Nurdle Count – A machine learning approach to nurdle classification and quantification - Year 1 Quarter 2 Report
(August 1st, 2024 – October 31st, 2024)

PI: Seneca Holland

October 23rd, 2024

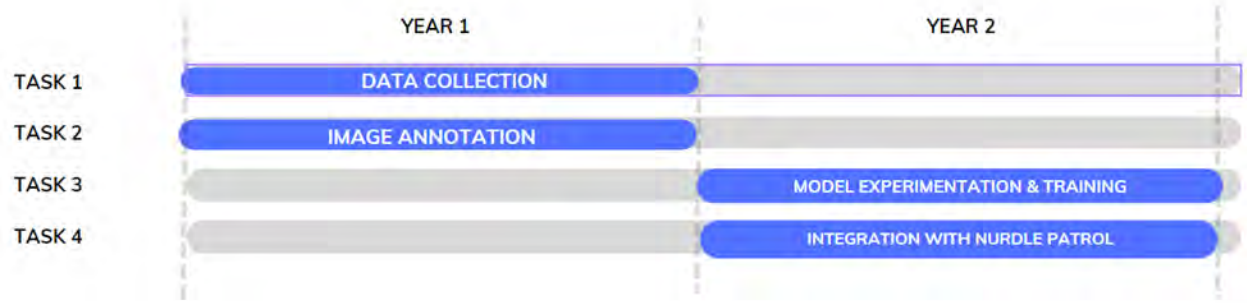
Administration:

The Nurdle Count – A machine learning approach to nurdle classification and quantification was approved for funding on January 8th, 2024, with a requested start date of May 1st, 2024.

Risks and Impacts:

None

Project Tasks:



1) Task 1 - Data collection:

a. **Collect training and test nurdle image data.**
b. **QA/QC collected nurdle image data.**
c. **Research and design AI training methods.**
d. **Develop a standard operating procedure (SOP) for capturing nurdle images.**

Task 1 – Subtasks 1a: Collect training and test nurdle image data

In Year 1, Quarter 1, the research team performed image capturing following the SOP developed for this purpose. Internally, using this SOP, 100 images were captured for Task 2—Image Annotation.

In Year 1, Quarter 2, this process was expanded with the help of middle school citizen scientists who are collecting images of nurdles in their classrooms and submitting them via the Nurdle Patrol Website using the QR code below.

Figure 1: Nurdle Count Image Submission QR Code

Students will continue to submit images in Year 1 Quarters 3 and 4 to expand the nurdle image library.

Task 1 – Subtasks 1b: QA/QC collected nurdle image data

In Year 1, Quarter 2, the research team executed the image capturing process as per the SOP specifically developed for this task. Utilizing this SOP, the team successfully captured 100 images designated for Task 2—Image Annotation. These images will be considered for initial experimental trials with various AI/ML model methodologies. This approach aims to assess the efficiency of the overall process and identify potential improvements. Through this experimentation, it is expected to uncover insights that will guide enhancements to the methodology and optimize the workflow.

Concurrently, Quality Control (QC) procedure is in the developmental phase, applying this dataset to test its efficacy. The research team is actively exploring a range of image analysis techniques to ascertain the quality and usability of the captured images. An SOP for the QC process is expected to be drafted based on the outcomes of these initial tests.
Further experimentation with the QC process will involve data collected from middle school citizen scientists.

Task 1 – Subtasks 1c: Research and design AI training methods

This task was completed in Year 1, Quarter 1.

Task 1 – Subtask 1d: Develop a standard operating procedure (SOP) for capturing nurdle images.

In Year 1, Quarter 1, and in preparation for collecting training and testing the Nurdle image data, the Nurdle Count team first developed two Standard Operating Procedures (SOPs). After extensive review, two Standard Operating Procedures (SOPs) were created, each tailored to different audiences: internal and external. The internal SOP is designed for use by the research team, while the external SOP is intended for 8th-grade students. Although both SOPs share similar content and workflow, the external SOP is written in language that is accessible and understandable at an 8th-grade reading level.

In Year 1, Quarter 2, project personnel developed a series of three videos detailing the nurdle capture process and made them available via YouTube for a wider audience. To ensure accessibility to a broader audience, YouTube settings enabled these videos to be viewed by kids, and closed captioning was enabled.

There videos are:

Part 1 – Setting up Nurdles in Nurdle Count: https://youtu.be/99pSZEfB37g
Part 2 – Capturing Pictures for Nurdle Count: https://youtu.be/rLRbYLwNVVg
Part 3 - Nurdle Count Image Submission: https://youtu.be/TyTd6OBw9HA

These videos were shared with the three middle school educators in coastal Texas communities who will be collaborating with their classes on Nurdle Count on the following tasks:

- Collecting nurdle image data via the external SOP
- Training the Nurdle Count AI by classifying nurdles
- Collect feedback and improve the Nurdle Count application

Teachers in these classrooms were also provided with a Nurdle Count Worksheet to aid in the counting and classification of nurdles. This tool familiarizes students with the shape, size, and color of nurdles which will be helpful when students move to Step 2 of their tasks - Training the Nurdle Count AI by classifying nurdles.

Additionally, teacher kits were developed and provided to each classroom. These kits will assist students in collecting nurdle image data and submitting it to the Nurdle Count Team.

These kits included the following items:

- 20 Nurdle Count Worksheets (also provided digitally)
- 20 Nurdle Count Instructions (also provided digitally)
- 20 Petri Dishes (for sorting and counting nurdles)
- 10 metal tweezers (for sorting and counting nurdles)

- 20 rulers for positioning the camera 8 inches away from the petri dish
- 10 Felt Rounds (for providing contrast in the Petri dishes when taking pictures of nurdles)
- 14-17 Vials Containing Nurdles (100-200 nurdles in each vial)
- Nurdle Patrol Postcards (for those interested in more info on the citizen science aspect of this research)
- Nurdle Patrol Business Cards (for those interested in more info on the citizen science aspect of this research)

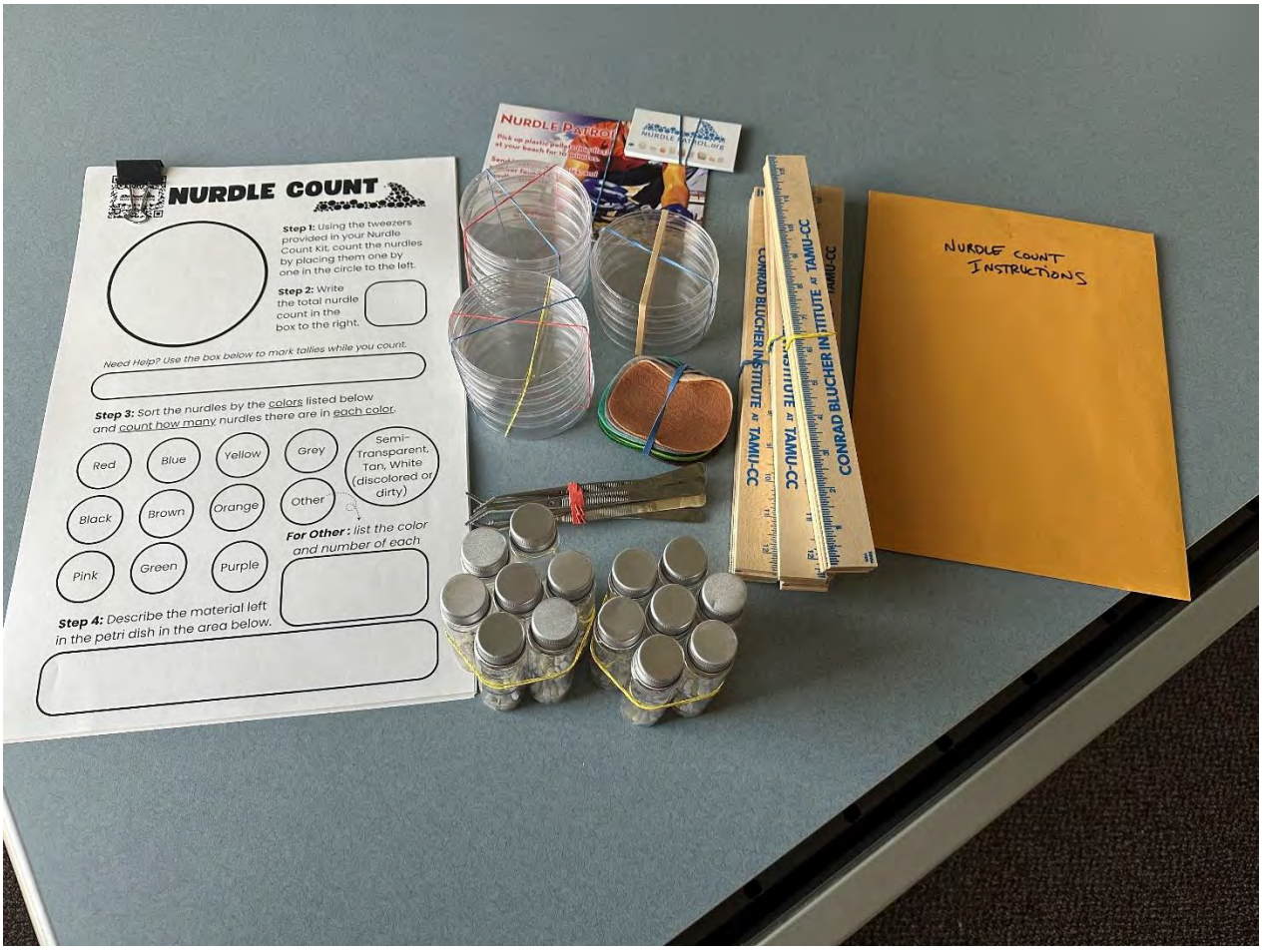


Figure 2: Nurdle Count Classroom Kit

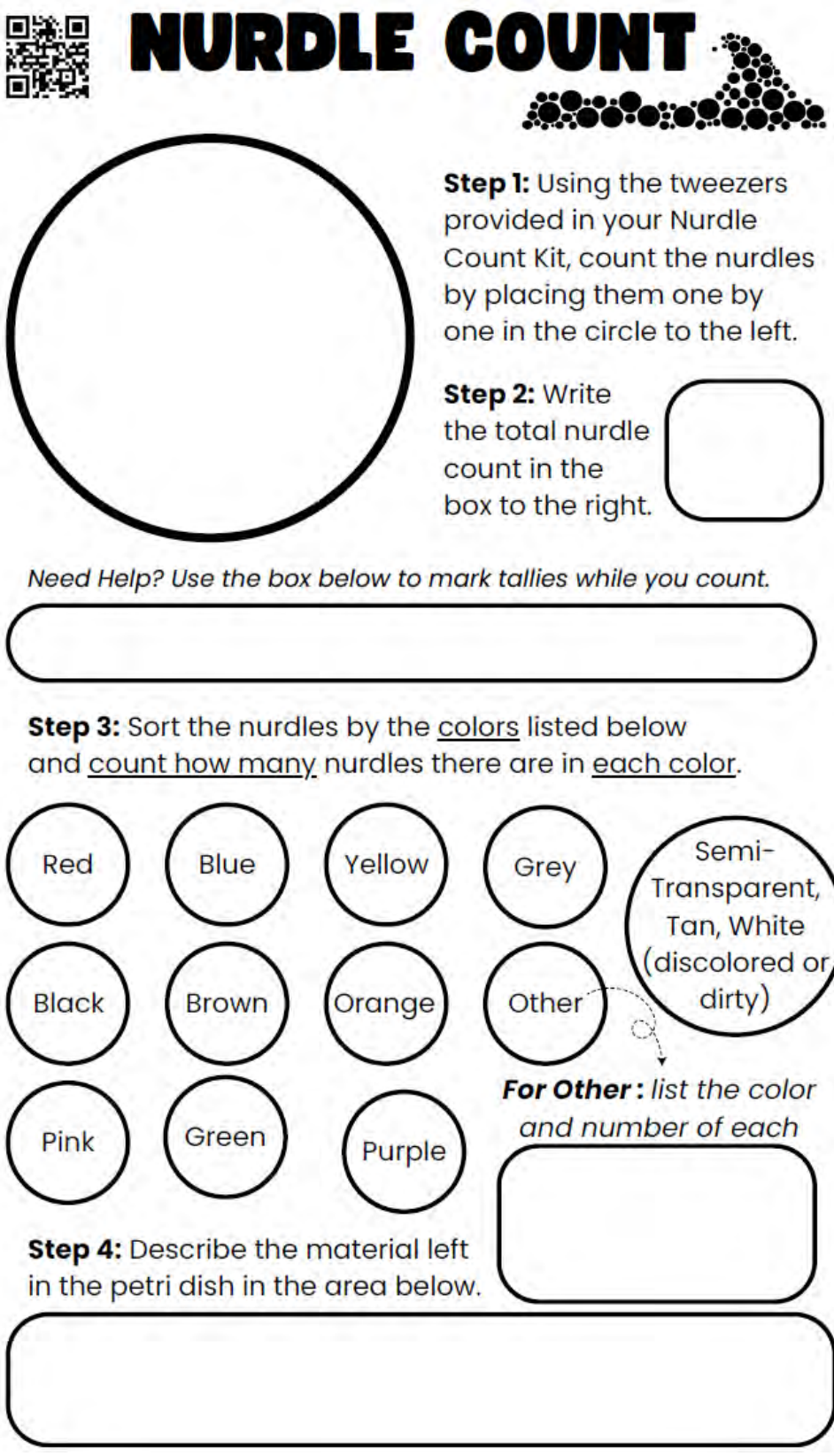


Figure 3: Nurdle Count Worksheet

Task 2 – Image Annotation

In the first quarter of the Nurdle Count project, the research team began working on data annotation, which is part of the next quarter's objectives. They performed image annotation on the CVAT (Computer Vision Annotation Tool) platform and experimented with exporting the annotation data for manipulation using Python, the programming language chosen to develop the AI/ML model for Nurdle Count.

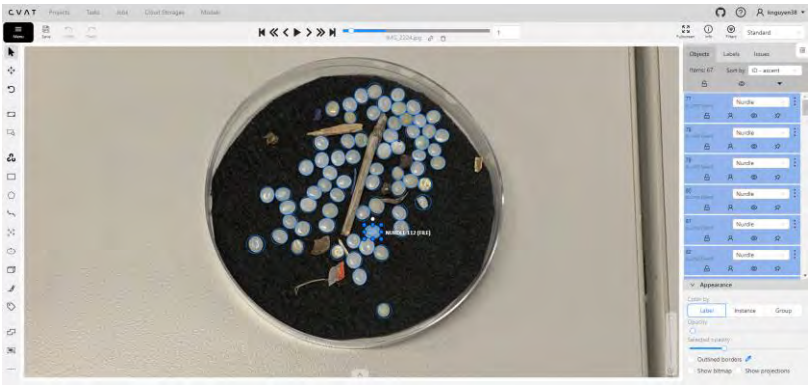


Figure 4: Annotated Nurdle Image

It is our hope that, in preparing the annotation data for use with Python and machine learning libraries, the collected nurdle image data can be used to train and develop an accurate, efficient, and useful model for identifying and quantifying nurdles in images taken by everyday citizen scientists.

In support of this task, the Nurdle Count Research Team has deployed a self-hosted CVAT platform. This platform provides a robust and efficient solution for the annotation process, giving users a friendly interface to annotate nurdles. The platform helps organize and manage the annotated images, making it easier for the team to analyze the data and draw conclusions about nurdle pollution.

By using CVAT, the Nurdle Count Research Team can more efficiently and accurately process large numbers of images, leading to better insights and more effective strategies to address the Nurdle Count problem.

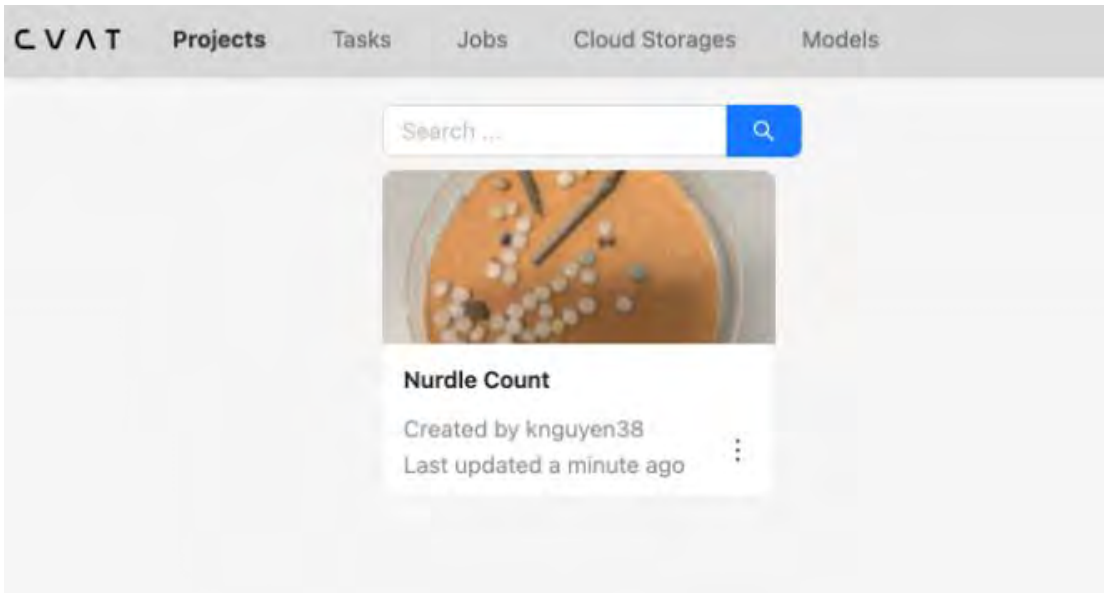


Figure 5: CVAT Platform for Nurdle Count Annotation

In Year 1 Quarter 2, with the self-hosted CVAT platform, the research team has completed annotation for fifty images thus far.

Figure 6: Example of An Annotated Image

The annotation process was enhanced by integrating image analysis techniques to maintain consistent image quality. These methods show potential for effectively identifying nurdles based on predefined characteristics such as shape, texture, and size.

A band-pass filtering technique has been applied to enhance specific features of the images while suppressing noise and irrelevant details. It helps in isolating nurdles from complex backgrounds, improving the visibility of smaller and translucent nurdles.

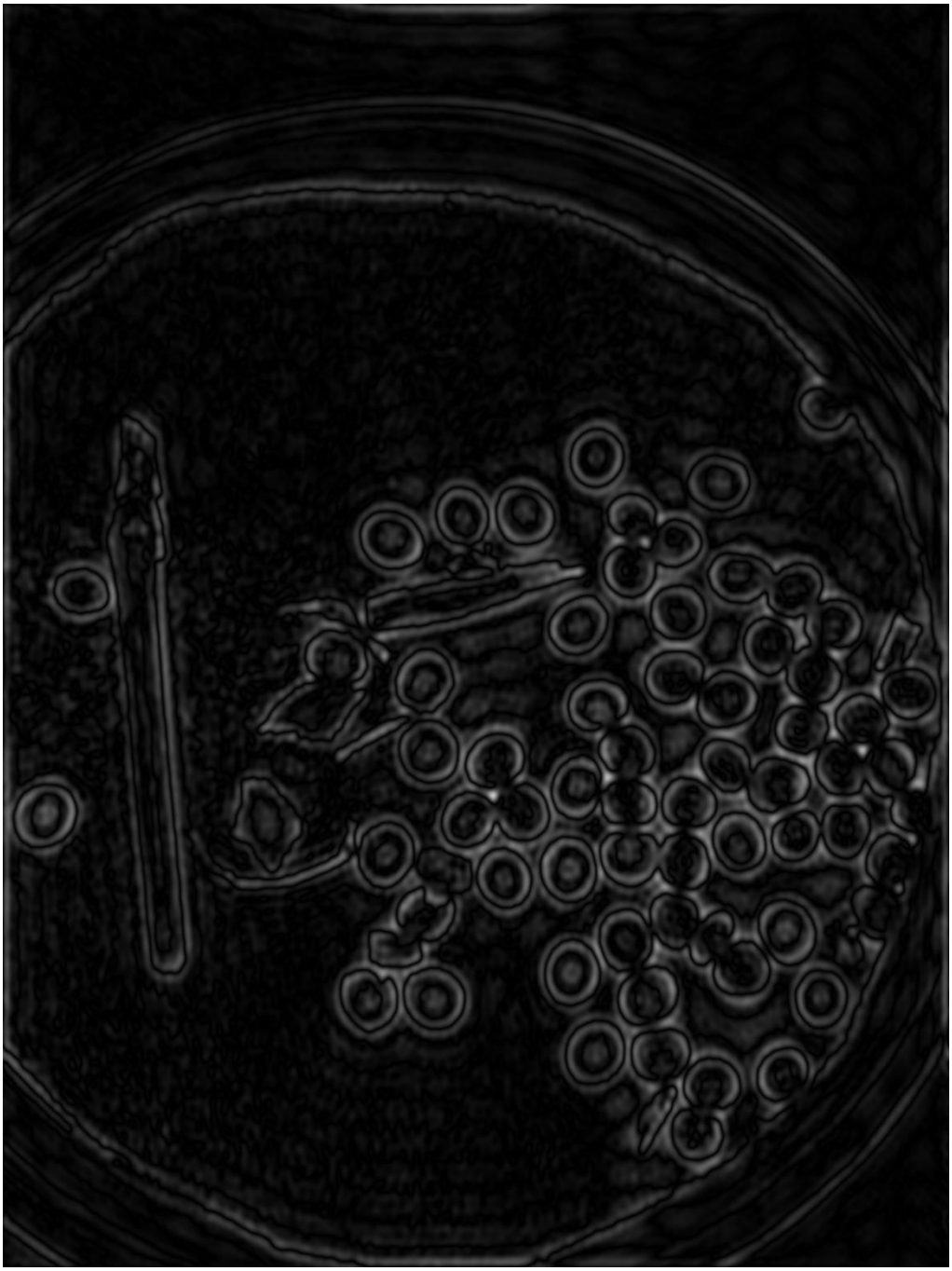


Figure 7: Example of An Image with Band-Pass Filtering

By employing edge detection algorithms, such as the Sobel operator, the team has enhanced the contours of nurdles. This facilitates more precise delineation and segmentation in subsequent processing steps.

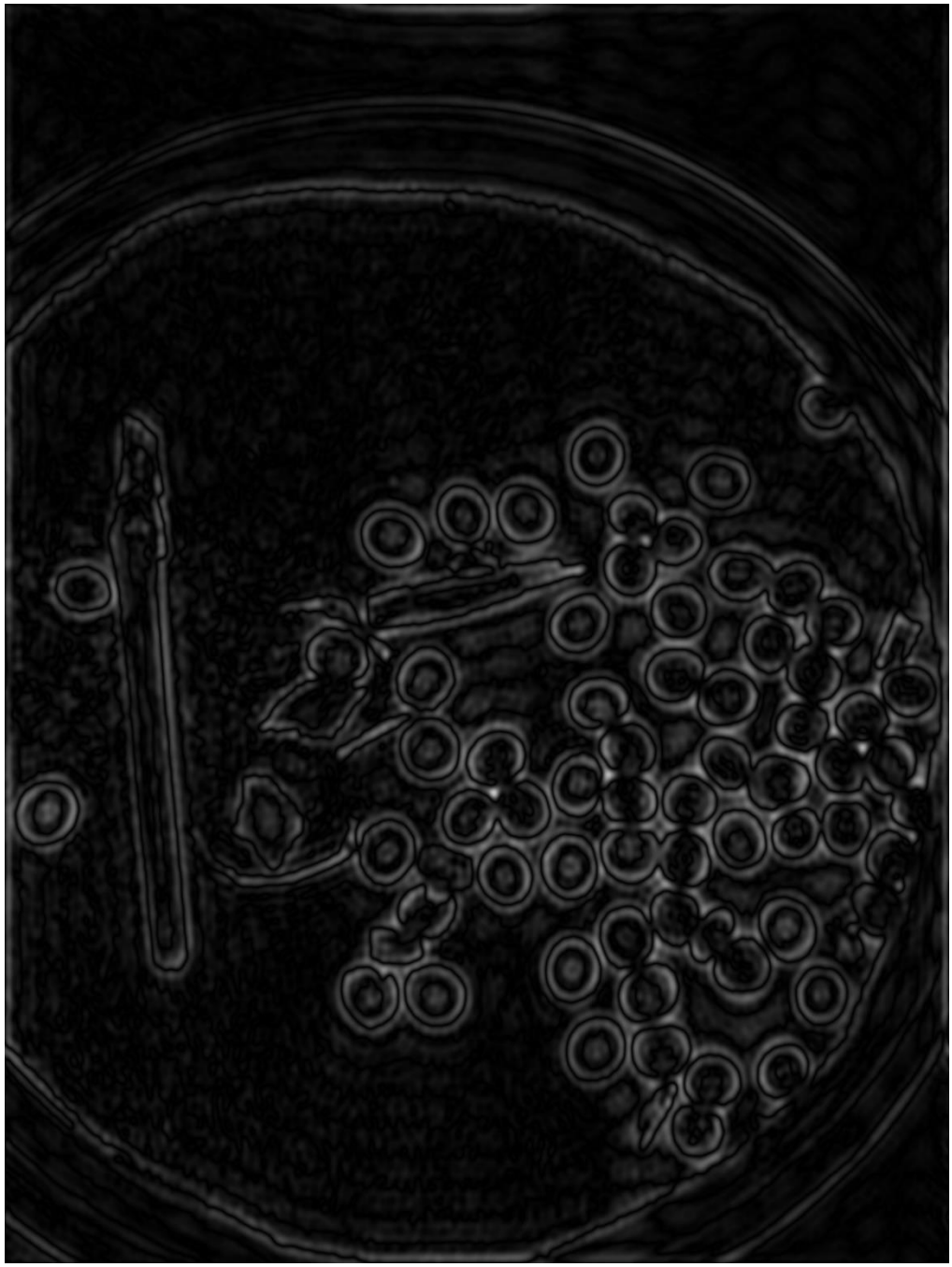


Figure 8: Example of Edge Detection

Adaptive thresholding techniques were used to refine the image processing further. This approach helps in differentiating nurdles from the background by converting grayscale images into binary images, making it easier for the annotation tool to recognize and mark the nurdles. This is potentially an efficient way to provide insights into how different types of backgrounds affect the thresholding process, helping in developing more robust algorithms that can adapt to changes in environmental conditions and background textures.

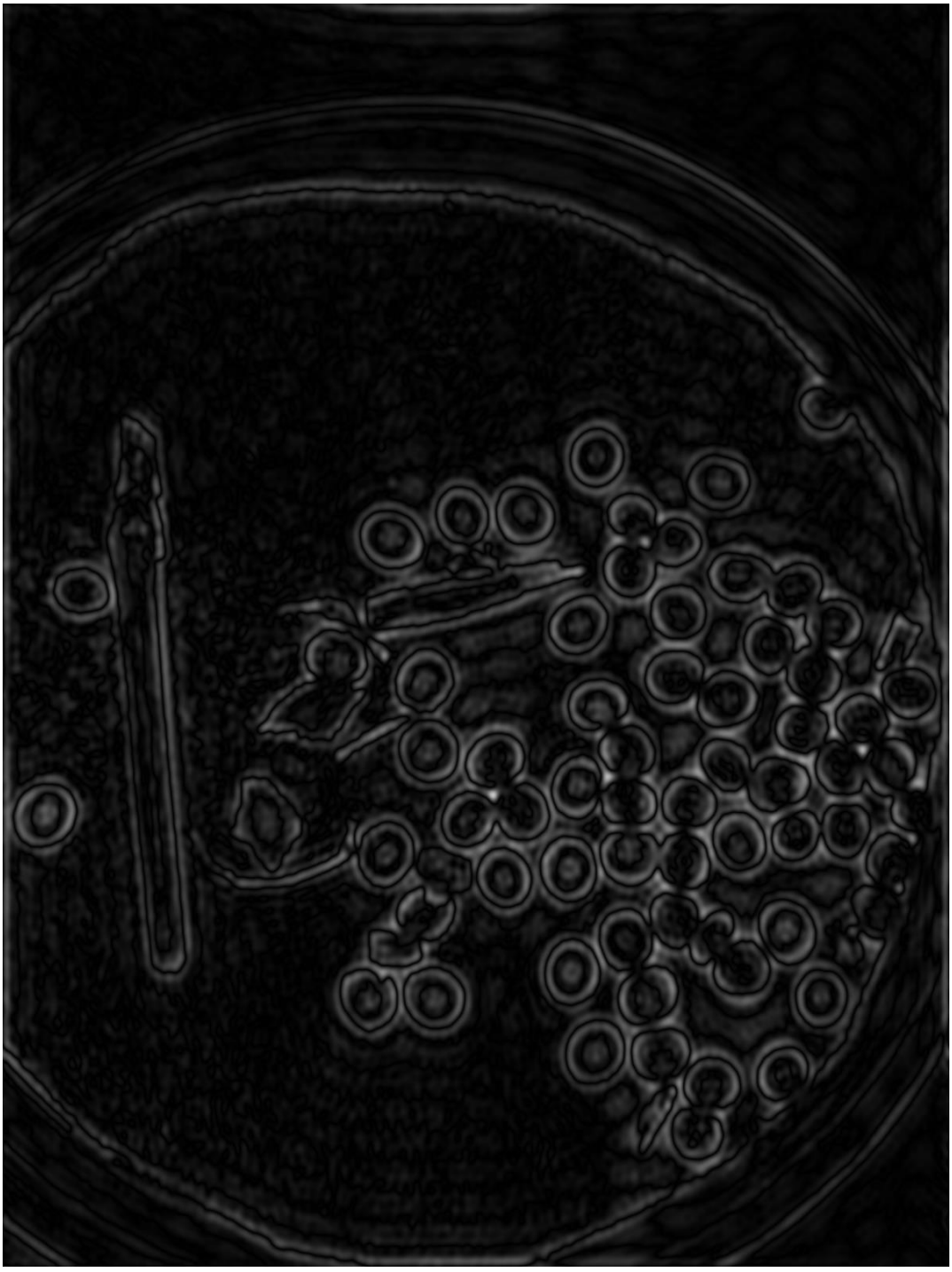


Figure 9: Example of Thresholding

To advance the project effectively, it is crucial to expand the datasets continuously. By gathering a more extensive array of images, the robustness and accuracy of the models for the next step can be significantly enhanced, allowing for a broader understanding of environmental variations and nurdle characteristics.

Simultaneously, following the extension of the datasets, ongoing annotation of the images is essential. Continuous annotation ensures that the data remains relevant and reflective of the diverse conditions under which nurdles are found.

Moreover, improving the procedures for data quality assurance and quality control (QA/QC) is imperative. By developing more rigorous QA/QC protocols, the research team can ensure the reliability and validity of the data used in their analyses.

Task 3 - Model experimentation and training: ***to be completed in year 2***

a. Train Nurdle Count AI.
b. Collect feedback and improve the AI model.

Task 4 - Integration with Nurdle Patrol: ***to be completed in year 2***

a. Implement the Nurdle Count feature on NurdlePatrol.org.
b. Implement the Nurdle Count feature in the Nurdle Patrol Apple iOS mobile application.
c. Implement the Nurdle Count feature in the Nurdle Patrol Android mobile application.
d. Publish AI model to the public.

Summary:

We have made substantial progress in Year 1, Quarter 2 with Tasks 1 and 2, with a marked increase in nurdle image data collection and foundational steps in image annotation, including the successful deployment of a self-hosted CVAT platform for streamlined annotation. In the next quarter, the focus will intensify on expanding the nurdle image dataset through additional contributions from middle school citizen scientists, which will enhance the diversity and environmental representation of the dataset. Concurrently, the project team will prioritize the development of a more rigorous QA/QC SOP for annotation, implementing advanced filtering and thresholding techniques to improve data reliability and optimize the model's accuracy in subsequent training phases. This approach aims to ensure the model is trained on high-quality, consistently annotated images, laying a stronger foundation for robust model development.

Obstacles: None