**Using Machine Learning to Measure Changes in Immunity in the Fall Armyworm at Varying Temperatures**

**Abstract:**

**Intro:**

An insect’s primary immune response against small viral particles is the use of hemocytes. Hemocytes’ primary function is to phagocytize or release enzymes to help fight off invading small foreign objects within the host’s body. In lepidopteran species like the fall armyworm, multiple uniquely shaped hemocytes are found to aid in immune response (Lavine and Strand, 2002). Hemocyte counting is a vital step in understanding how the immune system of the fall armyworm and other arthropod species react to pathogens. When insects are infected with a pathogen, hemocyte numbers are usually lower than healthier individuals (Smilanich et al., 2018). Additionally, hosts infected with higher viral concentrations are expected to have more hemocyte counts when compared to healthy or individuals with a low dose (Eslin & Prévost, 1998).

The fall armyworm (*Spodoptera frugiperda*) is a major agricultural pest thought to originate from the tropical environments of Central and South America. The larvae feed primarily on corn, rice, and grasses, but have been known to destroy crops like tobacco and cotton (Wang et al., 2022). In the last 10 years, fall armyworms have spread into Western and Central Africa causing massive agricultural damage to maize plots within the area (Goergan et al, 2016). Fall armyworms thrive in warmer climates, often most active in the late summers in North America (Ali et al., 1990). While pesticides are often used to help cull outbreaks and were successful in the past, development of resistance to common pesticides like carbamates and organophosphates has emerged making it harder to control the populations (Tay et al, 2023). Biocontrol practices like the use of predators, parasitoids, and pathogens have been shown to be successful in reducing the number of larvae. Compared to pesticides, biocontrol methods are unique in they can evolve to overcome host resistances themselves, leading to more long-term control options. However, few studies have investigated the use of biocontrol agents to control populations of the fall armyworm, with those available focusing primarily on parasitoids and pheromone traps (Allen et al., 2021; Varshney et al., 2021).

The current method for detecting and counting hemocytes requires manual tallying the presence of hemocytes by hand. This method is very tedious, taking up to 60 seconds per image to process counts. These methods are not always accurate due to the high potential for user error when counting due to and the presence of hemocytes and other cells out of focus. Programs like ImageJ are commonly used for counting objects in photos (Schneider et al., 2012) However, due to the unique and varied shape of hemocytes, other miscellaneous cells present in samples, and the placement of cells onto hemocytometer grids, ImageJ often struggles to give correct counts. Hemocytes will cluster together after a certain period of time, often leading to miscounts and confusion due to the overlap of cells and the potential for hemocytes to take unique shapes. Additionally, camera quality and contrast are more likely to skew results and lead to further inaccuracies in counts. Photos taken in low light can further skew counts done by hand due to the difficulty seeing hemocytes.

Object detection allows for the localization and classification of objects contained within photos. Machine learning is crucial for object detection, allowing users to develop models to automatically detect objects within images. Many object detection approaches use convolutional neural networks (CNNs) which allows for learning to occur directly from photo data. (J. Ren & Wang, 2022). Object detection algorithms are separated based on the complexity of the model and the steps required for detection. Single-stage detectors start by directly feeding image data into a convolutional network to detect the location and classification of the objects. Common single-stage detectors include You Only Look Once (YOLO) and Single Shot Multibox Detector (SSD) (Redmon et al., 2016; Liu et al., 2016). Two-stage detectors add an additional region proposal step during detection before classifying and drawing bounding boxes. Fast R-CNN and Faster R-CNN are both common examples of two-stage detectors (Girshick, 2015; S. Ren et al., 2016). In theory, two-stage detectors tend be more accurate at the sacrifice of time and computational resources required due to the complexity of the network, while single-stage are faster to train, but less accurate overall and struggle more on small objects. For example, a previous study found Faster-RCNN to the be the most accurate model compared to other single-stage detectors (Wang et al., 2022). However, other studies differentiating fish morphology and detecting cars from aerial images found single shot-detectors to be more accurate overall. Additionally, more recent developments in YOLO models (YOLO V8) improve the chances of overcoming two-stage detectors (Jocher et al., 2023).

Deep learning methods have been used in other previous studies to aid in the detection of cells other than hemocytes. For example, many studies have performed analysis of blood smears to aid in the detection of infected white blood cells (Poostchi et al., 2018). However, these studies often aim to detect large objects taking up a majority of pixels in the photos. One current obstacle faced in objected detection is the difficulty in predicting small objects. This is in part due to the limited data for use in training and the low resolution that often occurs due to the objects small size (Nguyen et al., 2020). Additionally, many CNN-based architectures rely on hierarchy feature mapping in predicting objects. While this may work for large and medium sized objects, small objects can easily be lost in the latter layers of these models due to the down sampling required (Liu et al., 2021). Therefore, it is important to consider how to account for the small size of hemocyte cells and compare the accuracy of multiple models, including both single-stage and two-stage detectors.

To our knowledge, no studies have investigated automating hemocyte counts in insects driving the need for new methods. Using machine learning should lead to accurate hemocyte counts that take significantly less time compared to manual counting. We expect single-stage detectors to be faster and more efficient during training, but we expect Two-Stage detectors like Faster-RCNN to be more accurate. For trial data on temperature comparisons and hemocytes, we expect temperature to influence hemocyte counts. Additionally, we expect caterpillar colonies exposed to virus to show a stronger hemocyte response than those who have not. Finally, caterpillars exposed to virus when dissected should show a stronger hemocyte response than those that are not.

**Materials and Methods:**

* **Rearing and Hemocyte Extraction**

The first generation of fall army worms were purchased from Benzon Research Incorporated (cite?). Caterpillars were reared on a mix of corn-based diet until fourth instars. For the first generation, caterpillars were split into two treatments: 26° C and 31° C. The 26° C treatment was kept at 26° C for 12 hours during the day and 16° C at night while the 31° C treatment was kept at 31° C for 12 hours during the day and 21° C at night. For each treatment, a dose response was initially performed to calculate the lethal dose of virus that 50% of the population dies at (LD50). We isolated and starved 180 caterpillars for 24 hours to ensure consumption of virus. Caterpillars were then transferred to individual cups containing small cubes of diet laced with 3 **μl of** virus solutions. The dosage of virus ranged from 100 cells/μL of virus to 106 cells/μL of virus, increasing by a factor of 10 for each virus treatment. The control tray was dosed with DI water containing no virus to account for aversion to wet diet. We left caterpillars on with the virus (or water) laced cubes for another 24 hours to ensure full consumption of diet. Once the 24 hours had passed, caterpillars were transferred back into individual containers half full of the corn-based diet. The date of mortality or pupation was recorded for each individual for use in calculating the LD50 for the subsequent generation. Once all caterpillars had pupated or died, individuals infected with virus were processed to collect remaining viral particles. This processed virus was then used to infect the next generation of caterpillars. For individuals that survived infection and pupated, caterpillars were transferred to cages containing forty to sixty individuals to ensure reproduction. Eggs were collected every 24 hours and labeled as the next subsequent generation. Neonates that emerged from the new generation were isolated into 1 oz cups filled with ½ ounces of diet. Once the neonates grew into fourth instars, they were then starved, and another dose response was performed to calculate the LD50 for that generation. This was repeated every generation infecting each subsequent generation with virus collected from the previous generation.

Within each temperature treatment, 27 caterpillars were set aside as controls while an additional 27 caterpillars were dosed with enough virus to kill 95 percent of the population of virus. Nine caterpillars from the control and nine caterpillars from the infected treatment were taken for collection every 24 hours, until 72 hours had passed and all caterpillars were used. As per USDA protocol for extracting hemocytes, larvae were initially placed at 4°C for 30 minutes. Once the insects were mostly immobile, they were transferred on ice for extraction. The fourth abdominal prolegs were sterilized with ethanol and then amputated using surgical needles. Hemolymph from the incision was allowed to pool for collection. Three caterpillars per treatment were bled and the total hemolymph of these caterpillars were collected. This was repeated two more times for both controls and infected, leading to three sets of photos. Once hemolymph was extracted, the solution was mixed with a chilled anti-coagulant saline solution in a ratio to insure at least 10 µl of solution were present. Generally, a ratio of 1 hemolymph: 2 anticoagulants was used for 48 hour and 72 hour treatments while a ratio of 1:3 hemolymph to anticoagulant was used for 24 hour treatments. The solution was then placed on a Hemocytometer for counting under a 40x magnification lens. A (MODEL OF CAMERA) was used to photograph 5 of the 25 squares for estimating the total number of hemocytes.

* **Machine Learning**

To gather data used in training the model, the same approach was used for collecting and gathering hemolymph. Nine photos were taken on the hemocytometer grid. Additional photos were taken off of the grid if hemocytes exhibited a unique shape or was in a large cluster. Light intake and camera zoom was varied for each set of photos to account for differences when capturing photos. 500 photos were collected taken for use in training and validating the model. Each photo was manually annotated using the program LabelImg (<https://github.com/qaprosoft/labelImg>) to ensure annotation accuracy. Halos of light present around hemocytes were included in boxes to increase the size of detection and give more context to the objects when training. Hemocytes out of focus or lacking definition were not included during annotation.

Photos were randomly separated into training, validation, and testing with 70% of photos entered as training data, 20% as validation, and 10% as test photos to ensure data was accurately read. Training data was used to train the models, the validation set was used during training to estimate the current model accuracy and help prevent overfitting, and the test set was used once training finished to measure how accurate each algorithm was when counting hemocytes. After data was sorted, the photos labeled training were augmented to increase the size of the dataset. Data in training was then augmented to flip 180 degrees to account for any unusual orientation of hemocytes and better account for changes in light. The contrast of photos was further exaggerated to ensure hemocyte detection could occur even with different levels of exposure. This tripled the size of our training data. Models were then trained twice, once with and once without the training data to determine if the additional computational time required to train more photos was beneficial to the overall accuracy.

All models were trained using the programming language Python. Two versions of YOLO models (YOLOV5x and YOLOV8x) were used to represent single-shot detectors. YOLO models were allowed to run until no improvement had been made in the past 50 runs, up to a maximum of 200 epochs. Code for training YOLO models was sourced from Ultralytics for use in training on our own models (Jocher, 2020; Jocher et al., 2023). All models other than YOLO based models were trained using MMDetection and code available on Github due to their extensive model library and easy to configure models (Chen et al., 2019). The faster-RCNN model with a backbone of R-101-FPN was run until 200 steps were performed and no significant changes in loss rates were observed. The results from YOLO models were compared to all other models to determine the fastest and the most accurate algorithm.

All models were trained using two NVIDIA V100S GPUs, each with 32 GB of memory. Additionally, all photos were condensed to a size of 1280 x 1280 pixels to reduce the overall training time. Batch sizes were kept constant at 8 across all models to better measure training times. Models were validated during and after training to calculate precision and recall. Once training and validation of models had occurred, detection was performed on a subset of the data. This test data was then compared to the counts performed by hand to estimate the accuracy of our counting models. Additionally, models were characterized using the recall, precision, and the Mean Average Precision when predicted boxes overlapped more than 50% (mAP50) to determine the best model for use. mAP50-95, or the average precision across multiple overlap thresholds ranging from .5 to .95, was used to further classify the best model. Training time and detection time were also recorded to determine the most accessible model for hardware with lower power. The definition of precision and recall are shown below.

Once all models were trained, the model with the highest mAP50 was used for counting and further comparisons. Counting was compared by taking the best model’s outputs on detections and comparing them to manual annotations on the test data. Temperature treatments were initially compared using all photos regardless of infection status and a T-test was performed to test for significance. Once all photos were cropped down to the hemocytometer square, counts were remade, and another T-test was performed to compare hemocyte counts. Once comparisons between temperatures had been performed, the cropped photos were further broken up into temperature and infection status and counts were plotted. An analysis of covariance was performed to calculate the significance of the data.

**Results:**

Initial comparison of models showed YOLOv8 has the highest mAP50 at 98.5% on unaugmented with a precision of 95.4% and a recall of 94.2% (Figure 2). This model had a 0.04% higher mAP50 than the most complex YOLOv5 model and faster R-CNN models. Additionally, compared to the Faster-RCNN model, both YOLOv5 and YOLOV8 had overall shorter training times (Table 1). YOLO8 also showed a higher mAP50-95 than both models. Augmenting photos to increase training data had surprisingly little effect on accuracy and precision. YOLOV8 trained on augmented photos had a slightly higher mAP50 of 0.01, but a lower mAP50-95 compared to YOLOV8 models trained on unaugmented data. Other models followed similar trends where augmentation had little to no effect on the accuracy of the model. Due to having the highest accuracy, YOLOV8 was used for counting hemocytes and any further analysis present in photos. Compared to actual counts, counts generated from the model were mostly accurate, only missing one or two hemocytes for the most part. The accuracy of counts also tended to decrease as more hemocytes were present within the photo (Figure 3). False positives tended to be small objects on the edge of photos, or cells with a similar size and shape to hemocytes. False negatives tended to occur when a hemocyte was out of focus or clustered together with many other hemocytes.

When using uncropped photos and comparing all photos from the 26 °C to the 31 °C treatment, caterpillars at temperature treatment 26 °C showed a higher mean hemocyte count of 70.67. Caterpillars reared in the 31 °C treatment showed a mean count of 62.16 (Fig 4). A t-test of the two treatments yielded a p-value of 0.0486, meaning that the lower temperature populations showed an overall stronger response regardless of disease presence. When cropping the photos down to squares on the hemocytometer, the 26 °C treatments had an average of 32.04 hemocytes and the 31 °C had an average of 29.65 hemocytes (Fig. 5). However, there was no significant difference between hemocytes when focusing specifically on squares (p = 0.216).

Further breaking down temperature treatments into coevolved and control vs infected showed that coevolved colonies tended to have higher hemocyte counts regardless of the presence of virus. No significant difference in hemocyte counts was observed between the infected coevolved populations at 26 and 31 °C. Uninfected coevolved populations saw a significantly higher number of hemocytes in the 26 °C treatments compared to the 31°C (Figure 5). Infected control populations saw a significant difference in counts with a higher mean in the 26°C population. Finally, the uninfected control showed the opposite effect where the 31°C treatment had a higher average hemocyte count than the 26°C treatment.

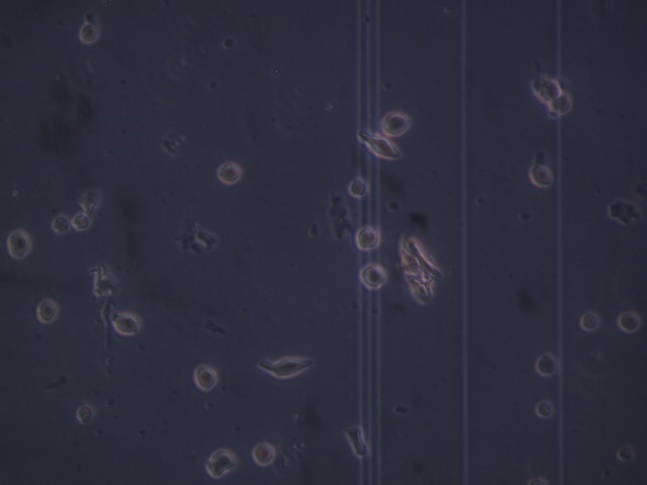
**Discussion:**

Based on our results, unless computational power is a limiting factor, YOLOv8 is the best model to use when detecting hemocytes. When compared to YOLOv5 and Faster-RCNN, YOLOv8 had higher accuracy and very short detection times overall. It is unsurprising that YOLOv8 is the most accurate compared to other single-stage models as it is the most recent model in the YOLO family and is predicted to be more accurate on the COCO dataset when compared to the others (Jocher et al., 2023). It is surprising that YOLO is more accurate than Faster-RCNN however due to YOLO being a one-stage model instead of two. This could be due in part to background contrast and with Faster-RCNN being an older model. More surprising however, was how little augmenting the training data changed the overall accuracy of predictions. The addition of augmented photos has been shown to improve predictions in other studies (Shorten & Khoshgoftaar, 2019). Our data

We need more here but holding off for now.

**Figures:**

A picture containing black

Description automatically generated 

**Figure 1.** Two examples of hemocytes placed on hemocytometer grids at different levels of brightness and saturation. A shows hemocytes on one square of the hemocytometer grid. B shows uniquely shaped hemocytes

Chart, scatter chart

Description automatically generated

**Figure 2.** The comparison between the hemocyte counts obtained by hand and the hemocyte counts as per Yolo8. The solid line represents a perfect count.

**Table 1.** Machine learning algorithms and information on the average precision, speed of training using the training set, and detection times without data augmentation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Architecture | Precision | Recall | Map50 | Map50-95 | Time to train | Detection time |
| Yolov5 | 0.964 | 0.958 | 0.981 | 0.773 | 1.9Hours |  |
| Yolov8 | 0.954 | 0.963 | 0.985 | 0.785 | 1.14 Hours |  |
| Faster R-CNN |  |  | 0.981 |  | 15 Hours |  |
|  |  |  |  |  |  |  |

**Table 2.** Machine learning algorithms and information on the average precision, speed of training using the training set, and detection times with data augmentation. The training size increased to 1371 after augmented photos were added.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Architecture | Precision | Recall | Map50 | Map50-95 | Time to train | Detection time |
| Yolov5 |  |  |  |  |  |  |
| Yolov8 | .96 | .953 | 0.986 | 0.77 | 1.70 Hours |  |
| Faster R-CNN |  |  | .981 |  | 15 Hours |  |
|  |  |  |  |  |  |  |

Chart, box and whisker chart

Description automatically generated

Fig 3. Average number of hemocytes in uncropped photos (1 is 26, 2 is 31)

Chart, box and whisker chart

Description automatically generated

Figure 4. The average number of hemocytes when focusing on one square of the hemocytometer. (1 is 26, 2 is 31)

Chart, box and whisker chart

Description automatically generated

Figure 5. Hemocyte counts at each temperature across treatments.

**Acknowledgements:**

We thank Kale Costanza and Scott Grimmell for help in data collection. This study was funded by (Funding Source)

**Author Contributions:**

Nathaniel Haulk and Bret Elderd designed the ideas and experiments. Nathaniel Haulk collected data and ran analysis. Nathaniel Haulk and Bret Elderd wrote the manuscript

**Conflict of Interest Statement:**

The authors have no conflicts of interest.

**Data Availability Statement:**

Photos and all code used are made available at <https://github.com/natom24/Machine-Learning>

**References:**