestinal Parasitic Detection

A total of 467 parasites across difference eggs and species were identified. The highest recorded count was observed in *A. lumbrioides* (Fertilized Eggs), totalling 141 parasite eggs, followed by 80 parasites for *S. mansoni* and 54 parasites for *Taenia spp*. Additionally, 37 parasites were identified for *A. lumbrioides* (Decorticated Eggs), and 34 for *A. lumbrioides* (Unfertilized Eggs). On the lower end of the spectrum, E. histolytica exhibited 36 recorded parasites, *Fasciola spp*. showed 33 parasites, *S. haematobium* had 32 parasites, and Hookworm presented the least with 25 recorded parasites as shown in (Figure 2).

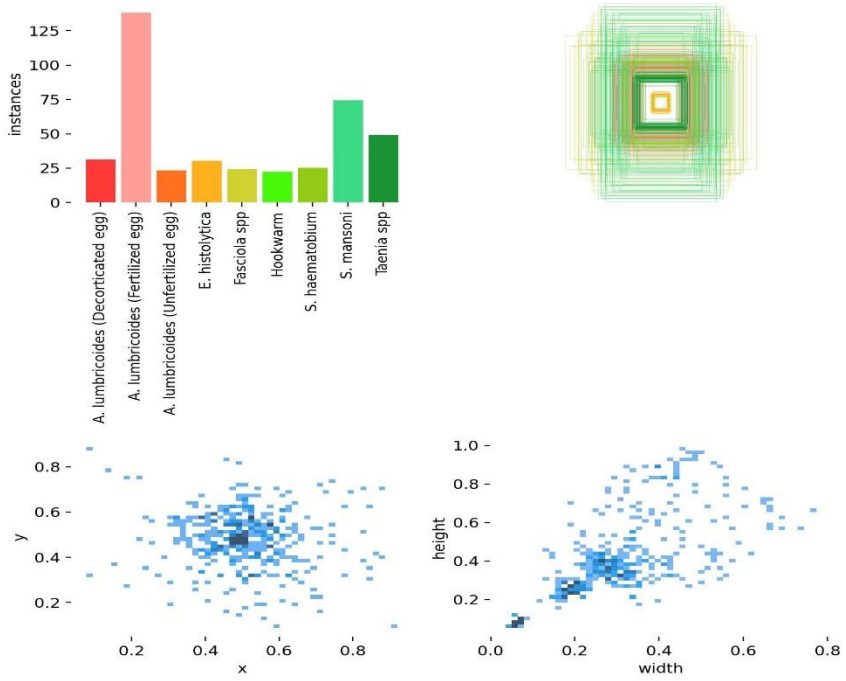


Figure 2. Dataset Distributions.

The scatter plot which depicts the size (width and height) of nine different categories of objects, including the eggs and different parasites. The x-axis represents the width of the objects, and the y-axis represents the height. Each data point is represented by a circle, and the size of the circle corresponds to the number of objects in that category with that particular width and height. The largest circle in the plot is at around (0.6, 0.4), which means there are many objects in that category that are 0.6 units wide and 0.4 units high. It is also worth noting that the data points are clustered

 doi:10.20944/preprints202402.1217.v2

into groups. This suggests that natural groupings of the objects based on their size being recognized to improve prediction of the parasites by YOLO.

The train/box loss, train/cls loss, and train/dfl loss metrics of the different aspects of the model's loss during training show decrease over time, indicating that the model is learning to make better predictions (Figure 3). Similarly, the val/box loss, val/cls loss, and val/dfl loss metrics of the model's loss on a validation dataset implying that the model has good generalization to unseen data (Figure 3). The decrease of the validation metrics over time, is not as much as the training loss metrics. However, this could be explain by the fact that the validation dataset is usually more challenging than the training dataset.

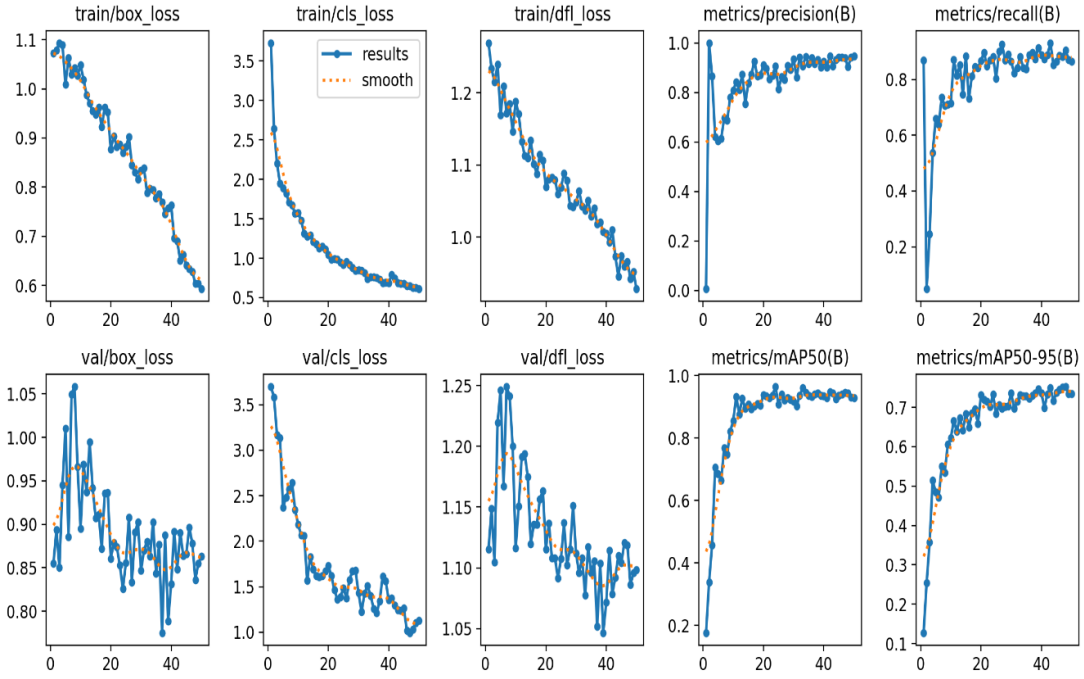


Figure 3. Model Training Performance Metrics.

The metrics/precision and metrics/recall metrics indicates a good measure of how well the model is able to correctly identify objects (precision) and how many true objects it misses (recall) during the training. Both metrics show increase over time, indicating that the model is getting better at detecting objects. Similarly the metrics/mAP50 and metrics/mAP50-95 metrics provides measure of the mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds indicating better performance as presented in (Figure 3).

The parasitic detection model exhibits strong performance across various parasite classes, achieving an F1 score of 90.0% at a predictions confidence of 56.4%. This indicates a well-balanced trade-off between precision and recall, showing the model's reliability in accurately detecting and classifying parasites as presented in (Figure 4). The Precision Confidence Curve for all the classes was 1.00 at 0.991 indicating that every positive prediction made by the model is correct. There are no false positives. The Recall Confidence Curve shows the recall for all classes of 0.98 at 0.00 indicating that the model is capturing a high proportion of the actual positive instances no matter how small confidence.

 doi:10.20944/preprints202402.1217.v2

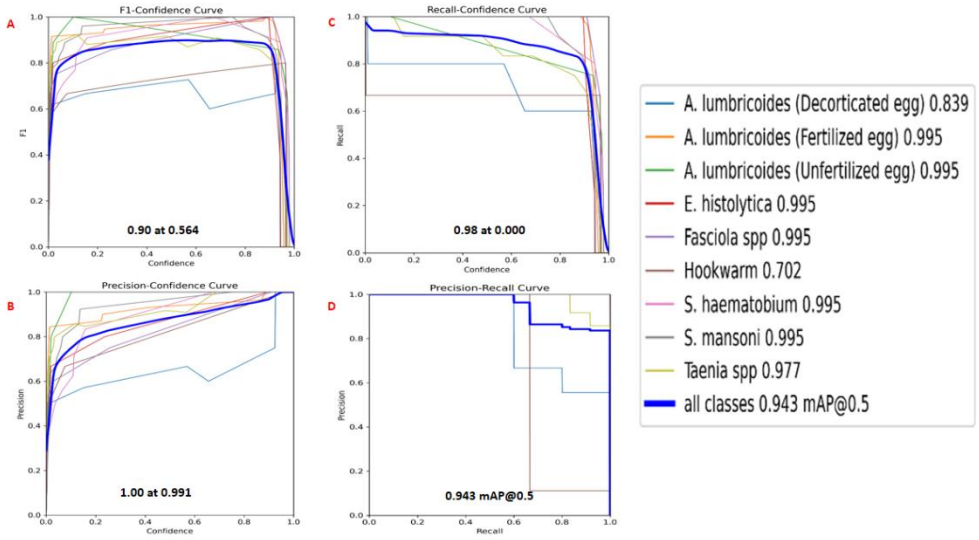
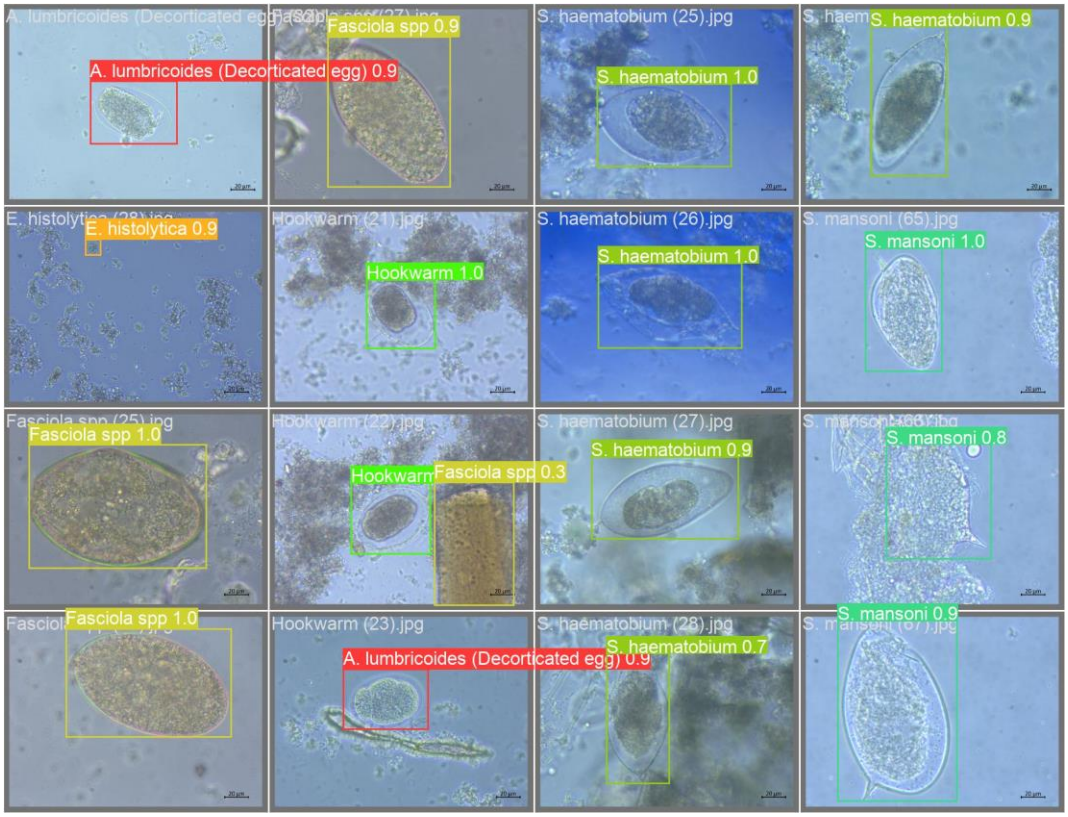


Figure 4. Performance Evaluation Metrics (A) F1-Confidence Curve, (B) Precision-Confidence Curve, (C) Recall Confidence Curve and (D) Precision-Recall Curve.

The "all classes" curve has a relatively high AUC, suggesting that the model performs well on average across all classes. Most of the classes, such as "*A. lumbricoides* (Fertilized egg)", "*A. lumbricoides* (Unfertilized egg)", "*Fasciola spp*", "*S. haematobium*", "*S. mansoni*" and "*E. histolytica*", have curves that are closer to the top-left corner with score of 0.995 each, indicating better precision and recall compared to other classes. Confidence for detecting "*Taenia spp*" recorded a score of 0.977, followed by "*A. lumbricoides* (Fertilized egg)" with a score of 0.839 and lastly, the least score of 0.702 was recorded for detecting Hookworm as shown in (Figure 4). The model was tested against unseen dataset and the results are shown in (Figure 5).



 doi:10.20944/preprints202402.1217.v2

Figure 5. Predicted Images of various Intestinal Parasites with predication accuracy scores.

3.2. Risk Factors Assessment

The risk factors associated with IPIs in this study shows that the "State" variable is significant according to both CIFE and CMIM methods, suggesting that geographical location may play a role in the prevalence of intestinal parasitic infections. This aligns with existing research that indicates variations in parasitic infections based on regional factors such as climate, sanitation infrastructure, and socioeconomic conditions (Abbas *et al*., 2023). Age is highlighted as a significant risk factor by CIFE, emphasizing that certain age groups may be more susceptible to intestinal parasitic infections. This corresponds with established literature indicating that children and the elderly are often more vulnerable due to weaker immune systems and different behavioral patterns (Eberemu and Magaji, 2017).

Family size is identified as a significant risk factor by three out of the four methods (CIFE, CMIM, and ICAP), implying that larger households may be at an increased risk of parasitic infections. This aligns with research that associates overcrowded living conditions with higher transmission rates of infectious diseases (Eberemu, 2018).

Table 2. Assessment of Risk Factors Associated with Intestinal Parasitic in North-Western Nigeria.

Risk Factors	CIFE	CMIM	DISR	ICAP
State	*	*	-	-
Age	*	-	-	-
Family Size	*	*	*	*
Father Education	*	*	-	*
Mother Education	*	*	*	*
Father Occupation	-	*	-	*
Presence of Pets at Home	-	-	-	-
Source of Water	-	-	*	-
Type of Toilet	*	*	*	*
Hand Washing after Toilet Use	-	-	-	-
Using Soap for Hand washing after Toilet	-	-	-	-
Hand Washing before Eating	*	*	*	*
Fingernails Cleanliness	-	*	*	*
Eating of Raw Vegetables	-	-	-	-
Taking-off Shoes while Playing	-	-	-	-
Playing with Soil		*	*	*
Eating while Playing	*	-	*	-
Sucking Fingernails	*	-	*	*
Open Defecation	*	*	*	*

*Significant -Not Significant.

The Parental education is significant by three methods (CIFE, CMIM, and ICAP), emphasizing the role of education in influencing hygiene practices and health awareness. Mother's education is particularly highly significant, suggesting that a mother's level of education may impact the household's overall hygiene practices. Father's occupation is identified as a significant risk factor by the CMIM method, indicating that certain occupations may pose a higher risk of exposure to parasitic infections. This finding supports the idea that occupational environments can contribute to the transmission of infectious diseases (Sani *et al*., 2024).

The presence of pets at home is not deemed significant by any method, suggesting that pet ownership may not be a major contributor to intestinal parasitic infections in the studied population.

The source of water is identified as a significant risk factor by the ICAP method, underscoring the importance of clean water sources in preventing parasitic infections. This aligns with established

 doi:10.20944/preprints202402.1217.v2

literature on waterborne diseases and the critical role of safe water supply in public health (Sani *et al*., 2024).

The type of toilet is recognized as a significant risk factor by three methods (CIFE, CMIM, and ICAP), emphasizing the importance of proper sanitation facilities in preventing parasitic infections. This finding supports existing research on the link between inadequate sanitation and the prevalence of intestinal parasites (Abbas *et al*., 2023).

Handwashing practices are consistently identified as significant risk factors by all four methods, particularly handwashing before eating and after using the toilet. Fingernail cleanliness is highlighted as a significant risk factor by CMIM and DIRS methods, emphasizing the importance of maintaining clean fingernails to reduce the risk of parasitic infections (Eberemu, 2018).

Playing with soil is identified as a significant risk factor by two methods (CMIM and DIRS), highlighting the potential transmission of parasites through contact with contaminated soil.

The findings regarding eating while playing, sucking fingernails, and open defecation suggest that these behaviors are significant risk factors for intestinal parasitic infections, as indicated by multiple asterisks across the feature selection methods (Eberemu and Magaji, 2017).

3.3. Machine Prediction of Parasitic Disease

The performance of the models were generally high, suggesting that the information-theory-based feature selection methods are effective for this prediction task. Naïve Bayes performs consistently across different feature selection methods, with scores ranging from 0.87 to 0.90, although it performs slightly better with CMIM.

Support Vector Machine (SVM) also shows consistent performance, with scores from 0.89 to 0.91. Interestingly, SVM performs best with the DISR feature selection method, suggesting that DISR might be particularly effective at identifying the most relevant features for SVM in this prediction task.

Random Forest shows the highest variability in performance based on the feature selection method used, with scores ranging from 0.86 (ICAP) to 0.97 (CIFE). However, Random Forest achieves the highest performance of all models and methods with CIFE, indicating a potentially significant synergy between CIFE's feature selection capabilities and Random Forest's prediction model for this task.

Decision Tree exhibits moderate performance compared to the other models, with scores between 0.84 (DISR, ICAP) and 0.92 (CIFE). This indicates that while Decision Trees can benefit from effective feature selection as seen with CIFE, they may be more sensitive to the choice of feature selection method than other models.

Table 3. Accuracy Scores of Machine Learning Models for Parasitic Disease Prediction using Information-Theory-Based approaches.

Items	CIFE	CMIM	DISR	ICAP
Naïve Bayes	0.89	0.90	0.87	0.89
Support Vector Machine	0.89	0.90	0.91	0.89
Random Forest	0.97	0.92	0.92	0.86
Decision Tree	0.92	0.89	0.84	0.84

For the Receiver Operating Characteristic (ROC) and Area under the Curve (AUC), Naïve Bayes + CIFE has an AUC of 0.9562, which indicates a very good predictive ability. It is outperformed slightly by Random Forest + CIFE. Random Forest + CIFE has the highest AUC value of 0.9900, suggesting that this model with this feature selection method has an excellent predictive ability. Decision Tree models generally have lower AUC values compared to Naïve Bayes and Random

 doi:10.20944/preprints202402.1217.v2

Forest. The best-performing Decision Tree model is when paired with CIFE, with an AUC of 0.9159 as presented in Figure 6.

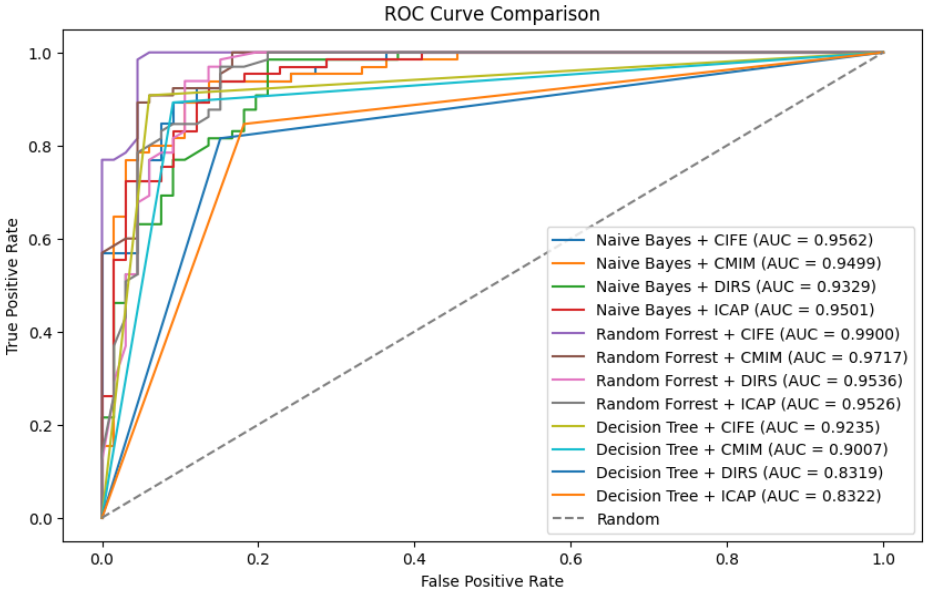


Figure 6. Performances (Receiver Operating Characteristics and Area Under Curve) of the models on the Information-Theory-Based Approaches. Showing Random Forest with CIFE has the best performance.

The performance of the models is generally high, suggesting that the information-theory-based feature selection methods are effective for this prediction task. The Random Forest algorithm, particularly when paired with CIFE, appears to be the most effective combination for predicting parasitic diseases in this study, achieving the highest. DISR and CMIM seem particularly effective for SVM and Naïve Bayes, respectively, while CIFE stands out for Random Forest, highlighting how different feature selection methods may favor different machine learning algorithms. This means that geographical location, age, family size, parental education, type of toilet, hand washing before eating, playing while eating, sucking fingernails as well as open defecation as identified by CIFE are the important risk factors associated with intestinal parasitic disease in Northern-Nigeria.

4. Conclusion

This study underscores the pressing need for innovative approaches to combat the high prevalence of intestinal parasitic infections in Northern Nigeria. The proposed AI-based platform, utilizing the YOLO V8 model, has shown promising results in accurately identifying and classifying various parasites in fecal samples. The model's robust performance metrics, including precision, recall, and F1 score, highlight its potential as an effective diagnostic tool. Also, the machine learning models shows that Information-Theory-Based approaches could provide an effective method of disease risk assessment in Northern-Nigeria. By overcoming the limitations of traditional methods, the AI model offers a more accessible and cost-effective solution. Implementation of this technology in healthcare settings could significantly improve the diagnosis and treatment of intestinal parasitic infections, particularly in regions facing resource constraints.

Acknowledgments: This research was funded by the Tertiary Education Trust Fund (TETFUND). We extend our heartfelt appreciation to TETFUND for their commitment to advancing educational research and development.

 doi:10.20944/preprints202402.1217.v2

References

Abbas, U., Eberemu, N., Orpin, J., & Kaware. (2023). Human Water Contact Behaviour and Schistosoma haematobium Infection among Almajiri School Children in Kurfi Local Government Area of Katsina State, Nigeria. *Nigerian Journal of Parasitology*, *44*(1), 37–47. https://doi.org/10.4314/njpar.v44i1.4

Abdulhadi, B. J. (2017). Survey on prevalence of intestinal parasites associated with some primary school aged children in Dutsinma area, Katsina State, Nigeria. *MOJ Biology and Medicine*, *2*(2). https://doi.org/10.15406/mojbm.2017.02.00044

Adebisi, S. A., Adebayo, K. A., & Adeleye, O. (2017). Prevalence and associated risk factors of Schistosoma mansoni and Ascaris lumbricoides among school-aged children in Ogun State, Nigeria. *International Journal of Tropical Disease & Health*, 14(1), 1-11.

Amin, R., Rajendran, C., & Sundar, S. (2020). DNA barcoding for the identification of human intestinal parasites. *Journal of Parasitology Research*, 2020.

Aydin, B., Ozer, S., & Yildirim, I. (2018). Automated detection of parasitic eggs in fecal samples using deep learning. *Journal of Medical Systems*, 42(11), 216.

Aydin, B., Ozer, S., & Yildirim, I. (2018). Automated detection of parasitic eggs in fecal samples using deep learning. *Journal of Medical Systems*, 42(11), 216.

Eberemu, N. C. (2018). Impact of *Schistosoma haematobium* Infection and Starvation on Some Neutral and Polar Lipids Content of *Bulinus truncatus*. *Fudma journal of sciences - ISSN: 2616-1370*, *2*(3), 224–232. http://journal.fudutsinma.edu.ng/index.php/fjs/article/view/336/0

Eberemu, N., & Magaji, H. (2017). The use of microscopy and rapid diagnostic test in diagnosing the prevalence of malaria among women attending antenatal clinic in Dutsin Ma, Katsina State, Nigeria. *Nigerian Journal of Parasitology*, *38*(2), 215. https://doi.org/10.4314/njpar.v38i2.15

Fasipe, K. A., Adebayo, K. A., Olowokure, B., & Ojo, S. A. (2020). Multiplex PCR for the simultaneous detection of multiple intestinal parasitic infections. *Journal of Parasitology Research*, 2020.

Fleuret, F. (2004). Fast Binary Feature Selection with Conditional Mutual Information. *Journal of Machine Learning Research*, *5*, 1531–1555. http://jmlr.csail.mit.edu/papers/volume5/fleuret04a/fleuret04a.pdf

Jain, P., Jain, P., & Jain, S. (2018). Automated detection of intestinal parasites in microscope images using deep learning. *Journal of Medical Systems*, 42(11), 216.

Jain, P., Jain, P., & Vatsa, M. (2018). A survey of deep learning methods for object detection. *arXiv preprint arXiv*:1803.03453.

Jakulin, A. (2005). *Machine learning based on attribute interactions : phd dissertation*. https://repozitorij.uni-lj.si/IzpisGradiva.php?id=24291

Kattula, D., Sarkar, R., Ajjampur, S. S. R., Minz, S., Levecke, B., Muliyil, J., & Kang, G. (2014). Prevalence & risk factors for soil transmitted helminth infection among school children in south India. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/24604041

Kuzborskij, I., Chen, T., & Wang, Y. (2020). Building high-quality datasets for medical imaging: A survey. *IEEE Transactions on Medical Imaging*, 39(6), 1499-1521.

Lin, D., & Tang, X. (2006). Conditional InfoMax Learning: an integrated framework for feature extraction and fusion. In *Lecture Notes in Computer Science* (pp. 68–82). https://doi.org/10.1007/11744023_6

Meyer, P., Schretter, C., & Bontempi, G. (2008). Information-Theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, *2*(3), 261–274. https://doi.org/10.1109/jstsp.2008.923858

Odinaka, K. K., Nwolisa, E., Mbanefo, F., Iheakaram, A. C., & Okolo, S. (2015). Prevalence and Pattern of Soil-Transmitted Helminthic Infection among Primary School Children in a Rural Community in Imo State, Nigeria. *Journal of Tropical Medicine*, *2015*, 1–4. https://doi.org/10.1155/2015/349439

Oh, T. I., Kim, D., Lee, S., Won, C. W., Kim, S., Yang, J., Yu, J., Kim, B., & Lee, J. (2022). Machine learning-based diagnosis and risk factor analysis of cardiocerebrovascular disease based on KNHANES. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-06333-1

Okoh, A. I., Mafiana, C. F., & Ejezie, G. C. (2011). Prevalence and intensity of Taenia saginata infection among cattle slaughterers in Abakaliki, Nigeria. *Journal of Helminthology*, 85(03), 307-314.

Olowokure, B., & Ojo, S. (2013). The burden of intestinal parasitic infections in Nigeria. *Journal of Parasitology Research*, 2013.

Olympus (2021). Olympus DP74 microscope camera. Retrieved from https://www.olympus-lifescience.com/en/microscope-resource/microscope-camera/dp74/

Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *DOAJ (DOAJ: Directory of Open Access Journals)*, *8*(3), 148–151. https://doi.org/10.4103/picr.picr_87_17

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779-788).

Sadauki, M. A., Dauda, A. B., & Yusuf, M. A. (2022). Prevalence of Gastrointestinal Helminths of African Catfish (Clarias gariepinus BURCHELL 1822) IN ZOBE Reservoir, Katsina State, Nigeria. *FUDMA Journal of Agriculture and Agricultural Technology*, *8*(1), 123–130. https://doi.org/10.33003/jaat.2022.0801.080

Sani, B. M., Auta, T., Orpin, J. B., & Yusuf, N. D. (2024). Gastrointestinal Protozoa and Geohelminth Infections and their Associated Risk Factors Among Primary School Pupils in Bindawa Local Government Area of Katsina State, Nigeria. *Trends in Medical Research*, *19*(1), 13–22. https://doi.org/10.3923/tmr.2024.13.22

Socioeconomic statistics - Nigeria Data Portal. (n.d.). Knoema. https://nigeria.opendataforafrica.org/iynrgrf/socioeconomic-statistics?states=1000210-katsina

Tefera, E., Belay, T., Mekonnen, S. K., Zeynudin, A., & Belachew, T. (2017). Prevalence and intensity of soil transmitted helminths among school children of Mendera Elementary School, Jimma, Southwest Ethiopia. *The Pan African Journal of Medicine*, *27*. https://doi.org/10.11604/pamj.2017.27.88.8817

WHO (2013). Laboratory diagnosis of intestinal parasites. Retrieved from https://www.who.int/publications/i/item/9789241548472

World Health Organization: WHO. (2023). Soil-transmitted helminth infections. *www.who.int*. https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections

Younes, N., Behnke, J. M., Ismail, A., & Abu-Madi, M. (2021). Socio-demographic influences on the prevalence of intestinal parasitic infections among workers in Qatar. *BMC Parasites & Vectors*, *14*(1). https://doi.org/10.1186/s13071-020-04449-9

Zafar, A., Attia, Z., Tesfaye, M., Walelign, S., Wordofa, M., Abera, D., Desta, K., Tsegaye, A., Ay, A., & Taye, B. (2022). Machine learning-based risk factor analysis and prevalence prediction of intestinal parasitic infections using epidemiological survey data. *PLOS Neglected Tropical Diseases*, *16*(6), e0010517. https://doi.org/10.1371/journal.pntd.0010517

Zhou, J., Xie, X., & Yang, Y. (2020). Color normalization in histopathological images: A review. *IEEE Journal of Biomedical and Health Informatics*, 24(5), 1667-1680.