Nurdle Count – A machine learning approach to nurdle classification and quantification - Year 2 Quarter 3 Report

PI: Seneca Holland

February 6th, 2026

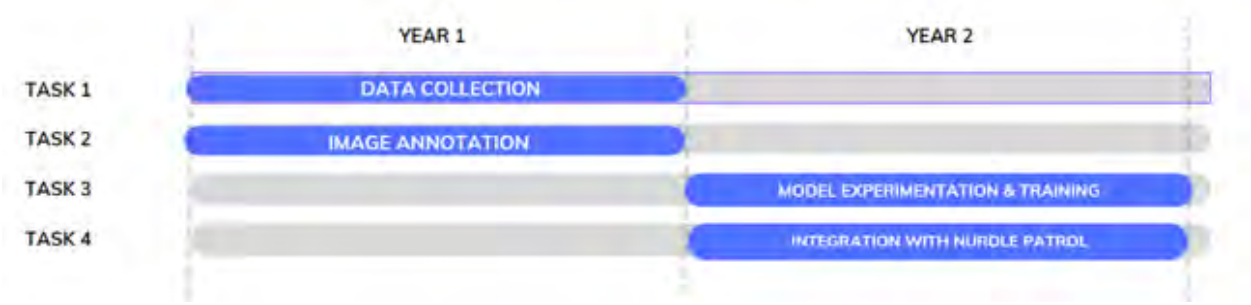
Administration:

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 which is 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.

In Year 1, Quarter 3, this process was expanded with the help of undergraduate students who collected images of nurdles in class and submitted them via the Nurdle Patrol Website using the QR code.

In Year 1, Quarter 4, this process was expanded with the help of several undergraduate students who added 700 images following strict collection parameters to the Nurdle Patrol Website using the QR codes (Figure 1).

Figure 1: Nurdle Count Image Submission QR Code

Task 1 – Subtask 1a was completed in Year 1, Quarter 4.

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

In Year 1, Quarter 4, Subtasks 1b (QA/QC of collected images) and 1d (development of the image capture SOP) became closely intertwined, forming an iterative and interdependent workflow. The QA/QC process required a finalized SOP to ensure consistent image quality and metadata, while the SOP's development relied on a fully functional Nurdle Swipe interface to validate and classify images. To support this integration, the Nurdle Swipe tool was upgraded to improve usability and streamline the review process. Notably, text-based buttons such as "swipe right" and "swipe left" were replaced with intuitive visual symbols to reduce user confusion and enhance accessibility (Figure 2).

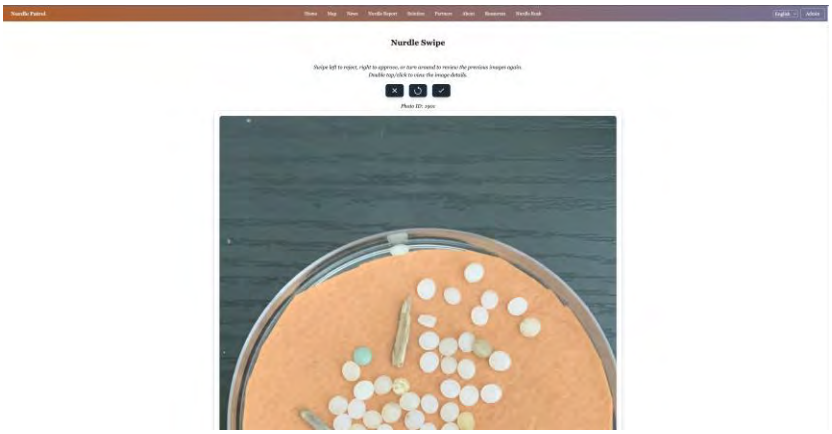


Figure 2: Nurdle Image

Additionally, a standardized list of disqualification reasons was created based on the most frequent issues identified in past image reviews. Validators can now select from this predefined list rather than entering reasons manually, streamlining the QA/QC process and promoting greater consistency (Figure 3).

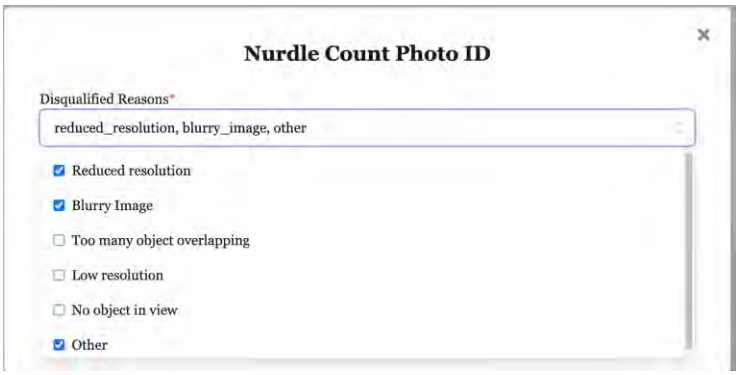


Figure 3: Nurdle Count Photo ID Disqualification Reasons

There is another option that can be used to point to disqualification reasons not yet included in the list, allowing validators to flag new issues. These entries will inform future updates by helping the research team expand and refine the standardized reason list (Figure 4).

Figure 4: Nurdle Count Photo ID Disqualification Reasons Comment

Researchers assessed each image based on clarity, resolution, object visibility, and the absence of obstructions or excessive overlaps. Specifically, a total of 638 images were reviewed through this effort, resulting in 545 images being marked as qualified and 93 as disqualified. Disqualified images were excluded due to reasons such as “reduced resolution” (60 images), “blurry image” (8 images), “too many objects overlapping” (8 images), and “no visible object in view” (19 images), with some images falling into multiple disqualification categories. The qualified set of images will be used for training AI models for nurdle identification. Figure xxx shows the Nurdle Swipe webpage interface, where two images were classed as qualified and disqualified, respectively. Task 1 – Subtask 1b was completed in Year 1.

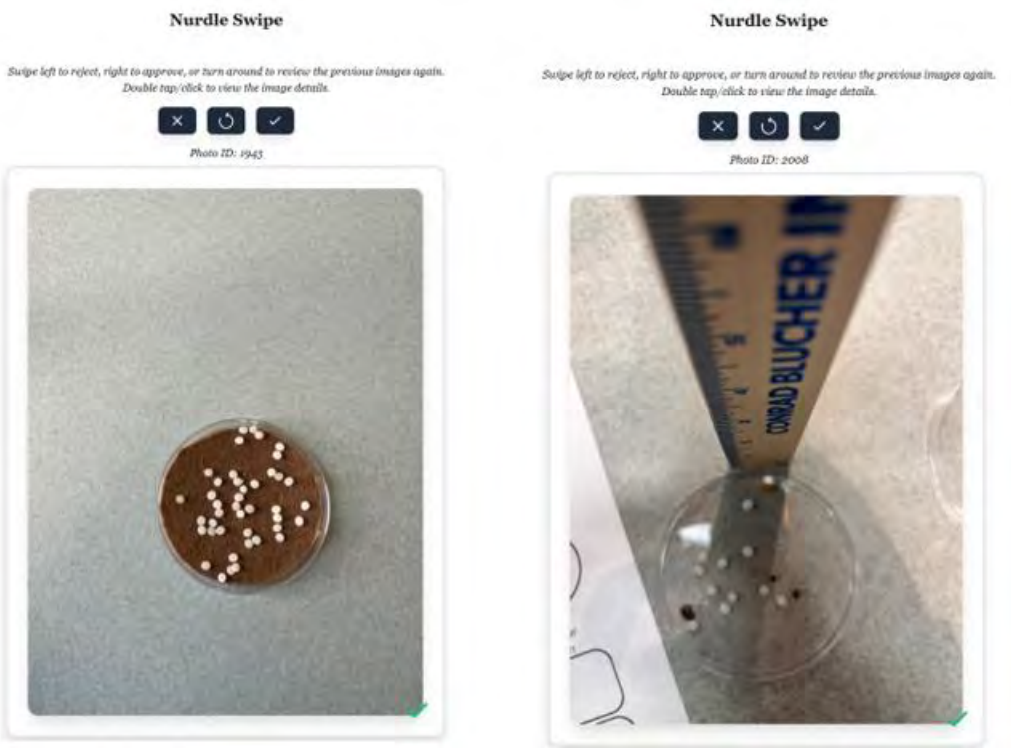


Figure 5: The Nurdle Swipe webpage interface. Left: a qualified nurdle image. Right: a disqualified nurdle image due to blurriness.

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 an 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.

These 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

In Year 1, Quarter 3, these videos and materials were leveraged to collect nurdle image data, collect feedback, and improve the Nurdle Count application. This subtask was completed in Year 1, Quarter 3.

For additional information about Year 1, Quarter 4, please see the Subtask 1b section above.

Task 2 – Image Annotation

In Year 1, Quarter 3, to enhance the quality assurance (QA) and quality control (QC) of images collected for training images for Nurdle Count, the Nurdle Swipe feature was developed and successfully integrated into Nurdle Patrol. This process is detailed in Task 1 above.

In Year 1, Quarter 4, we worked to define the preliminary model for detecting support for fast and automatic annotation. Several ML/AI models have been experimented on for nurdle detection. In addition, these model results can be counted as the preliminary results in the early phases and are valued on the way to detect and count nurdles accurately.

Several YOLO-family models, including YOLOv5n, YOLOv8n, and the latest YOLO11n, have been experimented with the current annotated set of images. The models were evaluated based on several metrics, as described in Table 1 below.

Table 1: Model Metrics

Metric	Description
Precision	Of all the objects the model says it found, what fraction are real objects? Higher is better.
Recall	Of all the real objects in the image, what fraction did the model actually find? Higher is better.
mAP@0.5	A combined score (mean Average Precision) that rewards finding objects with at least 50% overlap accuracy. Think of it as an overall “accuracy” at a loose overlap threshold. Higher is better.
mAP@0.5-0.95	Similar to mAP@0.5 but averaged over a range of tighter overlap requirements (from 50% up to 95%). This penalizes sloppy bounding boxes more heavily. Higher is better.

Precision, which indicates the proportion of reported detections that are actual nurdles rather than false alarms, was found to be similar across all three models at around 83%, so false alarms are seldom raised. Recall, which measures the proportion of real nurdles in an image that are detected, was highest for YOLO11n at 78%, compared with 60% for YOLOv5n and 69% for YOLOv8n, indicating that substantially fewer pellets were missed by YOLO11n.

Mean Average Precision at a 50% overlap threshold (mAP@0.5), which combines precision and recall into a single accuracy score under a relatively loose matching requirement between predicted and true nurdle locations, was highest for YOLO11n at 0.823—over ten points above YOLOv5n's 0.732. When a tighter matching requirement was imposed (averaging overlap thresholds from 50% to 95%, known as mAP@0.5–0.95), the improvement offered by YOLO11n became even more pronounced: a score of 0.466 was achieved, compared with 0.336 for YOLOv5n and 0.360 for YOLOv8n. These results indicate that not only are more nurdles detected by YOLO11n, but bounding boxes are also drawn around them more precisely.

In practical applications, the use of YOLO11n can result in far fewer pellets are missed. This combination is critical when undetected nurdles can contribute to pollution or signal production defects, and when false alerts can lead to wasted time and resources. Overall, YOLO11n is demonstrated to provide the best balance of thoroughness and reliability for accurate nurdle detection (Table 2).

Table 2: YOLO Results

Model	Precision	Recall	mAP@0.5	mAP@0.5-0.95
YOLOv5n	**0.83**	0.596	0.732	0.336
YOLOv8n	0.815	0.685	0.777	0.36
YOLO11n	0.828	**0.784**	**0.823**	**0.466**

In Task 3 of the project, the research team will continue to work on the automatic annotation workflow, integrate the model for automatic annotation, and experiment with the workflow on the new batch of nurdle images.

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

In preparation for Task 3, the research team advanced efforts to curate a high-quality image dataset through the Nurdle Swipe tool, a web-based platform hosted on the Nurdle Patrol website (https://nurdlepatrol.org/app/nurdle-swipe). The tool was developed to support AI model training by systematically reviewing nurdle images collected in accordance with the established image collection SOP. Using a swipe interface, researchers approved images that met quality standards (swipe right) or disqualified those that did not (swipe left).

Each image was evaluated for clarity, resolution, object visibility, and freedom from obstructions or excessive overlap. Through this process, 638 images were reviewed, of which 545 were classified as qualified and 93 were disqualified. Disqualifications were attributed to reduced resolution (60 images), blurry capture (8 images), excessive object overlap (8 images), or absence of visible objects (19 images), with some images falling into multiple categories. The resulting set of 545 qualified images will serve as a training dataset for AI-based nurdle identification models, ensuring that only rigorously vetted imagery is used to improve detection accuracy. Figure 5 illustrates the Nurdle Swipe interface, displaying examples of qualified and disqualified images.

Looking ahead, Task 3 will focus on model experimentation and training using this expanded, high-quality dataset. The research team will conduct comparative testing across multiple computer vision architectures to evaluate precision, recall, and mean average precision (mAP) metrics. The top-performing model will then be selected for iterative training and refinement. Feedback loops from annotation and quality control processes will be integrated to further improve performance. By the conclusion of Year 2, the trained Nurdle Count AI will be ready for deployment into the Nurdle Patrol website and mobile applications, providing a scalable solution for automated nurdle detection.

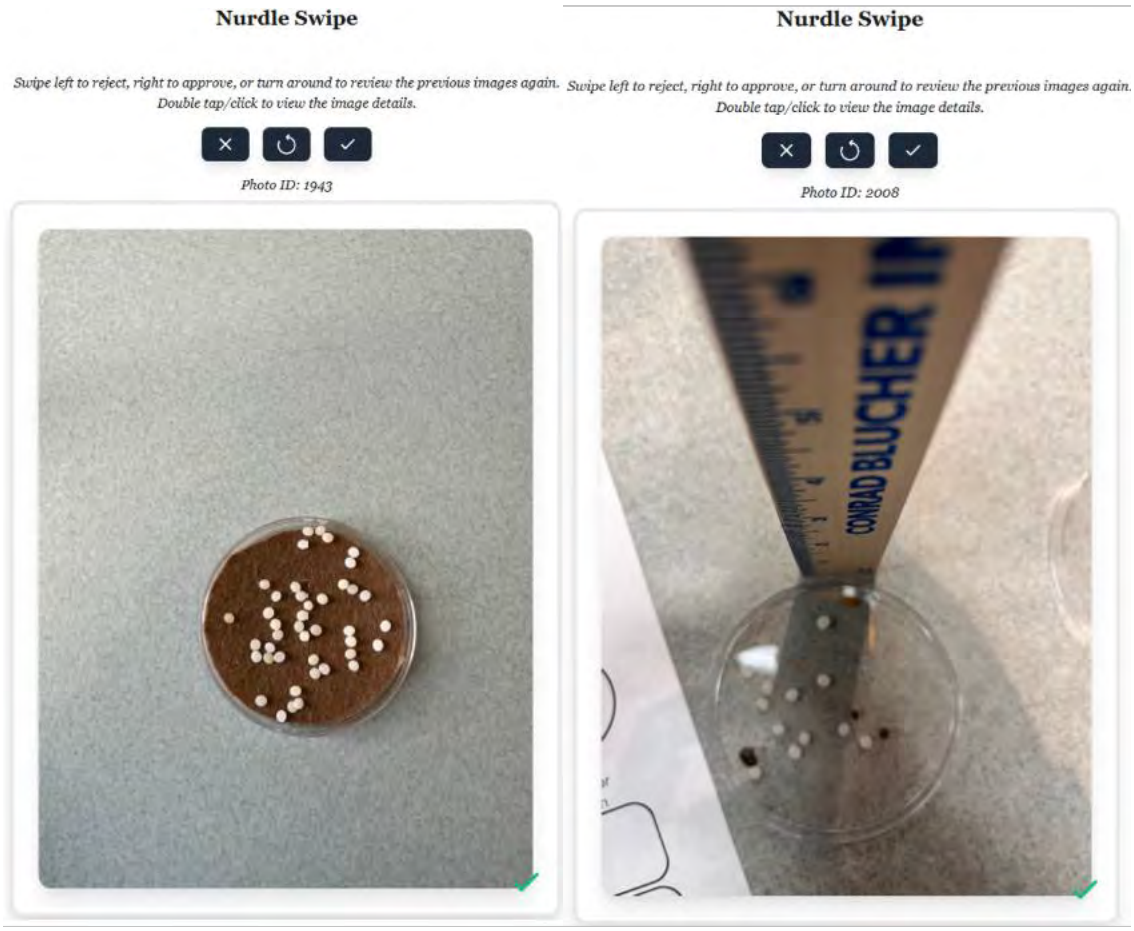


Figure 6: Nurdle Swipe Images

In Year 2, Quarter 2, work continued on the YOLO segmentation model experiment.

In the last quarter, the research team experimented with YOLO segmentation model. This approach required a different way of annotating in which the boxes fit more to the shape of the nurdles. A deployment of Facebook Segmentation Anything Model has been done to support the annotation process.

Figure 7: Example of an annotated image with SAM for YOLO segmentation model

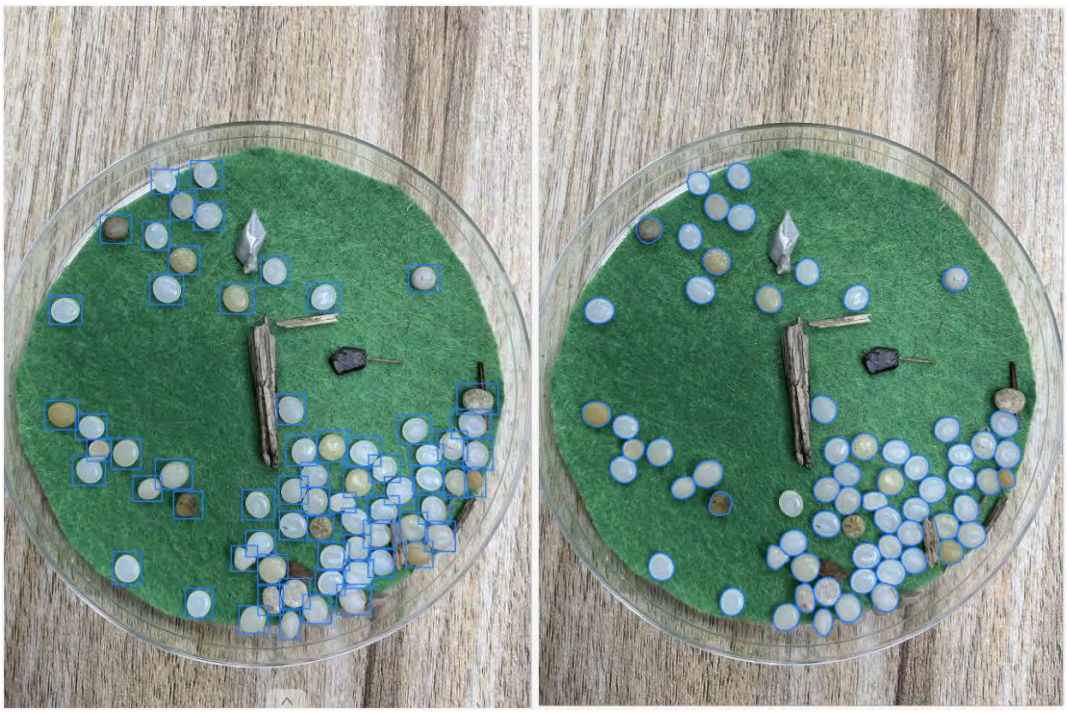


Figure 8: Original annotation for YOLO-based model (left) and annotation for YOLO segmentation model (right)

To assess the performance of the new YOLOv8n segmentation, it will be compared to the YOLO11n, the best model in the previous reports.

Table 3: Performance Metrics Comparison

Metric	YOLOv8n-seg	YOLOv11n	Difference
Overall Accuracy (mAP@0.5)	99.3%	82.3%	-17.0%
Strict Accuracy (mAP@0.5-0.95)	~90%	~40%	-50%
Correct Detection Rate (Precision)	~100%	~85-95%	-5 to -15%
Finding All Nurdles (Recall)	~100%	~87%	-13%
Balanced Score (F1)	0.98	0.81	-0.17
Confidence Needed for Reliability	57.3%	95.5%	+38.2%

The performance comparison reveals substantial differences between the two models. YOLOv8n with segmentation achieved an exceptional accuracy of 99.3%, meaning it correctly identifies nurdles 99 times out of 100. YOLOv11n reached only 82.3%, representing a significant 17 percentage point gap. This difference becomes even more pronounced when examining strict accuracy measures that require very precise box placement, where YOLOv8n achieves about 90% compared to YOLOv11n's 40%. When examining how correctly the models identify nurdles versus falsely detecting non-nurdles, YOLOv8n maintains nearly perfect accuracy at approximately 100%, while YOLOv11n ranges between 85-95%. The ability to find all nurdles similarly favors YOLOv8n at approximately 100% versus YOLOv11n's 87%.

Perhaps most critically for practical deployment, YOLOv8n achieves 100% accurate identifications when it's at least 57% confident, while YOLOv11n requires an extremely strict 95.5% confidence threshold. This means in real-world use, YOLOv8n can reliably detect nurdles with moderate confidence, while YOLOv11n must be almost completely certain before its detections can be trusted. This makes YOLOv11n far less practical, as most legitimate nurdle detections fall below this very high threshold, causing the model to miss many real nurdles just to avoid making errors.

Table 3: YOLOv8n Segmentation Results:

What's Actually There	Model Says "Nurdle"	Model Says "Background"
Nurdle	211 (96.8%)	7 (3.2%)
Background	2 (0.9%)	Perfect

Table 4: YOLOv11n Detection Results:

What's Actually There	Model Says "Nurdle"	Model Says "Background"
Nurdle	816 (80.9%)	193 (19.1%)
Background	0 (0%)	Perfect

The confusion matrix shows exactly how each model performs in different situations. For YOLOv8n segmentation, the model correctly identified 211 nurdles out of 218 presents, missing only 7 nurdles. This means it has a 96.8% success rate at finding nurdles that are there. The model also produced only 2 false alarms where it thought it saw a nurdle when there wasn't one, demonstrating excellent ability to distinguish between nurdles and background objects.

YOLOv11n detection tells a different story. While it correctly identified 816 nurdles, it missed 193 nurdles that were present. This means it has a 19.1% failure rate - nearly one in five nurdles goes undetected. Interestingly, YOLOv11n never produces false alarms, achieving a perfect 0% false positive rate. However, this "perfection" comes at a severe cost: the model is so conservative that it refuses to make a detection unless it's extremely certain, which causes it to miss many real nurdles. Missing one in five nurdles is 27 times worse than YOLOv8n's miss rate.

Detection Quality Analysis

When looking at the actual images with detection boxes drawn on them, the differences become visually apparent. YOLOv8n segmentation produces tight, well-fitted boxes that precisely outline individual nurdles. Even when multiple nurdles are clustered close together, the model successfully identifies each individual pellet separately. The detections remain accurate across various challenging conditions: different lighting (bright, shadowy, or dim), different angles and rotations of the objects, and different background materials (green surfaces, brown media, white plates, or wooden surfaces). The confidence scores the model assigns are generally high for correct detections, indicating it "knows" when it has found a real nurdle.

YOLOv11n detection exhibits several quality problems in its predictions. In many images, multiple overlapping boxes are detected on the same nurdle, showing that the model is detecting the same object several times instead of recognizing it as a single item. The boxes are generally less precisely fitted around the actual nurdle boundaries, often being too large or poorly positioned. The model particularly struggles when nurdles are packed closely together - it either groups multiple nurdles into one detection or misses individual nurdles within crowded scenes. The confidence scores vary widely from 30% to 90%, and because the model needs 95.5% confidence to be truly reliable, it ends up missing many valid nurdles that fall below this very strict threshold.

Precision-Recall Characteristics

The precision-recall relationship tells us how well a model can balance between being accurate and being thorough. YOLOv8n segmentation maintains over 95% accuracy in its identifications even while finding 95% of all nurdles present. Only when trying to find nearly every single nurdle (above 95% recall) does its accuracy begin to drop. This creates a nearly ideal performance curve where the model can be both accurate and thorough simultaneously.

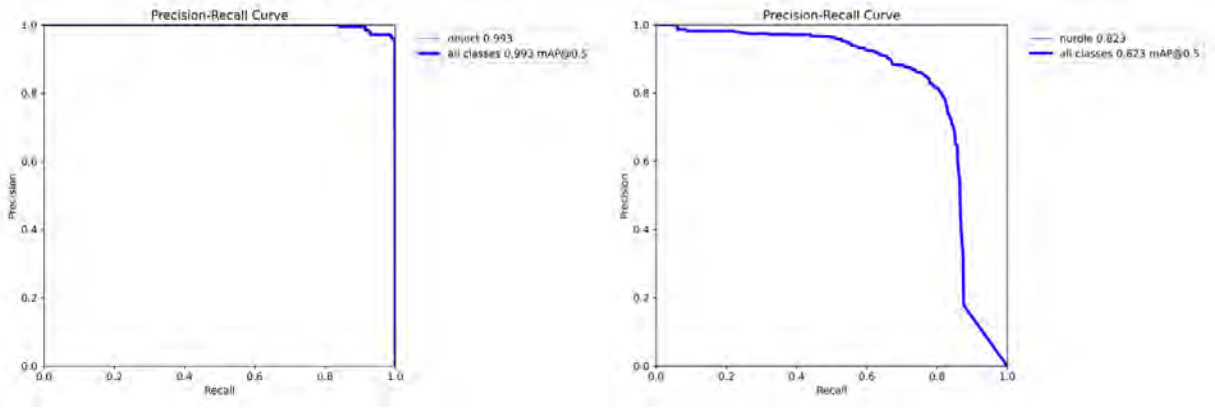


Figure 9*: YOLOv8n segmentation (left) achieves 0.993 mAP@0.5 with a nearly rectangular curve, maintaining over 95% precision while finding 95% of nurdles. YOLOv11n detection (right) achieves 0.823 mAP@0.5 with earlier precision degradation, dropping to 85% precision at 80% recall. The area under the curve difference (0.993 vs 0.823) represents a 17-point performance gap.*

YOLOv11n detection shows a much less favorable balance. It can maintain 98-100% identification accuracy only when it's being very selective and finding less than 40% of the nurdles present. As the research team lower the selectivity to find more nurdles, the accuracy degrades much earlier and more severely than YOLOv8n. By the time it finds 80% of nurdles, its identification accuracy has dropped to about 85%. This curved relationship indicates a harsh tradeoff where attempting to detect more nurdles rapidly compromises how accurately it identifies them. This makes it very difficult to find a setting that gives both reasonable coverage and reliable identifications.

Confidence Calibration Analysis

Confidence calibration determines how much we can trust a model's stated confidence in its predictions. Think of it like a weather forecast - if the forecast says 70% chance of rain, it should rain about 70% of the time for the forecast to be well-calibrated. YOLOv8n segmentation demonstrates excellent calibration. When the research team use a confidence threshold of 57.3% (meaning we only trust detections where the model is at least 57.3% certain), we get 100% accurate identifications. The model works well even at very low confidence thresholds, and its confidence scores align well with actual accuracy, making it straightforward to choose appropriate settings for different needs.

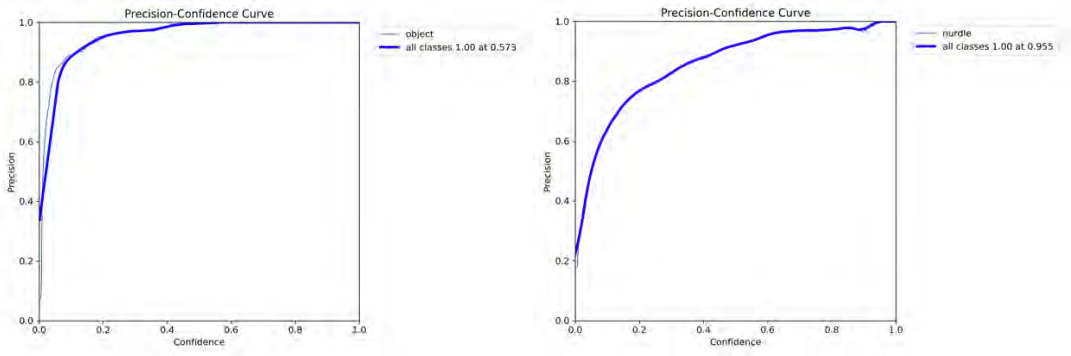


Figure 10: YOLOv8n segmentation (left) achieves 100% precision (perfect accuracy) at 57.3% confidence, with smooth calibration showing reliability at moderate thresholds. YOLOv11n detection (right) requires 95.5% confidence for 100% precision, making it impractical as most detections fall below this strict threshold. The 38-point difference in required confidence (95.5% vs 57.3%) represents a critical disadvantage for real-world deployment.

YOLOv11n detection has severe calibration problems. To get 100% accurate identifications, we must use an extremely strict confidence threshold of 95.5%, meaning we can only trust detections where the model is 95.5% certain. This is impractical because most real nurdle detections fall below this threshold, so we'd miss most nurdles just to avoid errors. Even when accepting all predictions regardless of confidence, the model still misses 13% of nurdles, showing that even its most confident predictions aren't finding everything. This poor calibration means users face a difficult choice: either use a very high confidence threshold and miss most nurdles or use a lower threshold and accept many unreliable detections.

Architecture Comparison

The architectural differences between these models are fundamental to understanding their performance. YOLOv8n with instance segmentation doesn't just draw boxes around objects - it also traces the exact outline of each nurdle at the pixel level, like carefully coloring within the lines. This segmentation capability provides several important advantages. It forces the model to understand object boundaries very precisely, not just approximately. The additional task of drawing outlines helps the model learn more detailed features during training. For small objects like nurdles, this detailed attention is especially valuable. The outlines also naturally help

separate overlapping objects since the model must trace each one individually. The main downsides are that it runs about 10-20% slower and uses slightly more computer memory.

YOLOv11n represents newer technology designed to be faster and more efficient. In theory, it should perform better through various technical improvements and optimizations. The simpler task of just drawing boxes (without outlines) should be easier to learn. However, in practice, these expected advantages haven't materialized for nurdle detection. The model performs significantly worse across all measures.

This suggests the improvements in YOLOv11 were focused on different types of detection tasks, possibly sacrificing the ability to detect small objects for gains in speed or efficiency. The segmentation component in YOLOv8n forces the model to learn precise shape and texture features at the pixel level - you can observe this in the detection images where bounding boxes fit tightly and strictly around each nurdle's actual boundaries. This detailed shape learning helps the model distinguish individual nurdles even in dense clusters and understand the subtle texture differences between nurdles and background. Without the segmentation head, YOLOv11n misses this extra learning signal and detailed feature understanding. The model learns only approximate locations rather than exact object shapes and textures, which proves particularly problematic for small, round objects like nurdles where shape precision is critical. The architectural changes that make YOLOv11 work well for some tasks apparently make it worse for detecting small, clustered objects like nurdles.

Potential Overfitting Concerns with YOLOv8n Segmentation

While YOLOv8n shows impressive 99.3% accuracy, this near-perfect performance raises an important concern: overfitting. Think of overfitting like a student who memorizes specific test questions instead of truly understanding the subject. The student might score perfectly on practice tests but fail when questions are worded differently. Similarly, YOLOv8n might have "memorized" the 1,300 training images rather than learning what nurdles generally look like, which could cause problems when encountering new situations.

Several warning signs suggest overfitting might be occurring. The small training set of only 1,300 images is concerning. The validation images likely came from the same collection effort using similar cameras, lighting, and locations as the training images, so they may not represent truly different real-world conditions. The segmentation approach, while accurate, requires learning very precise, detailed patterns that might make the model too sensitive to any changes in image quality or nurdle appearance.

There are practical situations where YOLOv8n's overfitting could make it perform worse than YOLOv11n. If deployed with a different camera or smartphone than used in training, YOLOv8n might struggle with different color profiles or image quality. When lighting conditions change, the model might fail to recognize nurdles it hasn't seen under those specific conditions. Different types of nurdles (different colors, sizes, or materials), dirty or damaged nurdles, wet versus dry surfaces, unusual camera angles, or new background types could all cause unexpected failures. YOLOv11n's more cautious, conservative approach might handle these surprises better because it hasn't learned overly specific patterns.

An overfitted model tends to be overconfident, making predictions even in uncertain situations. YOLOv11n's extreme caution, while causing it to miss nurdles, might be more appropriate when facing completely new scenarios. In truly different environments, YOLOv8n might produce many more false alarms than the 2 seen in testing.

To verify whether overfitting is a real problem, the model needs testing on completely independent images from different locations, times, cameras, and conditions that it has never encountered before. Currently, there's no evidence such comprehensive testing has occurred. Until then, the 99.3% accuracy should be viewed cautiously - it might represent genuine capability, or it might collapse when facing real-world diversity. The model should be expanded to at least 5,000-10,000 diverse training images and tested across multiple equipment types and environmental conditions before full confidence in deployment.

YOLOv8n appears significantly better than YOLOv11n based on current testing, but there's a risk this advantage might not hold up in all practical situations. The model might work perfectly in scenarios like training conditions but struggle when things look different. YOLOv11n's poorer performance but more conservative approach might be more reliable when encountering unexpected situations.

In Year 2 Quarter 3, the research team executed the re-annotation and retraining strategy outlined in the previous quarter's report. Following the planned approach of expanding the dataset with segmentation masks and implementing enhanced data augmentation, the YOLOv8n segmentation model was retrained and evaluated. The results demonstrate substantial improvements across all performance metrics, effectively addressing the overfitting concerns raised in Y2Q2.

In Year 2 Quarter 3, Task 3 activities focused on expanding the training dataset, retraining the segmentation model, and conducting a comprehensive performance evaluation to address overfitting concerns identified in earlier quarters.

Expanding the Training Data

The prior model was trained on approximately 1,300 labeled nurdle instances, which raised concerns about potential overfitting, where the AI memorizes specific images rather than learning generalizable nurdle characteristics. During Year 2 Quarter 3, the research team addressed this limitation by substantially expanding the training dataset to approximately 20,000 labeled nurdle instances through additional segmentation mask annotation and enhanced data augmentation.

While this expansion significantly improved dataset diversity and robustness, the research team acknowledges that the dataset continues to be dominated by controlled indoor photography. As a result, certain real-world beach conditions, including variable lighting, wet or reflective surfaces, and debris mixed with nurdles, may present challenges that are not yet fully represented in the current training data.

Overall Accuracy Metrics

The table below presents a comprehensive comparison between the Year 2 Quarter 2 baseline model and the improved Year 2 Quarter 3 YOLOv8n segmentation model, illustrating the impact of dataset expansion and retraining on model performance.

Metric	Y2Q2 YOLOv8n-seg	Y2Q3 YOLOv8n-seg	Change
mAP@0.5 (Box)	99.3%	99.5%	+0.2%
mAP@0.5-0.95 (Box)	~90%	~90%	Maintained
mAP@0.5 (Mask)	~99%	99.5%	+0.5%
mAP@0.5-0.95 (Mask)	~85%	~60%	See note
Precision (Box)	~100%	96-100%	Stable
Precision (Mask)	~100%	97-100%	Stable
Recall (Box)	~100%	99.0-99.5%	Stable
Recall (Mask)	~100%	97-99%	Stable
F1 Score	0.98	1.00	+0.02
Confidence for 100% Precision	57.3%	45.4%	-11.9% (improved)

Figure 10. ***Performance Metric Comparison Between Y2Q2 and Y2Q3 YOLOv8n Segmentation Models***

The apparent decrease in mask-based mAP@0.5–0.95 reflects stricter evaluation on a more diverse dataset rather than degraded performance. This outcome indicates improved model generalization, which is the intended result of the expanded training approach.

Performance Summary

Using a validation dataset containing over 6,200 labeled nurdle instances, the retrained model achieved a 99.7 percent detection rate, correctly identifying 6,234 out of 6,254 nurdles. A total of 20 nurdles were missed, and 14 false positive detections occurred where non-nurdle objects were misclassified as nurdles.

What Actually Happened	AI Said "Nurdle"	AI Said "Not a Nurdle"
Real nurdles present	6,234	20 missed
No nurdle present	14 false alarms	Correct

Figure 11. ***Confusion Matrix for YOLOv8n Segmentation Model Performance in Year 2 Quarter 3***

These results are encouraging; however, it is important to note that the validation dataset, while independent from training data, was collected under similar conditions. Model performance under true field conditions may differ.

Comparison to Previous Version

What We Measured	Previous Model (Y2Q2)	Current Model (Y2Q3)
Nurdles correctly found	96.8%	99.7%
Test dataset size	218 nurdles	6,254 nurdles
Overall accuracy score	99.3%	99.5%

Figure 12. ***Comparison of Nurdle Detection Performance Between Y2Q2 and Y2Q3 Models***

The ability of the model to maintain high overall accuracy (99.5 percent) on a validation dataset approximately 29 times larger than that used in Year 2 Quarter 2 provides encouraging evidence that the model is learning generalizable nurdle characteristics rather than memorizing individual images. However, this outcome does not guarantee equivalent performance on images collected from new locations, different camera systems, or under varied environmental conditions.

Confidence Calibration

As part of the Year 2 Quarter 3 evaluation, confidence calibration was examined to assess the reliability of model predictions. The retrained model requires only a 45 percent confidence threshold to achieve reliable results, compared to 57 percent for the previous version. This improvement suggests that the model can accept a broader range of detections without sacrificing accuracy. Final confidence thresholds for production deployment will require validation through real-world testing using citizen scientist submissions.

Comparison to Detection-Only Approach

Results from Year 2 Quarter 3 continue to demonstrate that the shape-tracing, segmentation-based approach outperforms the detection-only, bounding box method evaluated in earlier quarters.

Measurement	Shape-Tracing AI (Current)	Box-Only AI (Y2Q2)
Overall accuracy	99.5%	82.3%
Nurdles found	99.7%	78%

Figure 13. ***Performance Comparison Between Segmentation-Based and Detection-Only Nurdle Count AI Models***

The segmentation-based approach appears better suited for nurdle detection, likely because tracing exact object shapes enables the model to learn finer-grained visual features that distinguish nurdles from visually similar background objects.

Confusion Matrix Analysis

Confusion matrix results from Year 2 Quarter 3 further demonstrate strong detection performance.

Actual	Predicted: Nurdle	Predicted: Background
Nurdle	6,234 (99.7%)	20 (0.3%)
Background	14	—

Figure 14. ***Confusion Matrix Detailing True and False Classifications for the Y2Q3 Nurdle Count AI Model***

Key observations include a true positive rate of 99.7 percent, with 6,234 of 6,254 nurdles correctly identified, and a false negative rate of only 0.3 percent. False positives were minimal, with 14 background objects misclassified as nurdles. These results represent a substantial improvement over Year 2 Quarter 2, which showed 211 correct detections out of 218 instances. The combination of a dramatically larger test dataset and near-perfect accuracy provides strong evidence of improved model robustness under current validation conditions.

Limitations and Unknowns

Despite these promising results, several limitations remain and warrant careful consideration.

Dataset limitations include a continued emphasis on controlled backgrounds such as petri dishes and felt surfaces, with limited representation of challenging real-world environments including wet sand, mixed debris, and variable lighting conditions. Most images were captured using similar camera types and resolutions, which may limit generalizability.

Potential performance gaps include reduced accuracy for beach photography under natural lighting, difficulty detecting dirty, discolored, or weathered nurdles, and undercounting in dense clusters where nurdles touch or overlap. Images that do not follow established capture standard operating procedures may also yield lower accuracy.

Validation concerns remain, as test images were drawn from the same overall collection pipeline as training images. True out-of-distribution performance has not yet been evaluated, and the near-perfect validation accuracy warrants appropriate skepticism until confirmed under production conditions.

Areas for Improvement

Work planned for subsequent quarters will focus on addressing these limitations through targeted data collection and user-centered validation. Key priorities include expanding real-world validation using images submitted through the Nurdle Patrol website, collecting edge cases where the model underperforms to better understand failure modes, and integrating user feedback through a controlled beta testing phase. These efforts will be essential for refining model performance and ensuring reliability under diverse field conditions.

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

In Year 2 Quarter 1, initial design work began on the integration of Nurdle Count AI into the Nurdle Patrol Website (Design).

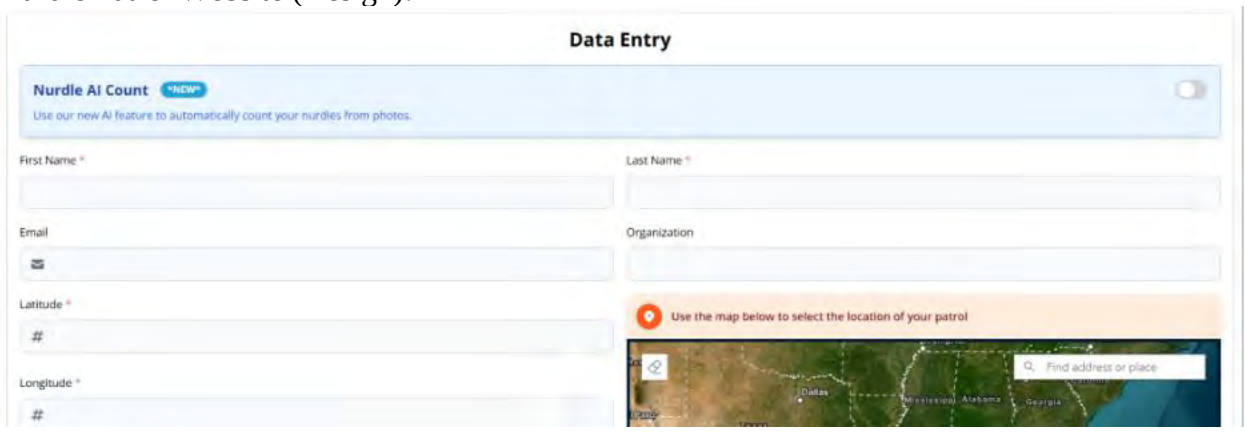


Figure 15: UI Data Entry

The Data Entry form on the Nurdle Patrol Website will include a toggle to enable the Nurdle Count AI feature. If it is enabled, the actions below will follow:

Figure 16: AI Detection & Confirmation

First, a pop-up will appear that uses Nurdle Count AI to automatically detect and estimate the total number of nurdles in the submitted image. The user will then be prompted to either confirm or reject the AI's detected count.

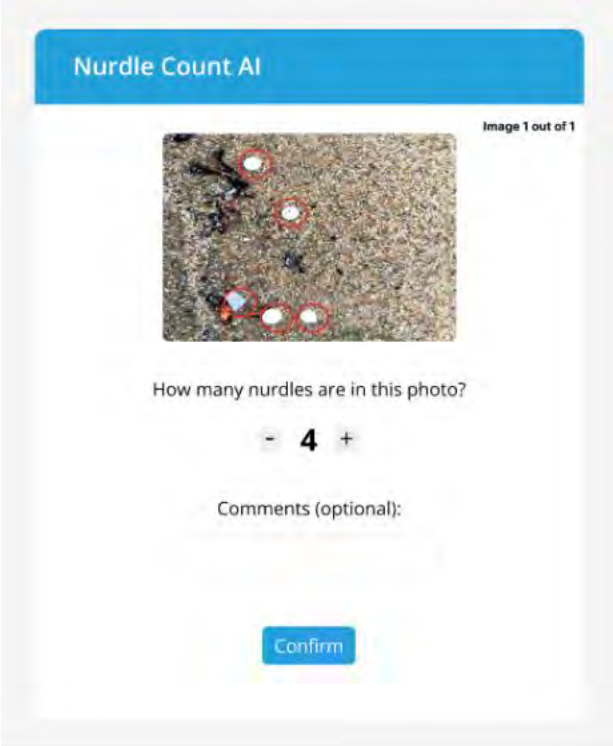


Figure 17: User Rejection & Manual Input

If the user rejects the AI's estimate, they will be asked to manually input the correct number of nurdles. An optional comment field will also be provided for additional notes or clarifications.

Figure 18: Form Auto-Population

Finally, once the user confirms, the form will be auto-populated.

In Year 2 Quarter 2, work continued to further the integration of Nurdle Count AI into the Nurdle Patrol website by developing workflows that support up to 2 photos (Figure x) (Figure x). The workflow allows users to submit their photo(s) and includes options to confirm, reject, or adjust the count as needed. Based on whether the user confirms or rejects the count, a different set of actions are presented to accurately account for all nurdles. Users can also leave comments within the interface, and finally, the count is auto-populated into the form, making the experience useful and easy.

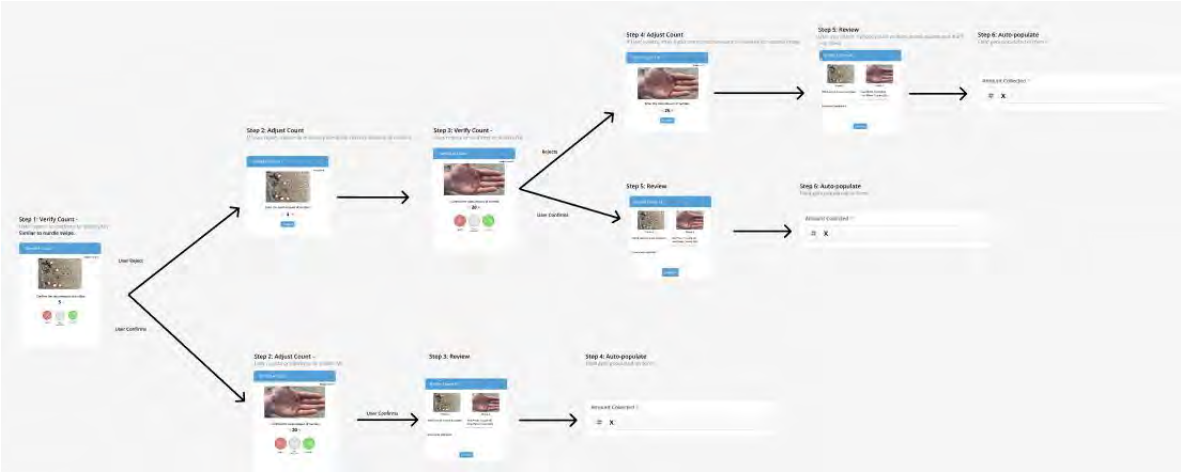


Figure 19: Workflow for 2 photos

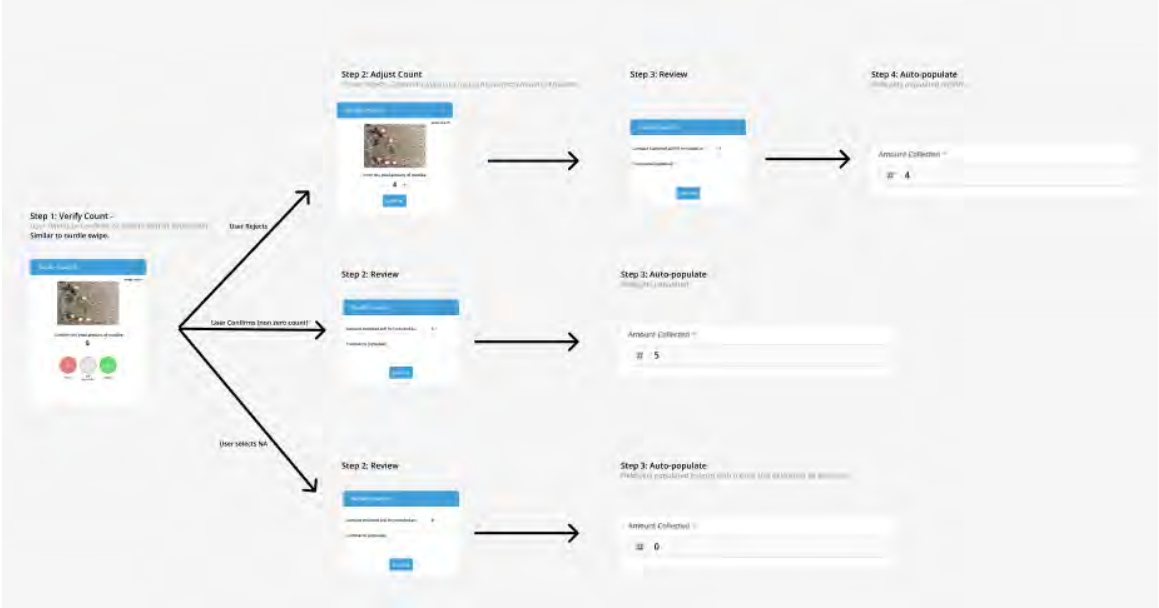


Figure 20: Workflow for 1 photo

Several popups were designed for ease of use and to guide users through every step of the process while using Nurdle Count AI to count their nurdles.

Integration Popup: Allows users to add counts for up to 2 photos. Users also have the option to select one count or the other if preferred. This is useful when users submit 2 photos of 2 different sets of nurdles that pertain to one location (Figure X).

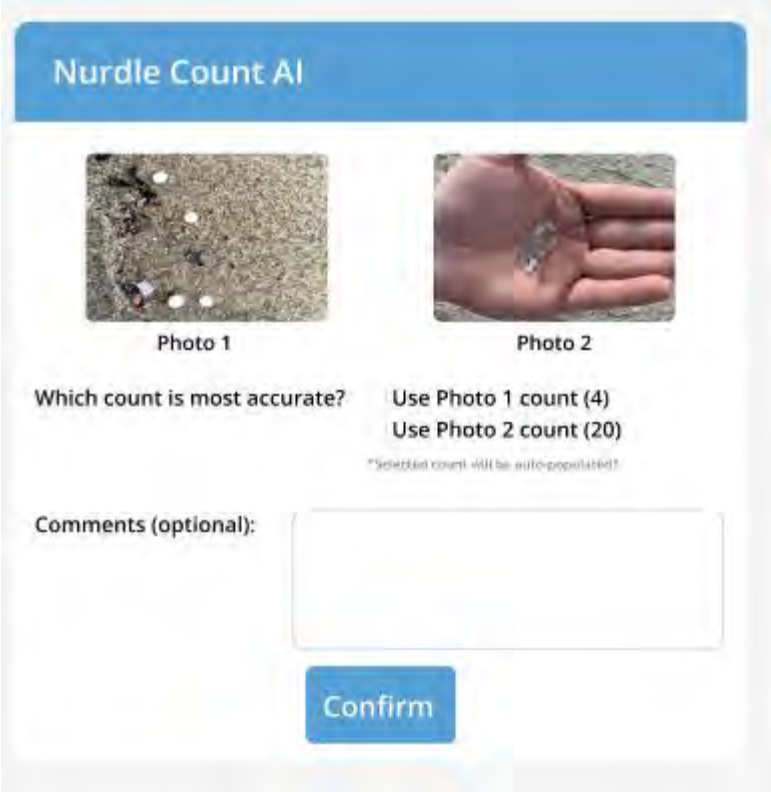


Figure 21: Workflow for 2 photos

Loading Popup: Displays while the Nurdle Count AI is detecting nurdles (Figure X).

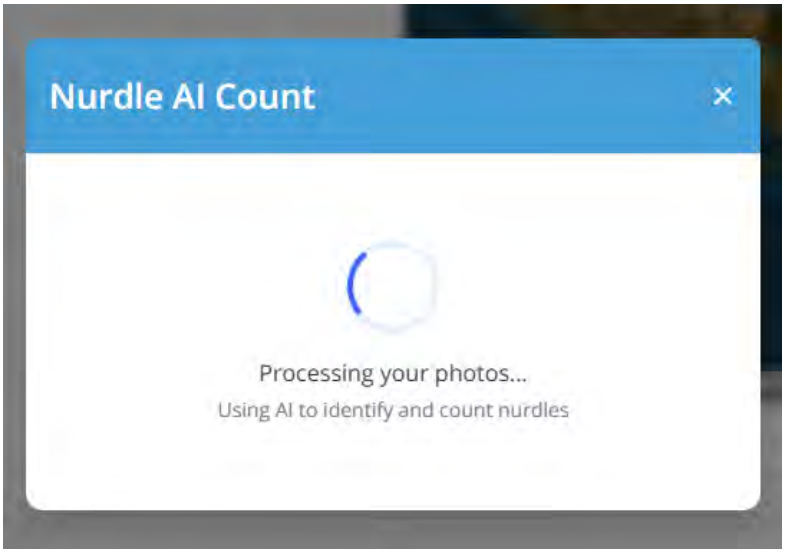


Figure 22: Pop-ups

Confirmation Popup: Displays once the user has finished confirming/adjusting the count (Figure X).

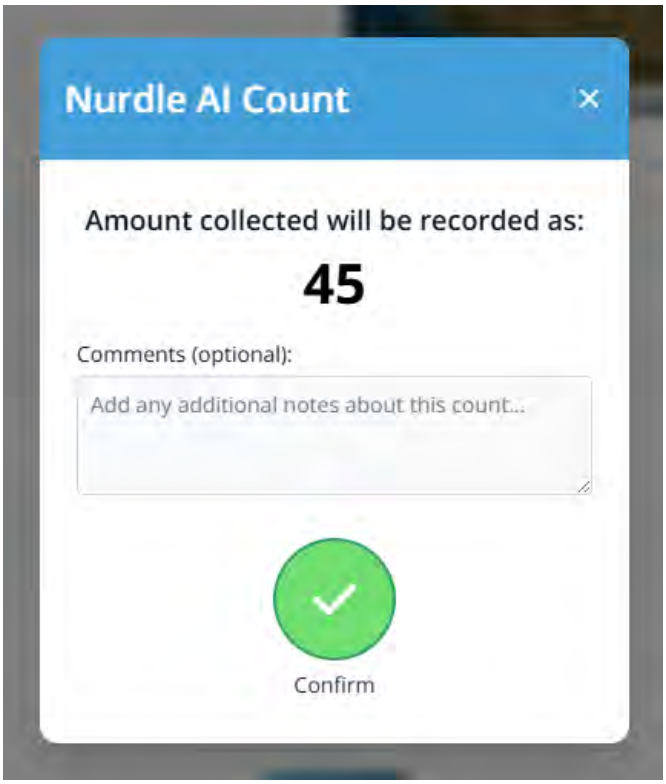


Figure 23: Confirmation Pop-up

In Year 2 Quarter 3, Task 4 work advanced from iterative development to workflow finalization. During this quarter, the Nurdle Count AI workflow was finalized, establishing a standardized and efficient pipeline for image ingestion, analysis, and output. This workflow now serves as the technical foundation for full integration into both the Nurdle Patrol website and mobile application.

In parallel, targeted user interface enhancements were implemented to support effective use of Nurdle Count AI. These include a dedicated toggle within the Nurdle Report to activate AI analysis, along with a series of interactive pop-ups that guide users through verification steps and required selections. Together, these improvements ensure a clear, intuitive user experience while maintaining data quality and user oversight of AI-assisted results.

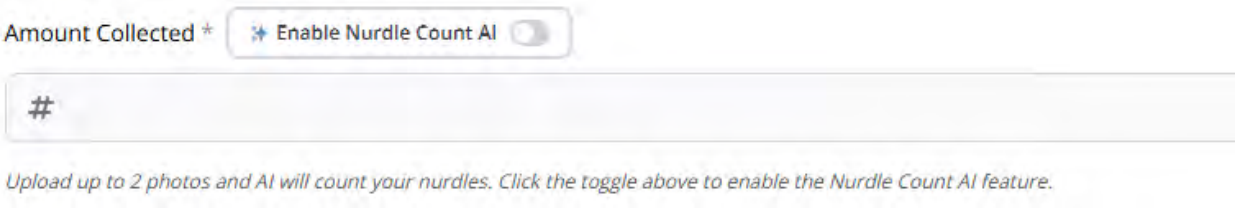


Figure 24: Enable Toggle for Nurdle Count AI

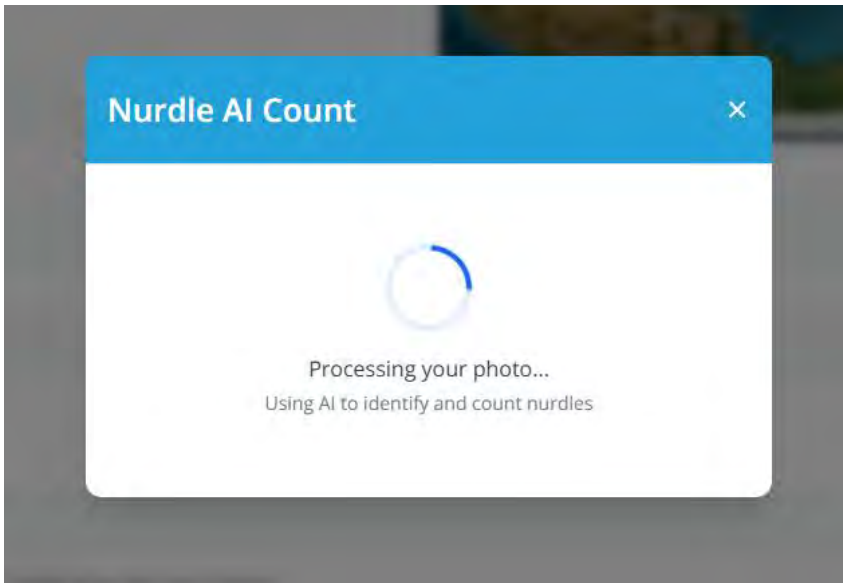


Figure 25: Analyzing Images

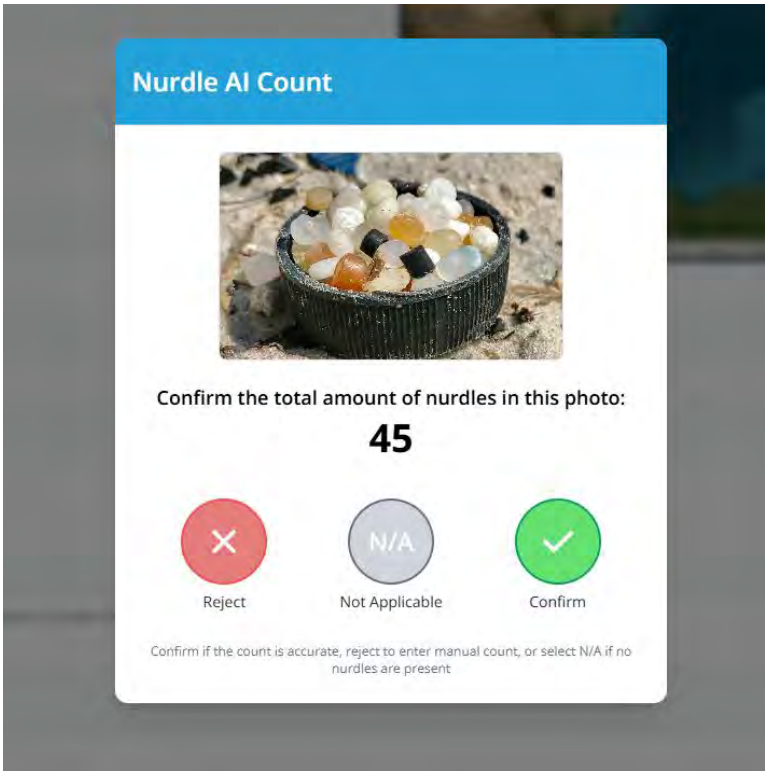


Figure 26: Verification Popup

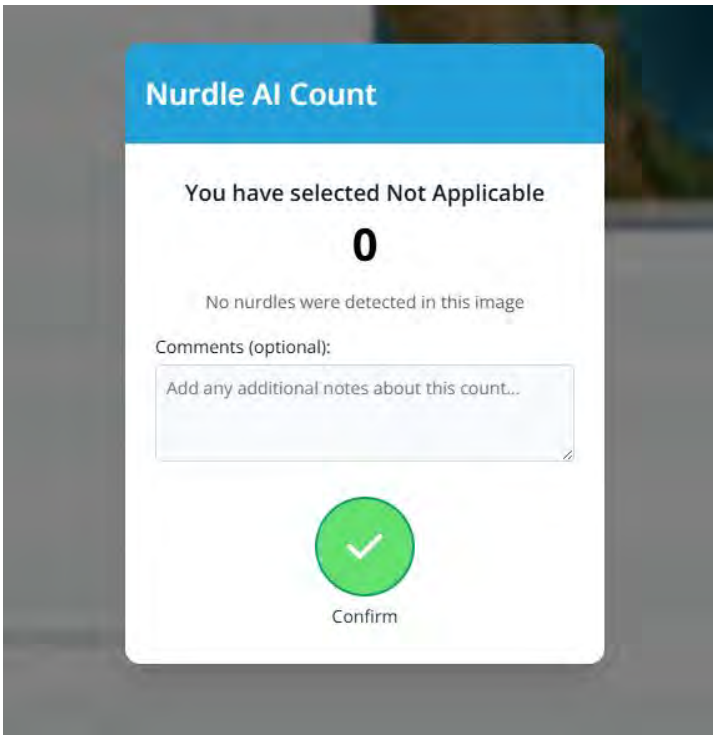


Figure 27: Example of User Selection

The current implementation supports dual-image processing, allowing users to submit up to two images for AI analysis. For each image, the system generates an AI-derived count of nurdles and provides structured comment fields for user annotations and contextual observations. Users retain full flexibility in how AI results are incorporated, including counts from one image, both images independently, or a combined result. Final control over submitted data remains with the user, ensuring data quality, transparency, and accuracy in all reported outputs.

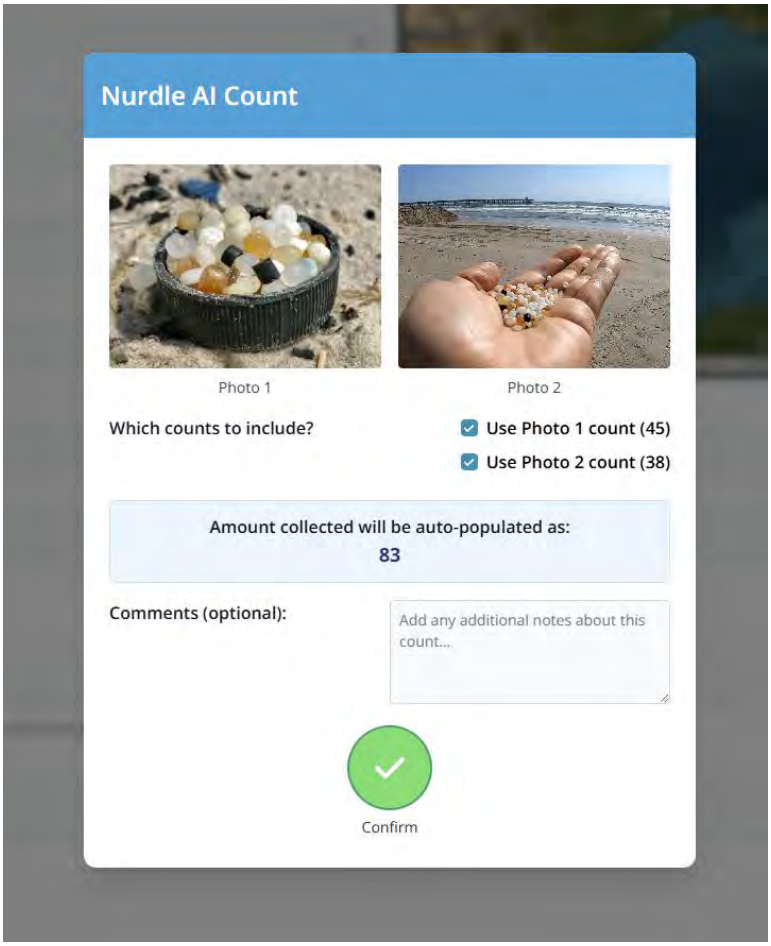


Figure 28: Confirmation Pop-up

The final project phase will focus on extending Nurdle Count AI integration to the Nurdle Patrol mobile application.

During Year 2 Quarter 3, Tasks 3 and 4 advanced from exploratory development toward validation, stabilization, and deployment readiness for the Nurdle Count AI system.

For Task 3, work focused on executing the re-annotation and retraining strategy identified in prior quarters and validating the performance of the YOLOv8n segmentation-based approach at scale. Building on earlier model comparisons, the research team expanded the training dataset,

retrained the segmentation model, and conducted a comprehensive performance evaluation using a substantially larger validation set. Results confirmed strong segmentation performance, with the YOLOv8n model achieving approximately 99.5 percent overall accuracy, a 99.7 percent detection rate, improved confidence calibration, and minimal false positives. Comparative analyses continued to show that the segmentation-based approach outperformed the YOLO11n detection model, which exhibited missed detections, weaker bounding box localization, and poor confidence calibration. Overfitting concerns identified in earlier quarters were directly addressed through dataset expansion and retraining, establishing a more robust and generalizable model while clearly documenting remaining limitations related to real-world environmental variability.

For Task 4, the team finalized and stabilized integration workflows for Nurdle Count AI within the Nurdle Patrol website. User-facing workflows supporting one- and two-photo submissions were refined and standardized, with guided interface pop-ups leading users through image upload, AI analysis, confirmation or adjustment of AI-generated counts, selection between multiple images when applicable, and automatic population of the data entry form with verified results. The finalized logic supports multiple user decision paths, including confirming AI estimates, rejecting estimates and entering manual counts, or selecting between counts derived from two separate images within a single submission.

Together, Year 2 Quarter 3 activities mark a transition from prototype development to validated modeling and cohesive, user-ready front-end integration. These efforts position Nurdle Count AI for controlled beta deployment while maintaining user oversight, data quality, and transparency consistent with the Nurdle Patrol citizen science framework.

Synergistic Activities:

In Year 2 Quarter 3, Jace Tunnell conducted a robust series of Nurdle Patrol outreach and education events across Texas and beyond, reaching a wide range of audiences from high school students to international organizations. These activities emphasized hands-on environmental education, citizen science engagement, and science communication.

Highlights include:

- 16 events delivered between August 1st and November 30, 2025, with a mix of in-person and virtual formats.
- Total reach of more than 1,260 participants**,** spanning K–12 students, teachers, community groups, scientists, and the general public.
- Local, regional, and national impact, with events hosted in the Texas Coastal Bend, at national and international venues such as Dallas College in person in Dallas, Texas, and the Universidad Autónoma de Baja California in La Paz, Mexico through virtual sessions that connected with broader audiences.
- Educational themes focused on beachcombing, plastic pellet (nurdle) pollution, science communication, oysters and water quality, and the role of citizen science in coastal stewardship.

Notable engagements:

- A large-scale event as a keynote speaker with 350 students and community members at Corpus Christi ISD science awards ceremony in December 2025.
- An outdoor booth event with 200 community members at the Briscoe Pavilion on North Padre Island, hosted by the Friends of Padre, held in November 2025.
- A radio show interview with a large audience reach in coastal areas.
- A community partnership with the Neighbor League of Corpus Christi (40 attendees).
- Virtual sessions to reach audiences beyond the Texas coast.

These events illustrate the strong demand for Nurdle Patrol's educational programming and the value of integrating science communication with citizen science opportunities. Collectively, this outreach fostered environmental awareness, built community connections, and encouraged active participation in monitoring plastic pollution across coastal environments.

Here is a full list of events conducted during this reporting period:

Date	Organization	Type	Subject	Title	Location	Number of Attendees
11/1/2025	Friends of Padre	In-person	Beachcombing/Nurdle Patrol	Booth	Briscoe Pavillion on North Padre	200
11/11/2025	Veteran's Memorial High School	In-person	Beachcombing/Nurdle Patrol	Beachcombing and Nurdle Patrol	Veteran's Memorial High School	120
11/14/2025	Austin school	Virtual	Nurdle Patrol	Nurdle Patrol Citizen Science Project	Zoom	35
11/20/2025	Island Moon Live Radio Show	In-person	Beachcombing and Nurdle Patrol	Radio	Doc's	
11/21/2025	Marvin Middle School	In-person	Nurdle Patrol	Nurdle Patrol	CBI	40
12/15/2025	Corpus Christi ISD	In-person	Beachcombing/Nurdle Patrol	Beachcombing and plastic pollution	Veteran's Memorial High School	350
1/12/2026	Gulf of America Alliance	Virtual	Marine Debris	Texas Marine Debris Projects	Zoom	47