Smart Agriculture - Capstone Project for Analytika

Jarvis Nederlof, Rita Ni, Shivam Verma, and Vignesh Chandrasekaran

June 29, 2020

1 Executive Summary

Using state-of-the-art object detection and image segmentation techniques we are able to infer the yield on different plant experiments. We demonstrate and produce an effective proof of concept model employing transfer learning that is capable of accurate predictions on a variety of Tomato and Basil plant species across 4 different experimental datasets. Finally, we visualize our yield predictions and results in an interactive dashboard combining the predictions with the operational experiment data (sensor data), showing the accuracy of the predictions and providing a solid base for future work.

2 Introduction

Whether dealing with arable land, fertilizer use, or many other macro aspects, food technology has seen a lot of progress in terms of efficiency gains. These efficiency gains, however, haven't manifested completely into the microcosmic and organism level. To this end, plant-phenotyping has emerged as a science which attempts to understand a plant's interaction between the macro and micro aspects, between the environmental and genetic conditions. The plant phenotype is demonstrated when there is a dynamic reciprocity established between the genotype of the organism and the plant's physical environment, measured by a plant's commercial yield. Currently, scientists are working to understand the yield variation based on recipes developed using the various environmental conditions [1].

3 Project Description

Analytika is a Calgary based AI & IoT company, specializing in Industry 4.0 solutions to optimize and transform industrial processes. As part of their commercial offerings they are developing a smart agriculture solution to optimize overall life cycles of crops in Canada and Mexico. They collaborated with researchers in Mexico to obtain sensor data corresponding to environmental conditions (air temperature, water pH levels, etc.) and temporal images of plants housed in controlled containers. The company is attempting to use state-of-the-art computer vision techniques to measure the yield of the plants. The yield will then be combined with the sensor and environmental data to identify high performing crop recipes. This is analogous to a machine learning process whereby the target/response is inferred from the images with the sensor data acting as predictive variables.

4 Project Objective

Our contribution comes in the form of a proof of concept concerning the use of computer vision object detection and image segmentation techniques to extract a plant's yield. To measure the yield from the images we predict the plant's leaf area. Leaf area is an important metric in determining plant growth since larger leaf areas usually result in higher light inception and yield [2]. In addition to the leaf area, we also predict the number of leaves of a plant. Number of leaves is an important metric in this particular experiment as confounding factors can make identifying leaf area difficult, and the combination of leaf area and number of leaves could be a better indication for plant yield and growth characteristics.

We formalize our results in the form of an interactive dashboard. In the dashboard users can see the predictions overlaid onto the plant's image with predicted leaf count and area clearly displayed. Users can sequence through the different experiments and predictions, and examine the predictions on an average basis, smoothing out some of the noise in the predictions. Additionally, we cross reference the predictions with the sensor data based on image timestamps so the user can see what the environmental conditions were in the container for each prediction.

5 Data & Architecture

There are two types of crops we looked at, Tomato and Basil. Tomato had two experimental groups/recipes: blue & and basil (i.e., tomato grown with basil). Basil had two experimental groups/recipes: control-1 and control-2.

5.1 Sensor Data

Electronic sensors were used in each experiment to measure different experimental and environmental factors. The sensors measured factors relating to water, light, and air metrics. Each experiment measured a different combination of factors which is summarized in the table below:

Metric	Units	Basil 1	Basil 2	Tomato Control	Tomato Blue	Tomato Basil
Water Temperature	Celsius	NA	V	V	V	V
Water Potential H	pH	C	V	V	V	V
Water Electrical Cond.	ms/cm	NA	V	V	V	V
Illumination Distance	cm	V	V	V	V	V
Light Spectrum NM	%	V	V	V	V	V
Light PPFD	µmol/m2/s	V	V	V	V	V
Air Carbon Dioxide	ppm	V	V	V	V	V
Air Humidity	%	V	V	V	V	V
Air Organic Compounds	ppb	NA	NA	V	V	V
Air Temperature	Celsius	V	V	V	V	V

Figure 5.1 Sensor Data: V - Varied, C - Constant, NA - Not Measured

Each sensor recorded information at a sampling frequency of 5 minutes, but had significant noise in the frequency of measurement. Noise in sensor data is a common occurrence in IoT devices due to various effects such as network issues and/or delays in data transmission, etc. The noise in our experiments resulted in out of sync timestamps which made it difficult to merge all the sensor data. To mitigate the data issues we incrementally rounded-off the time difference between measurements to remove noise in the data to a point where the number of missing values and number of records arising due to the outer join of all the sensor data is minimized. We further treated the data for outliers and then imputed missing values due to the sensor's missing readings using a moving average imputation strategy. Appendix 9.1 shows the effect of rounding timescale on merged sensor data for Basil control-2.

5.2 Image Data

Images were captured every hour and were used to infer the yield characteristics - namely, the number of leaves and area. Like with the sensor data, the image data had it's own unique challenges:

- Camera position and lighting: Some images were cropped, blurred, etc.
- Overgrowth: Some plants overgrow and blocked other plants in the same container which made it difficult to distinguish between adjacent plants
- Plant replacement & trimming: Plant leaves were trimmed during the experiment, and plants were sometimes removed or replaced.

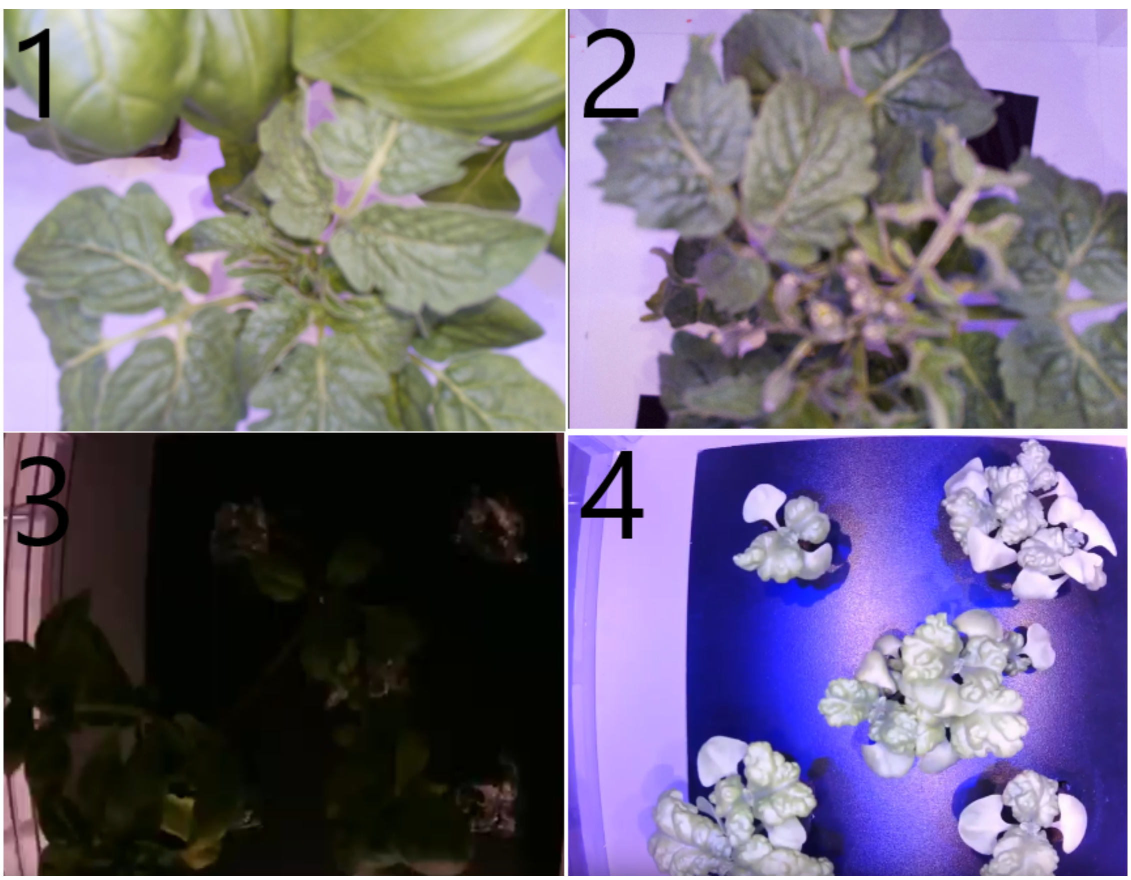

Figure 5.2 Image Issues: 1 - Overgrowth, 2 - Periodic trimming, 3 - No lighting, 4 - Different lighting

Apart from the challenges with the images themselves, we had a lack of processed images that we could use in the model training. In order to get satisfactory results we needed many leaf annotated images of plants that a model could learn on. To satisfy this requirement we acquired annotated and ground-truth images from several external sources, which are described in the project architecture (section 5).

6 Project Architecture

The project cloud architecture was setup on Microsoft Azure and was dependent on two modules - Blob Storage and Machine Learning Studio. The entire image data was housed on the blob storage (~12,500 images), which included the partner's data and additional data needed to train the model on. The Machine Learning Studio was used to build, evaluate and make predictions via Jupyter notebooks on hosted compute instances (high performing CPU and GPU instances). The output from Azure was connected with Microsoft Power BI to build the dashboard.

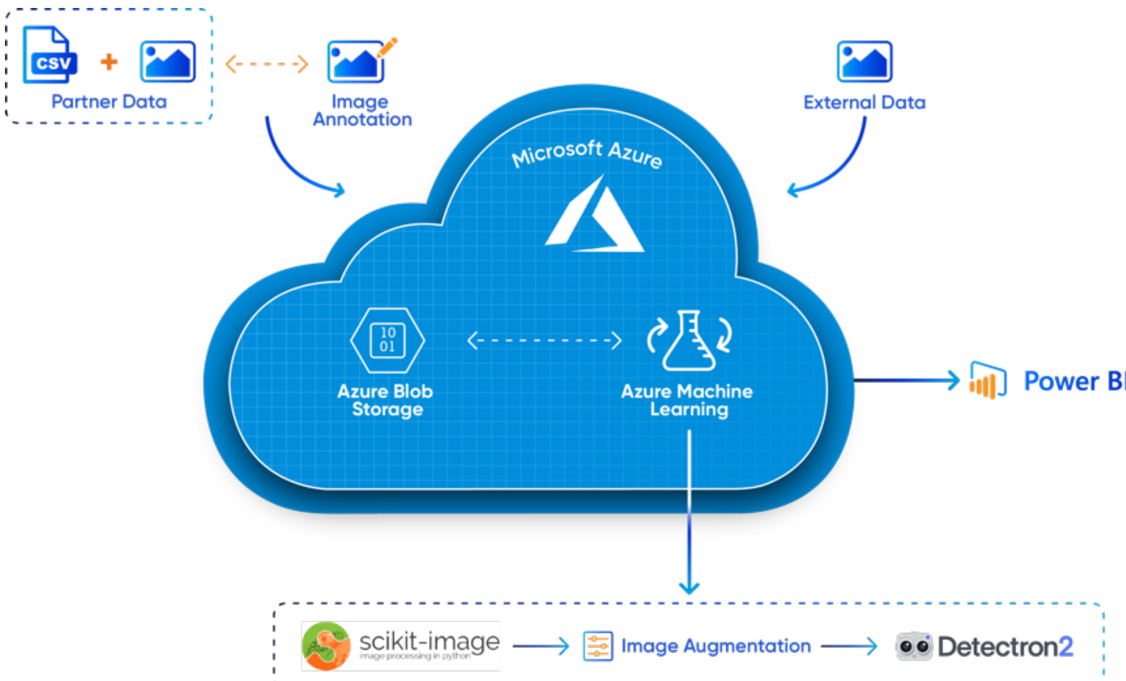

Figure 5.3 Project architecture workflow: Source - Melissa Egan, 2020

6.1 Model Training Data

Apart from the challenges with the image data given by the partner, we also didn't know the ground truth of the images, a necessity for training the image segmentation model. Therefore, we identified publicly available data sources that gave annotated leaf masks from plant images. Data from different sources gave a diverse set of plant growth stages. We organized the data into primary and secondary training data.

6.2 Primary Training Data

We sourced three separate external datasets of annotated ground-truth masked plant images to aid in the training and to increase model performance:

- **Synthetic Arabidopsis Dataset**: This dataset provides 10,000 top-down synthetic images of Arabidopsis plants. The images are derived using random variations from a 3D plant model. Random variations include changes in length, width, angle of the leaf, curl of the leaf, etc. [3].
- **Plant Phenotyping Dataset**: This dataset consisted of 1344 annotated images of real plants of varying varieties, sizes and plant species [4].
- **Aberystwyth Leaf Evaluation Dataset**: This dataset has 56 annotated images of plants grown in trays for a total of 916 arabidopsis ground-truth plant segmentations [5].

6.3 Secondary Training Data

In addition to the external training datasets, we also required annotated images from the experimental dataset in order to evaluate the model, and increase accuracy. Using images from the experimental dataset would help the model to learn about type, size and texture of the leaves specific to our images. Annotating images, however, represents a time-consuming task, therefore, we chose to only annotate 60 images. We used an open-source polygon selector and annotation tool `makesense.ai` to manually annotate 60 images and get the ground truth in the form of a JSON file containing XY coordinates of annotations. The steps for annotations are included in the Appendix.

In order to increase the number of training images in an efficient manner (without annotating more images) we implemented image augmentation which randomly flipped, rotated, saturated, and contrasted the images. Holding out 20 images for testing, we augmented 40 images which increased the secondary training batch to a total of 160 images.

6.4 Test Train split

To train the model, we implemented a split training technique. We first trained the model on all 11,400 external images for 1,000 iterations – the primary train. This gave us a good base for identifying leaf and plant components, but was too broad in its predictions. In order to fine-tune our model to our specific dataset, we trained a further 500 iterations on our 160 annotated and augmented images – the secondary train. We used 20 of our annotated images to evaluate the model performance. A summary of the above is depicted in Figure 6.1 below

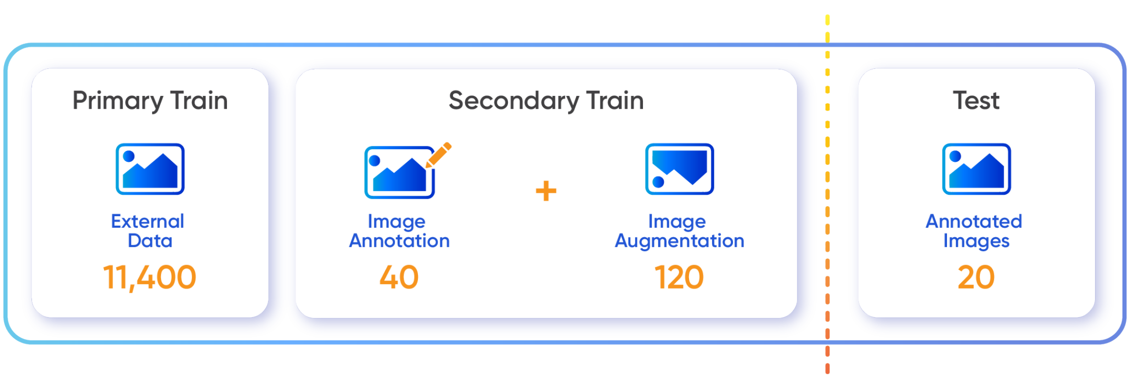

Figure 6.1 Test Train Split: Source - Melissa Egan, 2020

6.5 Evaluation metrics

To understand the output of the model and further tune it, the metric that was explored and found to be a good fit was Average Precision. The average precision considers both the object detection and image segmentation needs of the model. By varying the prediction thresholds for object detection, we obtain the precision-recall curve and by varying the intersection-over-union (IoU) scores we obtain many such precision-recall curves. By finding the mean across all the thresholds (ranging from 0.1 to 1.0, increased by 0.1) and IoU (0.5 to 0.95, increased by 0.05), we find the mean Average Precision which demonstrated the stability of the predictions across various thresholds and area. This is finally reported as AP. AP50 is also interesting which is for an IoU of 0.5. The model results show a good increase in performance moving from the primary train to the secondary train execution. Overall, we were able to achieve AP and AP50 scores of 61.7 and 87.0 respectively, depicted below:

Model Instance	Iterations	AP	AP50
Primary Train	1000	32.8	65.2
Secondary Train	500	61.7	87.0

Table 6.2 Average Precision (AP & AP50)

6.6 Data Product

The data product of this project is a user friendly dashboard that can dynamically display predictions from our leaf detection and segmentation model on top of plant images (i.e. experiment images). It also shows the corresponding operational parameters for any time point during the experiment period. The goal of this dashboard is to visualize the predictions generated by our leaf detection model and assist researchers in performing experimental inference on different growing recipes.

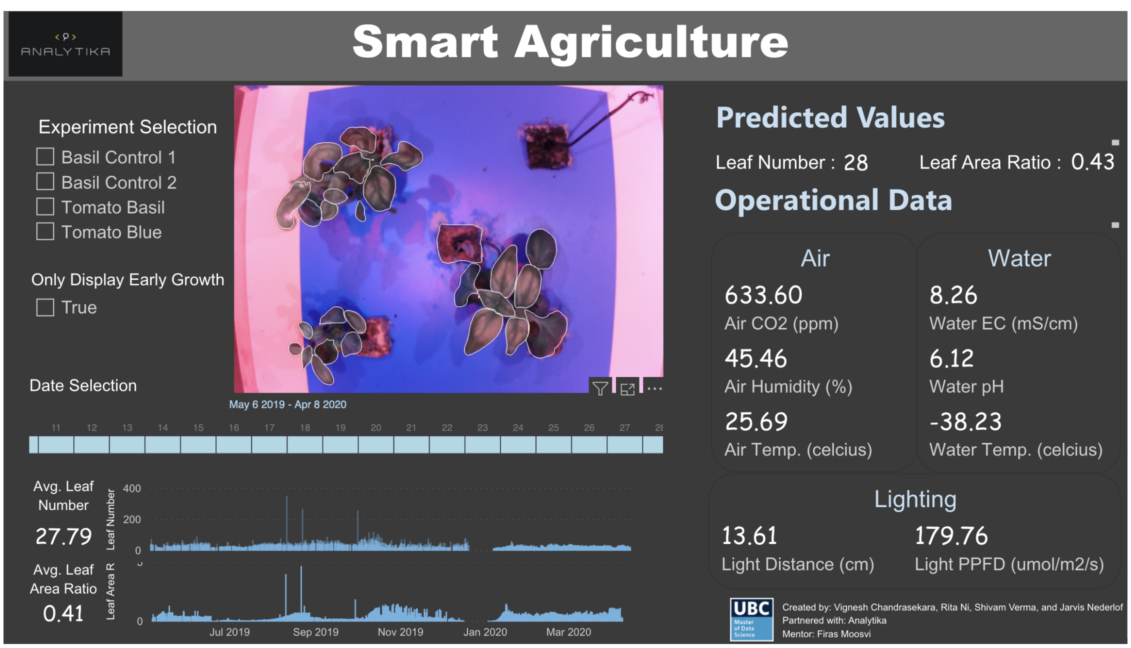

Figure 7.1 Smart Agriculture Dashboard

To visualize model predictions, the boundary of each predicted leaf is outlined in white, which overlays on top of the original image to help evaluate the quality of the prediction. The leaf area ratio is calculated by dividing the total leaf area on the image by the total image area.

Since plant appearance and image quality varies at different time points, single predictions are sometimes prone to errors as the result of poor image quality or over/under segmenting. With this in mind, we offer dashboard users the option to view daily average prediction metrics as well as average predictions over multiple days to provide stabilized prediction results. The image data and leaf predictions are also connected with the operational data based on timestamps, so that plant growth data and growing parameters of the plants can be linked. The operational data consists of 8 different parameters, which can be difficult to read altogether, so we separated the 8 parameters into three categories: air, water, and lighting. This corresponds to the three essential aspects of plant growth and is easier for users to locate specific parameters when using the dashboard. By connecting the image data and operational data with model predictions, this dashboard can help the user to understand how the plant responds to different growing parameters and used to help develop future growing recipes to optimize plant yields.

The performance of our leaf prediction model is very promising on both tomato and basil plants and it is able to produce relatively accurate results even with low quality image data. Under artificial lighting, the plant leaves can have a colored hue and light reflection, and under no artificial lighting, the image can be too dark. Our model is flexible enough to produce very good results under both extremes given the quality of the data.

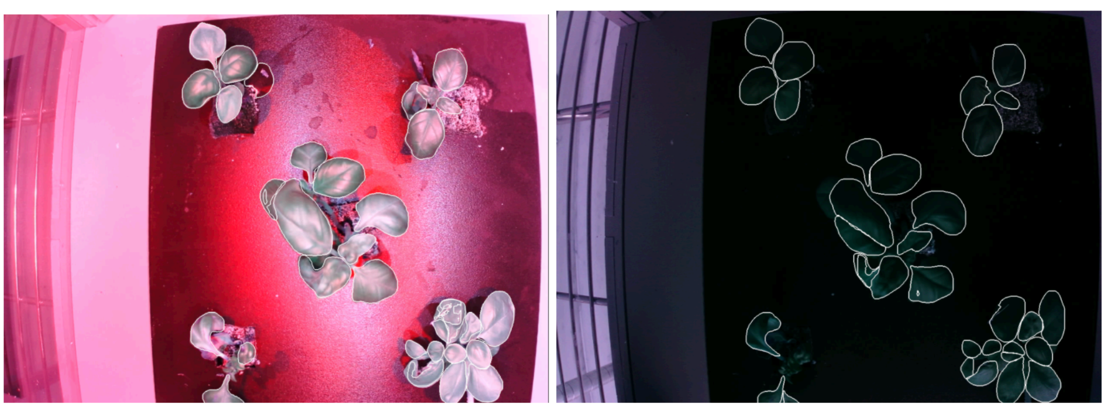

Figure 7.2 Predictions under colored lighting with reflection (leaf) and no artificial lighting

7 Project Challenges and Improvements

There were several challenges we encountered while developing the predictive model. Most of the challenges pertained to the image data and could be reasonably mitigated in future experiments. Therefore, we have listed several of the challenges and our associated comments/recommendation for how to deal with them in future experiments:

- The leaf count gets impacted by overlapping leaves from the top-view of the container. If we had images from side views we could get a more accurate estimate of the count.
- In basil control-2 images, we notice that the leaves are trimmed while they are growing, and some plants are replaced mid-experiment, impacting both the area and number of leaves.

This should be avoided during the experiment to maintain data consistency.

- We do not have a perception of depth as leaves closer to the camera seem larger than ones further away. The use of LIDAR measurements in combination with the images could help establish depth measurements.
- The lighting conditions vary significantly between images. Though our predictions are still able to predict in low light, we should be able to achieve higher precision if the feature maps the model has to learn are less variable in terms of lighting.
- There are plenty of blurred images that make object detection even more challenging since it gets hard to identify the edges. Also, the focal point of the camera varies across images, sometimes focusing on a specific plant, the floor, or the highest leaf.
- To ensure fidelity of the data, all images need to be taken at a more appropriate frequency, given that the current temporal aspect is highly erratic.
- Some of the plants, especially the tomato ones, grow quite tall and camouflage the camera. This severely limits the ability to measure much at all. Reducing the number of plants in the experiment, with an adjustable camera should help mitigate this issue.
- To make valid causal inferences, we require significantly more recipes to measure the impact of individual features in the operational data accurately.

8 Future and Conclusion

Going forward, the segmentation model can be implemented on a camera with a GPU so that the predictions can be sent to a live dashboard in real-time via an integrated pipeline. This would result in a fully developed IoT solution that could be deployed to researchers in the field, and possibly production ready farmers. The model will likely need to be retrained periodically in order to extend the prediction to different kinds of plants (other than tomato and basil). This could be implemented by selecting images with prediction masks identified correctly and retraining on those images. Also, the number of experiments can be increased to establish causal inference on the experimental parameters. Hypothesis testing with appropriate multiple comparison adjustments can then be applied to obtain the best recipe given the experiment's environmental conditions. These additions build off of the proof of concept that we established and represent significant future value for the project.

Overall, we demonstrate an excellent proof of concept model to effectively identify plant yield characteristics through object detection and image segmentation. The results of the model are impressive in terms of the average precision scores and visual predictions despite the many challenges with the datasets. We expect the dashboard to be an effective tool that the partner's can use to further develop the project.

9 Appendix

9.1 Missing data analysis

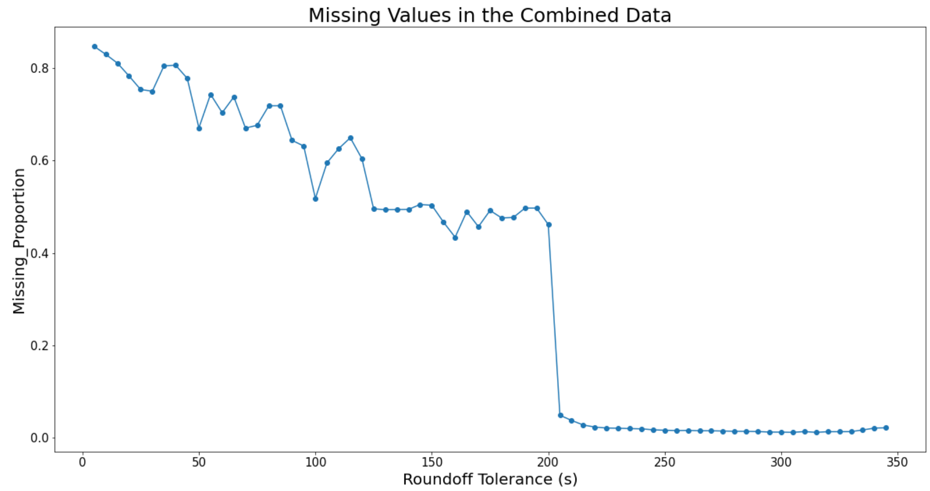

Figure 5.1.1 Missing proportion in merged data vs timescale roundoff tolerance

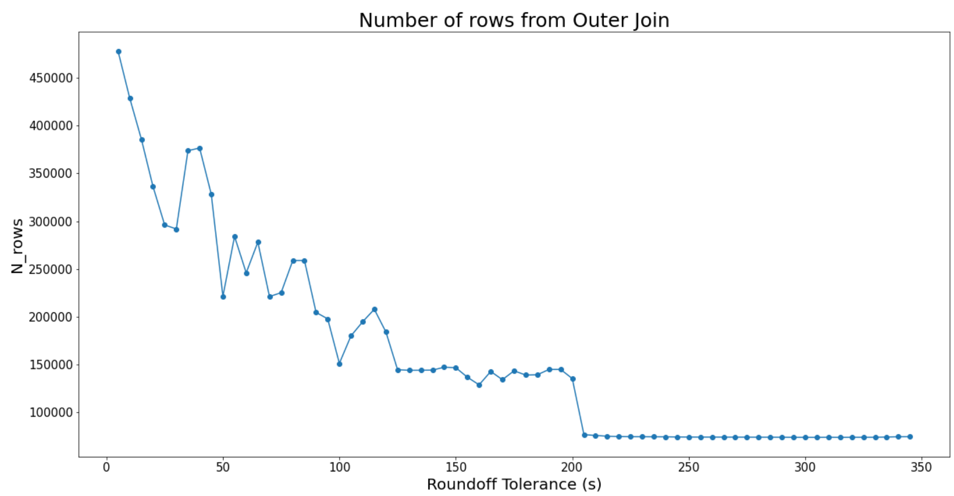

Figure 5.1.2 Number of rows in merged data vs timescale roundoff tolerance

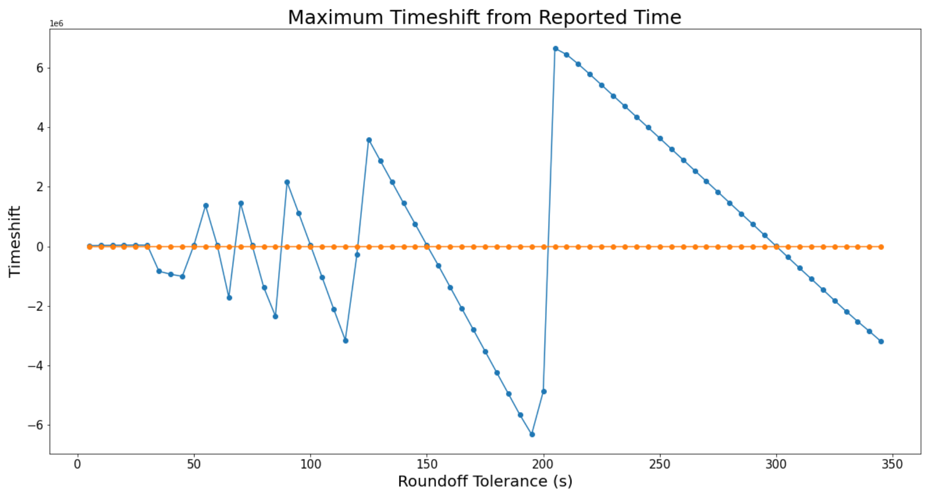

Figure 5.1.3 Timeshift in merged data vs timescale roundoff tolerance

9.2 Steps to Annotate Images

- Open www.makesense.ai
- Click on get started
- Drop all your images
- Click on object detection
- Create label list by clicking the + sign and name it leaf
- Choose 'I'm going on my own' for models
- Choose polygons on the right
- Start drawing!
- Once you're done click on Export labels and choose the polygon option.
- Export as VGG JSON format

References

[1] Dionysia A. Fasoula, Ioannis M. Ioannides, and Michalis Omirou. Phenotyping and plant breeding: Overcoming the barriers. *Frontiers in Plant Science*, 10:1713, 2020.

[2] Sarathi M. Weraduwage, Jin Chen, Fransisca C. Anozie, Alejandro Morales, Sean E. Weise, and Thomas D. Sharkey. The relationship between leaf area growth and biomass accumulation in arabidopsis thaliana. *Frontiers in Plant Science*, 6:167, 2015.

[3] Daniel Ward and Peyman Moghadam. Synthetic arabidopsis dataset. v4. csiro. data collection, 2018.

[4] Massimo Minervini, Andreas Fischbach, Hanno Scharr, and Sotirios A. Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recognition Letters*, pages –, 2015.

[5] Jonathan Bell and Hannah M. Dee. Aberystwyth leaf evaluation dataset, November 2016.