Chapter 3

Dataset of Microscopic Images of Parasite Eggs and Ground Truths

The majority of computer vision approaches and models are data-driven. A dataset with sufficient data is necessary for the training and validation process of any computer vision model. For the automatic parasite egg detection and identification task, a standard dataset that contains various types of microscopic images of parasite eggs is very much required. However, during our research, it has been observed that most of the works used datasets that were prepared by themselves or collected from various organizations, which are not publicly available. These datasets vary in terms of the number of images and classes of parasite eggs. The size of the parasite eggs, the amount of debris, and the colour of the images also vary in different research studies.

Among the various works, it is observed that very few used images that contain a significant amount of debris and multiple different types of parasite eggs in a single image. This research attempts to collect and prepare a dataset of microscopic images of fecal samples from pigs that contain multiple parasite eggs of the same and/or different types with varying amounts of debris. Microscopic images captured from a fecal sample of pigs of the three most prevalent types of parasite eggs, namely, Roundworm (Ascaris lumbricoides), Hookworm (Ancylostoma duodenale / Necator americanus) and Whipworm (Trichuris trichiura) are collected. The first set of data containing images of Roundworm and Hookworm eggs is provided by *Dr. Nagappa S. Karabasanavar, Karnataka Veterinary, Animal & Fisheries Sciences University*, under the project *"E-Varaha Information System for Safe Pork Production in North-Eastern Region of India"*, sponsored by *Information Technology Research Academy, India.* These images contain single as well as multiple parasite eggs, mostly of the same species, with little to a heavy amount of debris. They also provided a few images that do not contain only debris without any parasite eggs. A few examples of the images are shown in Figure 3-1.

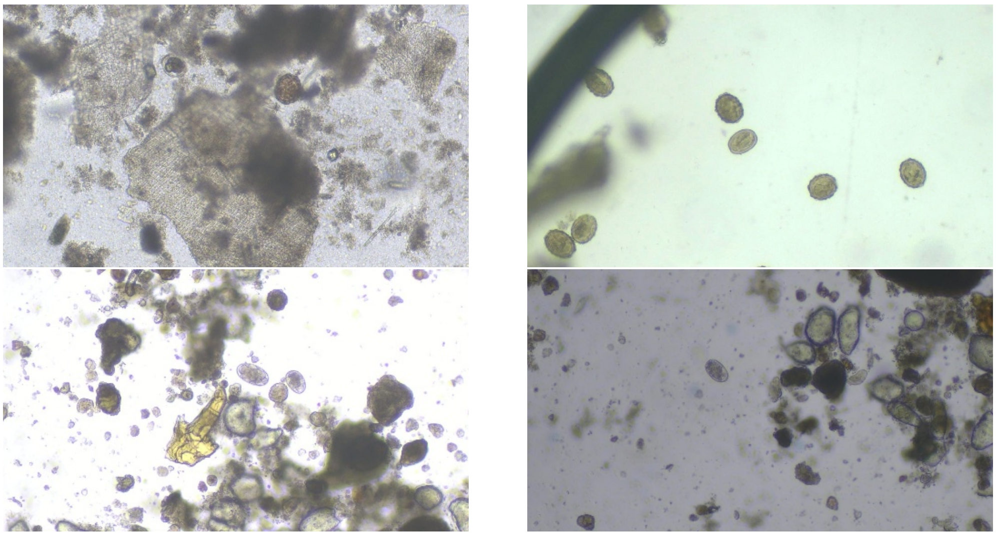

Figure 3-1: A few microscopic images of fecal samples containing parasite eggs of Asacaris (first row) and Neactor (second row), provided by Dr. Nagappa

The second set of images, which contains Whipworm as well as Roundworm eggs in different sizes, is provided by Raafat Salih Hadi from Universiti Malaysia Pahang. The difference between the images from the earlier source is that a few images in this dataset contain parasite eggs with a higher zooming effect, as shown in Figure 3-2. However, in most of the images, the size of the parasite eggs ranges from 100 to 450 pixels, approximately. Some of the images contain single, variable-sized parasite eggs, while others contain multiple eggs of the same or different types with a heavy amount of debris.

These images are used in various tasks such as segmentation, classification, and object detection. A dataset of microscopic images along with the ground truths to perform various image processing, machine learning, and deep learning approaches is prepared. The following sections describe the different approaches we used to create datasets for image segmentation, classification, and object detection.

3.1 Dataset for traditional image segmentation

Microscopic images serve as the primary source of information for studying parasite eggs in an automatic system. However, due to variations in imaging conditions

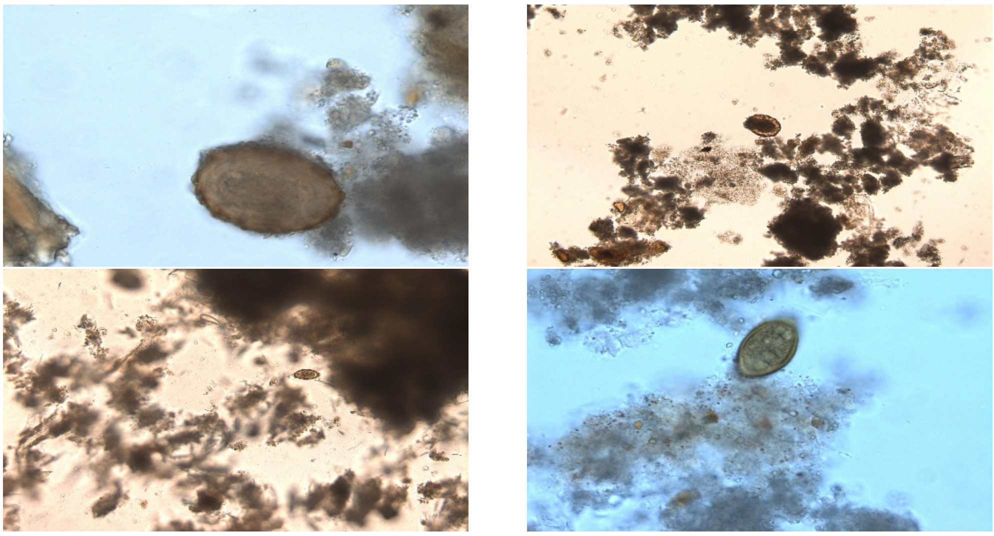

Figure 3-2: A few microscopic images of fecal samples containing parasite eggs of Asacaris (First row) and Trichuris (Second row), provided by Mr. R. S. Hadi [2]

and specimen preparation, these images often exhibit discrepancies in quality. Some images may be blurry, overly bright, or dark, which can impact the performance of the segmentation approach. In our dataset, there are a few images that are excessively blurry, where the shape and texture of the eggs are distorted, as shown in Figure 3-3. To standardize the dataset, the images are carefully examined, and those where parasite eggs appear excessively blurry are removed. Finally, a total of 1,444 images containing three types of parasite eggs and 73 images without any eggs are selected. Table 3.1 presents the total number of images for each type of parasite egg in the final dataset.

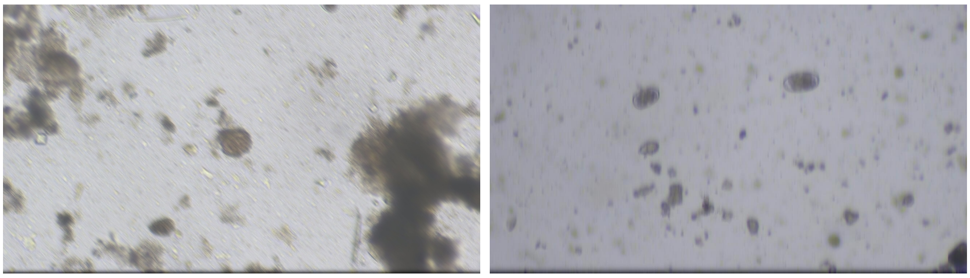

Figure 3-3: Example of blurry images that are discarded from the dataset

Table 3.1: Dataset containing microscopic images of different types of parasite egg

Parasite Egg Type	**Number of Images**
Ascaris (Roundworm)	839
Necator (Hookworm)	230
Trichuris (Whipworm)	375
Not Containing Egg	73
Total	1517

3.2 Dataset for classification

Preparing a dataset for machine learning and convolutional neural network (CNN)-based classification, aimed at identifying distinct types of parasite eggs is a detailed process. The dataset is prepared in several steps. Most of the steps are performed after the image segmentation process. Following the image segmentation process, four types of objects from the images, including three types of parasite eggs and non-egg objects, are extracted. It is observed that the segmentation process failed to detect a few parasite eggs from some of the images. These images are identified, and the parasite eggs that aren't successfully segmented are manually cropped and added to the dataset. Subsequently, various types of parasite eggs and non-egg objects are separated. A few examples of segmented objects of different categories are shown in Figures from 3-4 to 3-7.

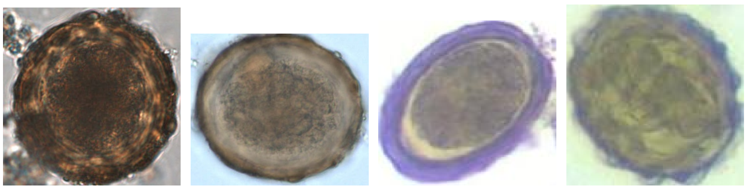

Figure 3-4: A few segmented images of Ascaris (Roundworm) eggs

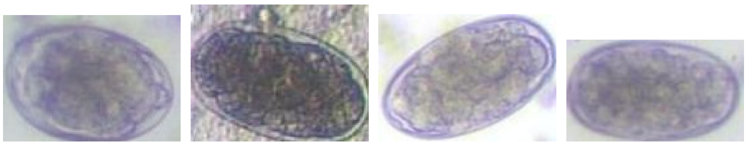

Figure 3-5: A few segmented images of Necator (Hookworm) eggs

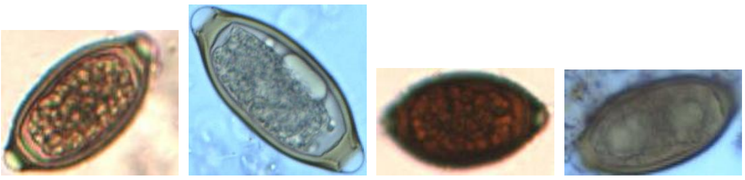

Figure 3-6: A few segmented images of Trichuris (Whipworm)

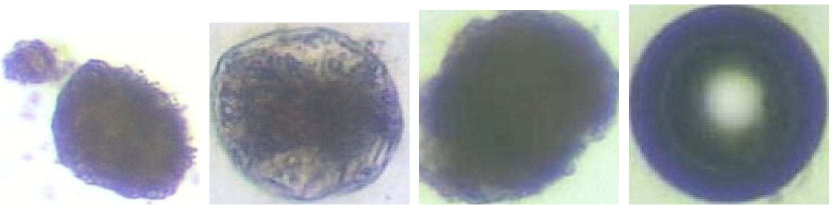

Figure 3-7: A few segmented images of non-egg objects or debris

After the segmentation process, a significant number of objects from the microscopic images of parasite eggs are extracted. It is observed that the counts of Ascaris eggs and non-egg objects are notably higher than the other two classes of parasite eggs, as mentioned in Table 3.2. To address this, a few data augmentation techniques are implemented. Data augmentation involves generating new training images through various transformations, including rotation, flipping, blurring, and shifting. By adding augmented images along with the existing images, this work aims to balance the representation of different object categories and improve the performance of classification models. This approach not only enriches the diversity of the dataset but also enhances the robustness of the classification algorithm, enabling it to better generalize to unseen data and accurately identify the objects.

Rotations of the images at angles of 45°, 90°, 135°, 180°, and 270° are applied. Additionally, horizontal and vertical flipping, along with Gaussian blurring, are also employed. Left and right shifts ranging from 10% to 30% of the image dimensions are used. To achieve a balanced distribution of samples across all classes, rotation and flipping are used exclusively for the Ascaris eggs. Conversely, for the other two types of parasite eggs, all the mentioned augmentation methods are used. However, due to the significantly larger quantity of non-egg objects, it is decided not to apply any data augmentation techniques to this particular class of objects. Following the data augmentation, a total of $14,640$ segmented images are prepared for classification tasks. Table 3.2 provides a breakdown of the segmented images for each class post-data augmentation process.

Table 3.2: Number of Segmented Objects

Object Category	**Quantity Before Data Augmentation**	**Quantity After Data Augmentation**
Ascaris (Roundworm)	1253	3655
Necator (Hookworm)	352	3618
Trichuris (Whipworm)	436	3644
Non-Egg (Debris)	3723	3723

3.3 Dataset for deep learning-based segmentation

Deep learning models for image segmentation heavily rely on the quality and characteristics of the dataset, including annotation accuracy, sufficient sample quantity, and a balanced class distribution for training. To maintain a balanced dataset, a few data augmentation techniques are employed, including flipping (horizontal and vertical), rotation, shifting, and blurring, specifically on the original microscopic images containing hookworm and whipworm. This is because their quantity is lower compared to roundworm egg images. A few similar-looking images, containing roundworm eggs, are discarded to ensure the robustness of the dataset. Finally, a dataset is prepared for applying CNN-based segmentation approaches, as detailed in Table 3.3.

Table 3.3: Image dataset containing different types of parasite eggs for CNN-based segmentation

Parasite Egg Type	Number of Images
Ascaris (Roundworm)	895
Necator (Hookworm)	880
Trichuris (Whipworm)	890
Total	2665

After preparing the dataset, a graphical image annotation tool called LabelImg is used to generate ground truths for these images [105, 106]. The tool allows to draw outlines around each object and specify the class of the objects to define a ground truth mask. With this approach, all three classes of parasite eggs present in the images, along with various non-egg objects that exhibit similar properties, such as shape, size, and texture, to the parasite eggs are annotated. The annotations are saved in Pascal VOC XML format, which is widely supported by most deep learning-based approaches. Examples of ground truth segmentation masks are shown in Figures 3-8 to 3-10.

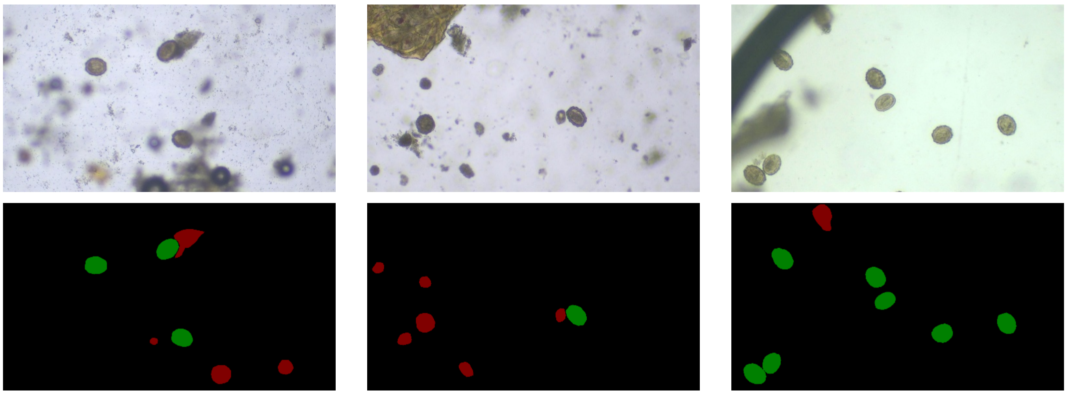

Figure 3-8: Original Image (Row 1) and Corresponding Segmentation Mask (Row 2) indicating Ascaris or Roundworm Eggs in Green and Non-Egg objects in Red

A multi-class segmentation approach relies primarily on the quality of the segmentation masks. Additionally, these masks can also be adapted for a binary segmentation task. In this alternate approach, all annotated objects, regardless of their class, are considered foreground objects, as shown in Figure 3-11. This modification extends the usability of the segmentation masks, enabling their application across various segmentation techniques and analysis methods.

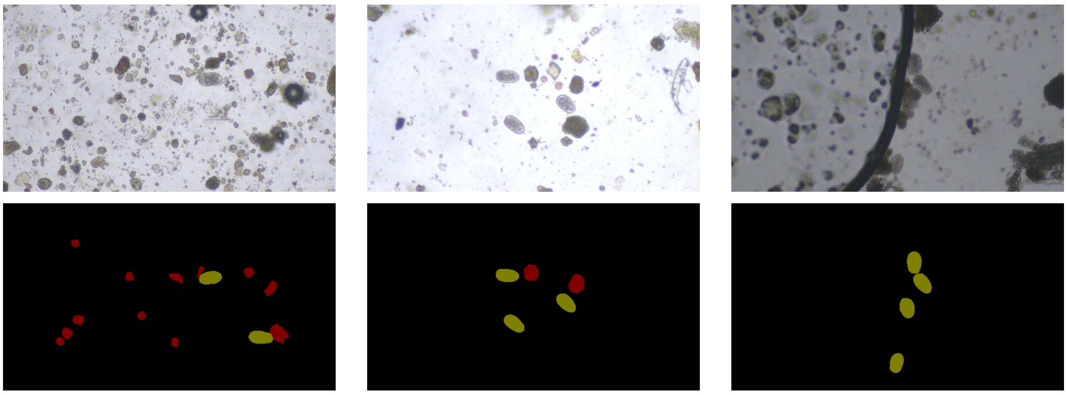

Figure 3-9: Original Image (Row 1) and Corresponding Segmentation Mask (Row 2) indicating Necator or Hookworm Eggs in Green and Non-Egg objects in Red

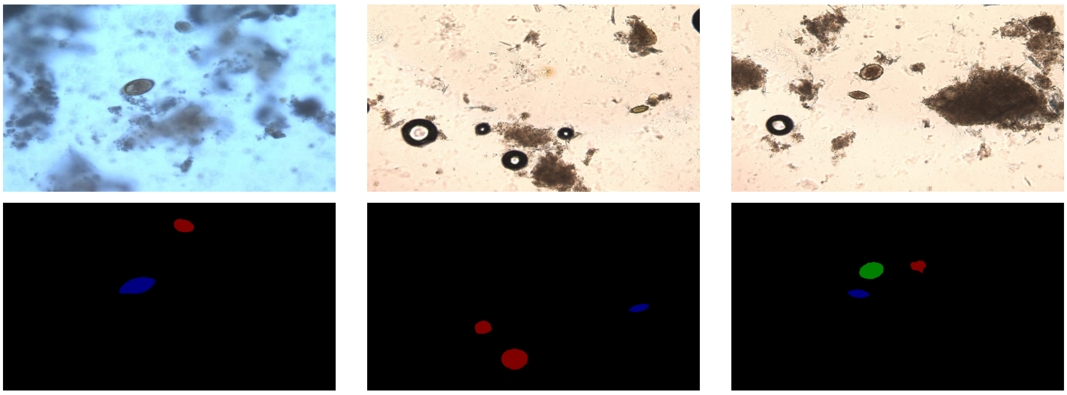

Figure 3-10: Original Image (Row 1) and Corresponding Segmentation Mask (Row 2) indicating Trichuris or Whipworm Eggs in Green and Non-Egg objects in Red

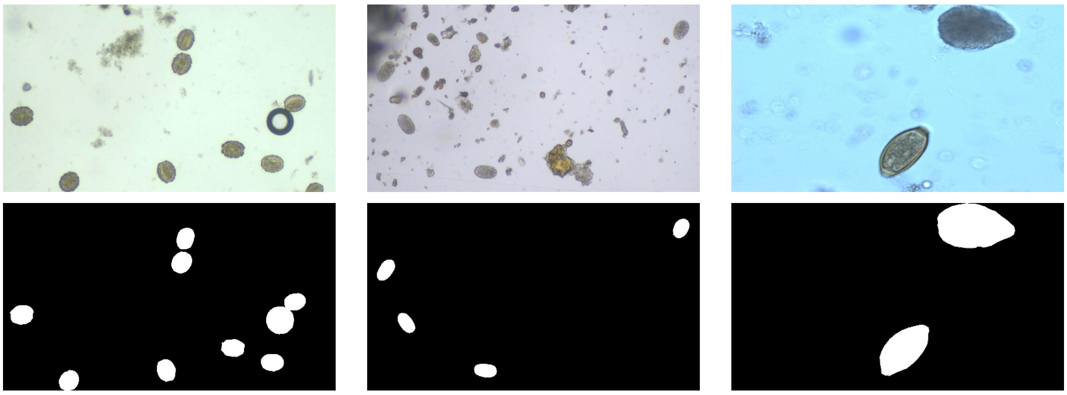

Figure 3-11: Original Image (Row 1) and Corresponding Binary Segmentation Mask (Row 2) indicating parasite eggs and other non-egg objects in white

3.4 Dataset for Deep Learning-Based Object Detection

For the object detection task, accurate annotation of the objects is crucial, since it forms the foundation for subsequent analysis. Utilizing the same dataset and tool referenced in the preceding section, objects within the images are carefully annotated by bounding boxes. Additionally, each bounding box is assigned a class label, indicating the type of object it represents (e.g., Ascaris egg, Necator egg, Trichuris egg, and Non-Egg object). Once all objects within the image are annotated, these are saved in a structured PASCAL VOC XML format, along with the corresponding image file. This annotated dataset serves as the training data for object detection algorithms, enabling them to learn to detect and localize objects within new, unseen images. A few examples of annotating the images in our dataset are shown in Figure 3-12.

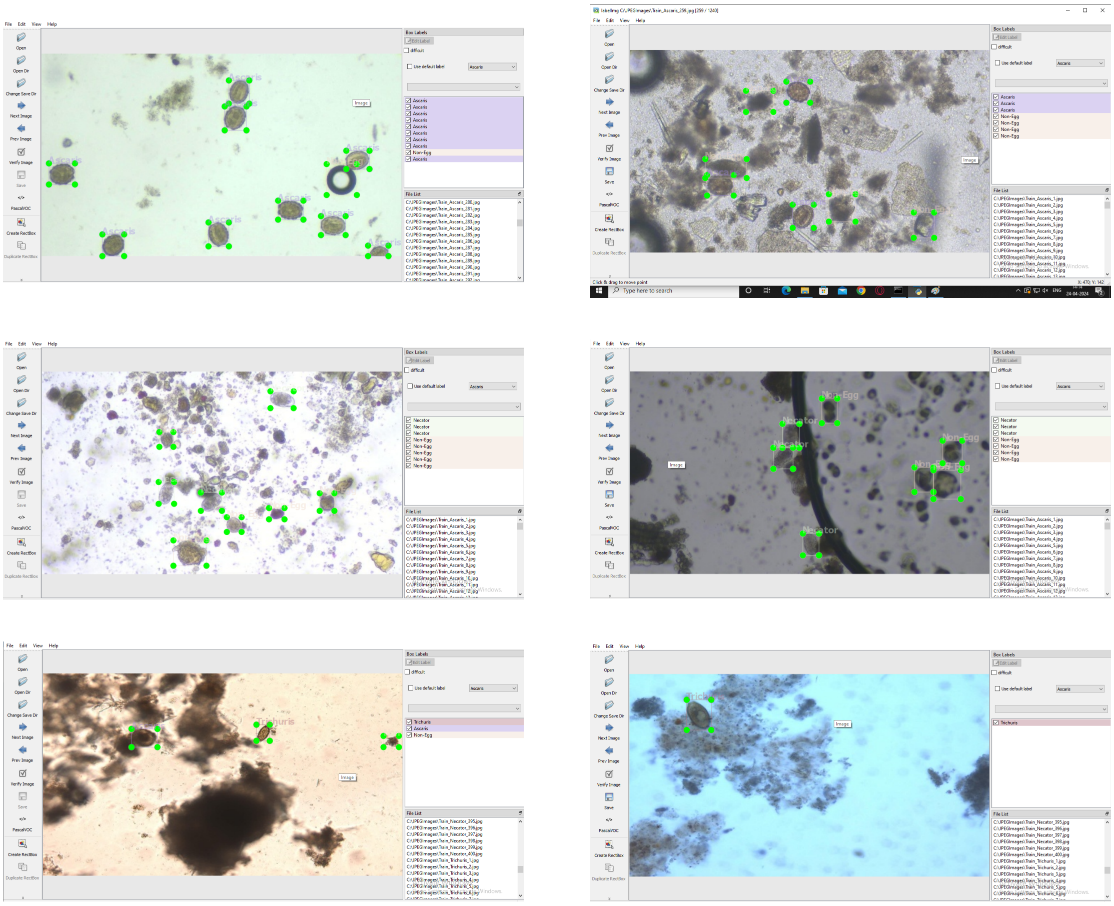

Figure 3-12: Example of annotating different classes of objects using bounding boxes in LabelImg

Each XML file corresponds to a single annotated image and follows a hierarchical structure. The file contains various elements for storing essential information about the image, such as the filename, image size, and image format. Each annotated object in the image is represented by a tag that contains information such as the object's name (class label), bounding box coordinates (xmin, ymin, xmax, ymax), and optionally, a difficult flag, which indicates whether the object is difficult to recognize. This tag helps to distinguish between objects that are relatively easy to detect and those that may pose difficulties for the object detection algorithms. In the COCO (Common Objects in Context) format, annotations are typically stored in a structured JSON (JavaScript Object Notation) file, organizing information into key components. The keys contain image metadata, including ID, filename, width, height, class labels, etc. The bounding box coordinates of each annotated object are usually included within the annotation object using the top-left corner's (x, y) coordinates, width, and height of the bounding box.

3.5 Conclusion

This chapter outlines the process of creating datasets for analyzing microscopic images of parasite eggs in fecal samples and identifying different types of parasite eggs. Key contributions and steps are summarized below:

- The work begins with acquiring images containing three types of parasite eggs: roundworm (*Ascaris lumbricoides*), hookworm (*Necator americanus* and *Ancylostoma duodenale*), and whipworm (*Trichuris trichiura*).
- Pre-processing steps, such as resizing and removing blurry images, are applied to optimize various methods in the automatic detection and identification of parasite eggs.
- Initially, datasets are created after the segmentation process. However, due to limitations in traditional segmentation, some parasite eggs are remain undetected. To address this, the overlooked eggs are manually segmented and included in the classification dataset.
- Parasite eggs and similar non-egg objects are annotated using the LabeImg tool to create ground truth masks for binary as well as multi-class semantic segmentation and object detection tasks.
- Data augmentation techniques are also employed to ensure a balanced dataset for training deep learning models.

Each section of this chapter plays a vital role in the preparation of datasets aiming at analyzing parasite eggs within microscopic images of fecal samples. Through detailed discussions, a solid foundation for advancing research in this field is established. By addressing challenges and ensuring dataset balance, we lay the groundwork for further exploration and advancement in parasite egg analysis in microscopic images.