Increasing yield of crop production in Greenhouses using Artificial Intelligence

Group 5

Capstone Project

Project Sponsor: Prof. Manjari Maheshwari

Group Members:

- Odunayo Agagu – 0810500
- Yellai Aditya Venkat Murty – 0801899
- Suman Sunny Rathod – 0815907
- Khushi Patel – 0790434
- Naveen Kumar Reddy Sama – 0804315

PROBLEM STATEMENT

The agricultural industry faces the major obstacle of controlling crop diseases and increasing crop yield in the midst of growing global population and declining resources. In Canada, where topography and weather present special difficulties, greenhouse farming becomes a key strategy for boosting productivity, guaranteeing year-round supply, and lowering dependency on imports. Greenhouses are the major source of growing crops since they provide controlled environments that reduce the risks associated with weather. However, there are a number of serious issues that need to be addressed due to the ongoing threat of disease outbreaks in important greenhouse crops, including crop damage, resource inefficiency, environmental degradation, and financial losses from waste and excess production.

PROJECT MOTIVATION

This project's main motivation is the urgent need to improve agricultural sustainability and yield, especially in light of growing population and resource depletion. AI is transforming a number of industries for greater ease and efficiency, so it is essential that the agricultural sector uses these technologies to handle the challenges of contemporary farming. To help farmers, policymakers, and other agricultural sector stakeholders for making well-informed decisions that improve output, minimize the adverse impacts on the environment, and provide food security in the face of evolving global challenges is the ultimate goal.

PROJECT GOAL

This project aims to enhance crop production in Canadian greenhouses by leveraging artificial intelligence (AI) to identify common diseases and optimize yield. The motivation behind integrating agriculture and machine learning is driven by the critical need for improved yield and sustainability amid growing population and limited resources. Greenhouse farming in Canada is particularly significant due to its role in reducing reliance on imports and ensuring year-round supply, supported by advanced automation systems.

Utilizing the data collected from these systems, the project demonstrates how AI can be used to train models for crop yield optimization and disease detection. The first phase focuses on predicting crop yield through deep learning models, employing feature extraction and prediction methods. In the second phase, deep neural networks are trained on labeled images of crops and diseases to detect and identify lesions, targeting major greenhouse crops such as tomatoes, lettuce, bell peppers, cucumbers, and strawberries. This selection is based on data analysis from Stats Canada. Overall, the project aims to showcase the potential of AI in revolutionizing greenhouse farming techniques in Canada.

PROJECT OBJECTIVES

The primary objectives of the initial phase of research include:

- Research on the historical data and gather the relevant data that can be used for performing EDA and data analysis.
- Analyze the type of crop production, quantity of crop growth and consumptions in greenhouses across Canada.
- Perform the analysis on the major crops grown in Canada, their production, the regions producing these major crops, the production amount, etc.

- Perform the analysis on the available greenhouses in Canada, the regions with major operational greenhouses and the estimated geographical areas in which they have been spread, greenhouses' production, sales and resales, etc.
- Perform thorough Exploratory Data Analysis based on the gathered data and provide actionable recommendations.

The primary objectives of the next phase include:

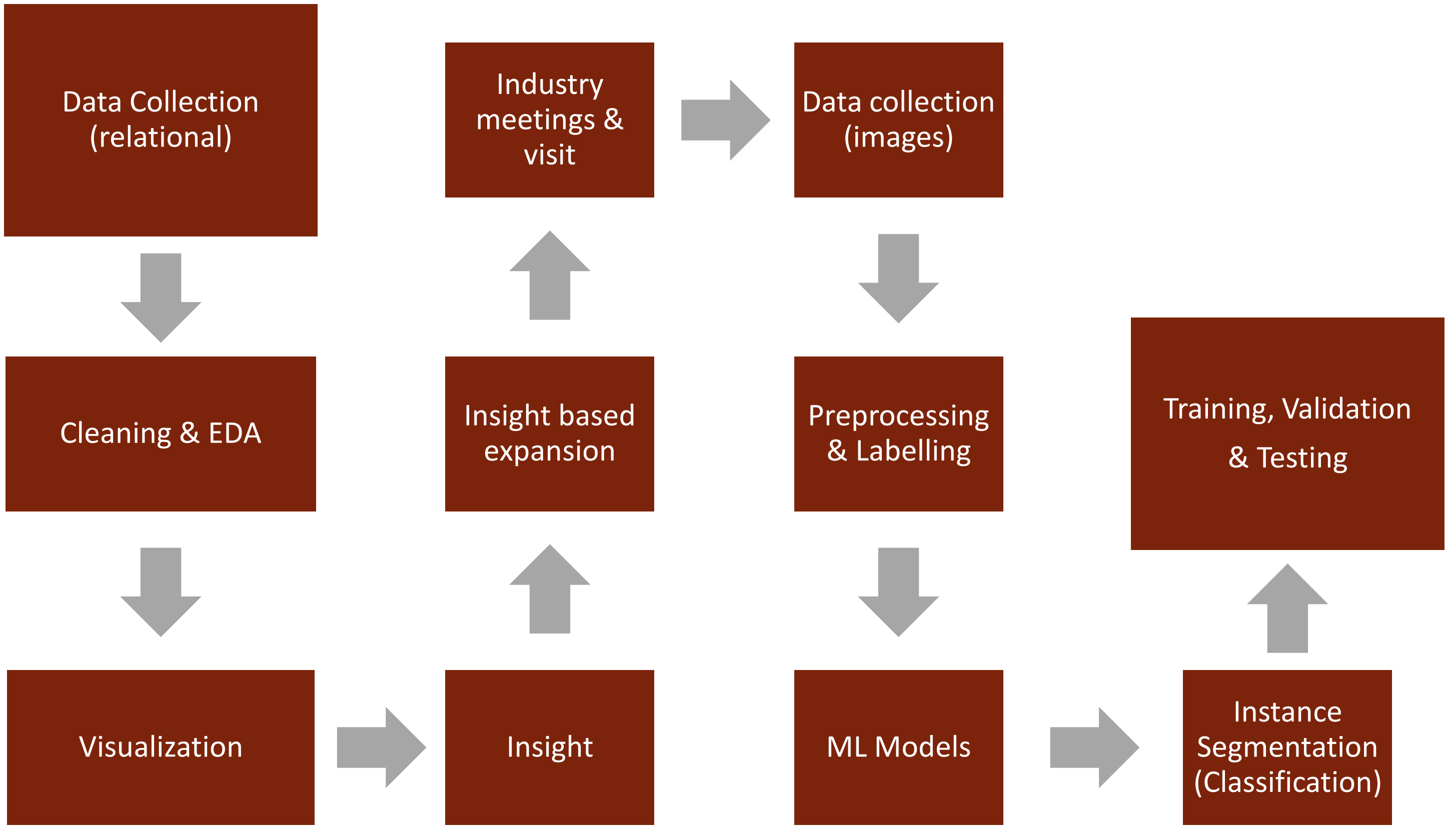
- Research on common diseases on selected crops.
- Gather the images of healthy crops and diseased crops for the selected 5 crops (tomatoes, lettuce, bell peppers, cucumbers, and strawberries).
- For each of the five crops, classify the images as healthy crops, unhealthy crops. For unhealthy crops, further classify based on the specific diseases. Prepare the image datasets based on the classification.
- Perform image preprocessing through image resizing, renaming and labelling based on the classes.
- Prepare the appropriate model, train and test the model and provide actionable recommendations.

DATASET INFORMATION

https://www150.statcan.gc.ca
https://universe.roboflow.com/search

- The relational dataset used for this phase-1 is sourced from statistics Canada with has public data, as noted from the provided links. Data has been collected by government from the period of January 2007 to January 2022, in Excel format with dataset size ranging from 17KB to 4.80MB.
- The dataset is a collection of different datasets which cover different aspects of the topic.
  1. Estimates of greenhouse total area and months of operation.
  2. Total value of greenhouse products.
  3. Estimated areas, yield, production, average farm price and total farm value of principal field crops, in metric and imperial units.
- The image dataset for phase-2 is a collection of different image datasets bucketed on the basis of healthy crops (healthy fruit/veggie and healthy leaves) and the various types of diseases among these crops (diseases fruit and diseased leaves). The images gathered from the various sources such as roboflow, google, etc.

METHODOLOGY



Data Collection (relational)

Cleaning & EDA

Visualization

Insight

Insight based expansion

Industry meetings & visit

Data collection (images)

Preprocessing & Labelling

ML Models

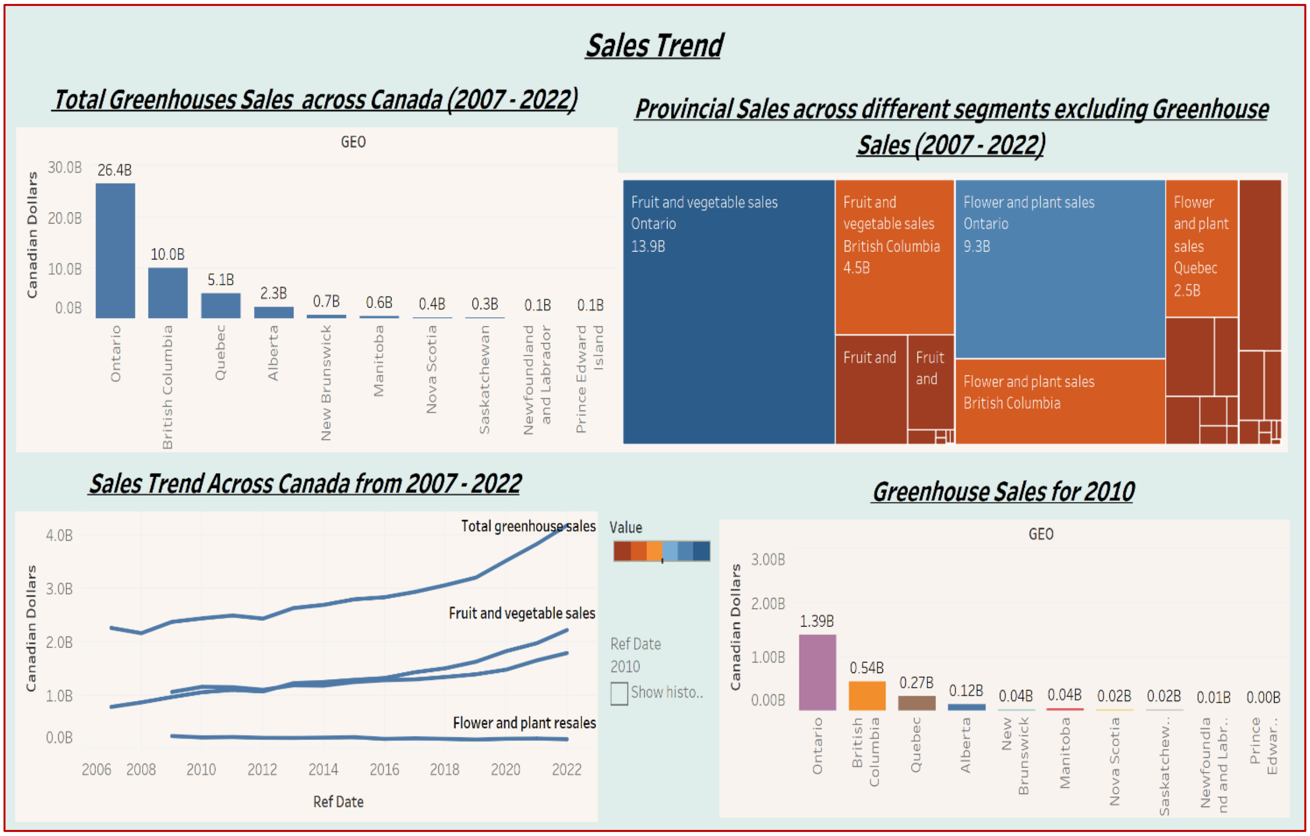
Instance Segmentation (Classification)

Training, Validation & Testing

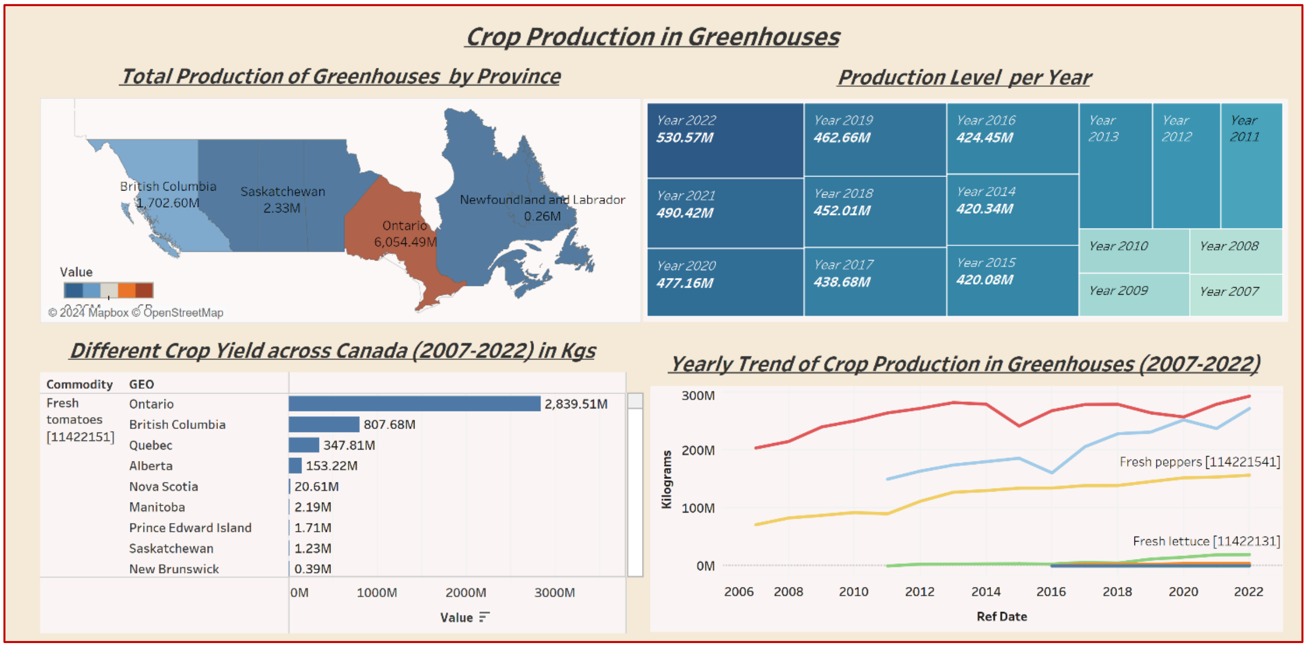
VISUALIZATION (PHASE-1)

Some of the important dashboards created for phase-1

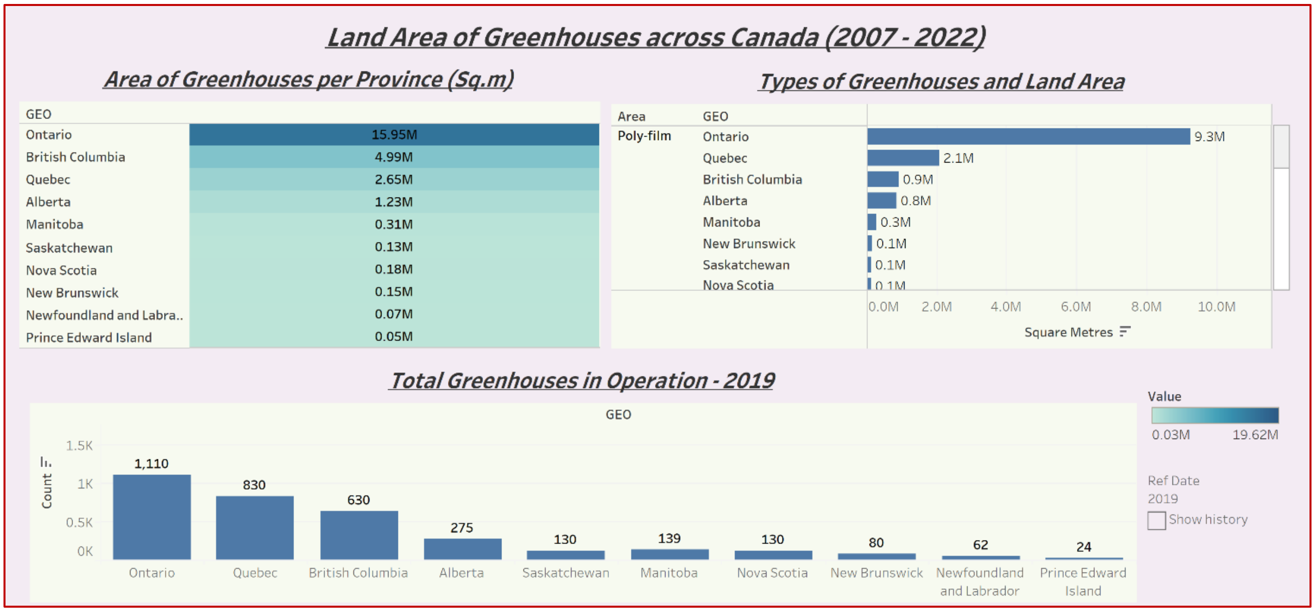
Dashboard 1: Greenhouses' Sales across Canadian provinces



Dashboard 2: Greenhouses' Crop production across Canadian provinces



Dashboard 3: Types and land area covered by Greenhouses across Canadian provinces



Land Area of Greenhouses across Canada (2007 - 2022)

Area of Greenhouses per Province (Sq.m)

GEO	
Ontario	15.95M
British Columbia	4.99M
Quebec	2.65M
Alberta	1.23M
Manitoba	0.31M
Saskatchewan	0.13M
Nova Scotia	0.18M
New Brunswick	0.15M
Newfoundland and Labra..	0.07M
Prince Edward Island	0.05M

KEY INSIGHTS

- The three provinces that account for the highest greenhouses (in terms of number, total area, production and sales) in Canada are Ontario, British Columbia and Quebec.
- The major commodities produced in greenhouses are tomatoes, cucumbers and peppers. Followed by lettuce, strawberries and eggplants.
- The greenhouses are either made of poly-film, glass or rigid plastic. Poly-film made greenhouse cover the largest land area

GREENHOUSE VISIT INSIGHT

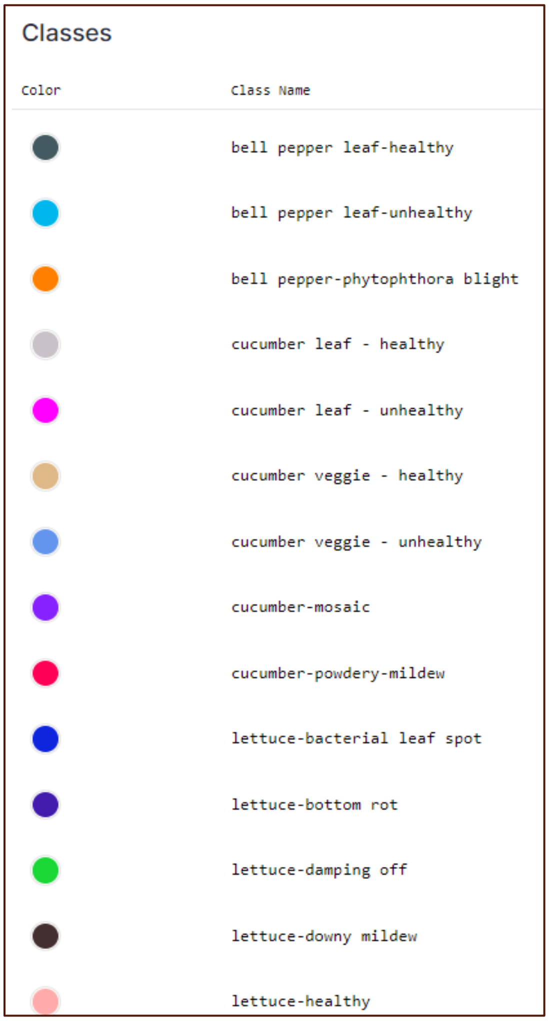
- Two greenhouse visits (DiCiocco – Sonny Fresh Farms and Nature Fresh Farms) were carried out by the group members and Prof. Manjari Maheshwari.
- During the visits, very helpful information has been extracted such as what are the most common diseases found in the selected crops, what problems the greenhouses are facing, clarity on new methods to achieve goals and how to adopt a better strategy.
- After a keen discussion with the greenhouses, this modelling can help reduce the disease detection time & improve the quality of produce.
- Image classification can help set the disease type into different stages helping the greenhouses to act differently by sending them to the lab for examination and understand which has to be treated in what way.
- Overall, will boost efficiency, reduce labor and time consumed, and can act faster to reduce damage due to disease leading to crop yield optimization.

IMAGE DATA COLLECTION (PHASE-2)

DATA COLLECTION AND CLASSIFICATION

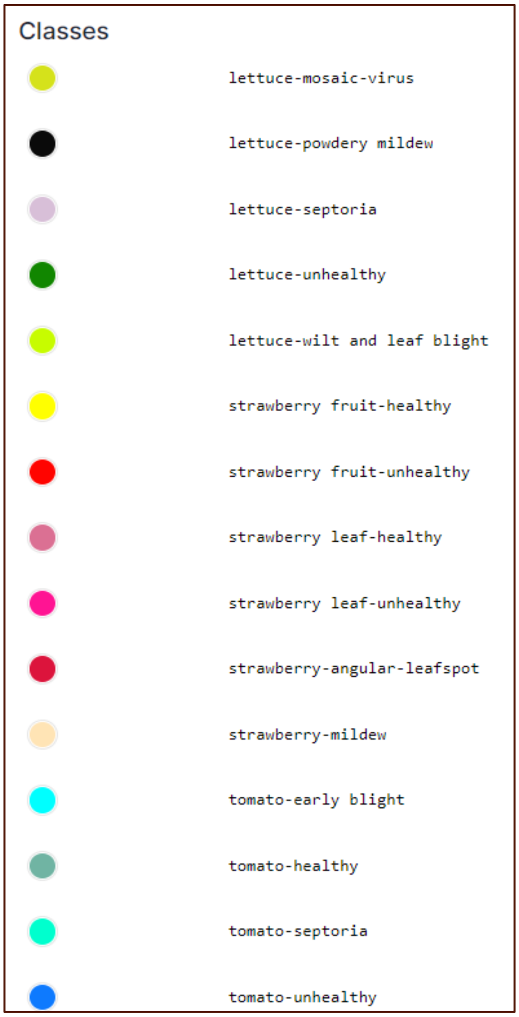
The phase-2 image dataset comprises various collections of images categorized into healthy crops (including both fruits/vegetables and leaves) and different types of diseases affecting these crops (both diseased fruits/vegetables and leaves). These images have been sourced from diverse platforms such as Roboflow, Google, and other relevant sources. Below are the various categories formed and the images are bucketed to these classes based on the appropriate match:

1. Strawberry healthy fruit, Strawberry healthy leaves, Strawberry unhealthy fruit, Strawberry unhealthy leaves, Strawberry mildew, Strawberry angular leafspot
2. Lettuce healthy, Lettuce wilt and leaf blight, Lettuce septoria, Lettuce powdery mildew, Lettuce mosaic virus, Lettuce downy mildew, Lettuce bacterial leaf spot
3. Cucumber healthy veggie, Cucumber unhealthy veggie, Cucumber healthy leaves, Cucumber unhealthy leaves, Cucumber mosaic
4. Bell pepper healthy leaves, Bell pepper unhealthy leaves, Bell pepper phytophthora
5. Tomato healthy, Tomato unhealthy, Tomato healthy, Tomato unhealthy, Tomato septoria, Tomato early blight



Classes

Color	Class Name
●	bell pepper leaf-healthy
●	bell pepper leaf-unhealthy
●	bell pepper-phytophthora blight
●	cucumber leaf - healthy
●	cucumber leaf - unhealthy
●	cucumber veggie - healthy
●	cucumber veggie - unhealthy
●	cucumber-mosaic
●	cucumber-powdery-mildew
●	lettuce-bacterial leaf spot
●	lettuce-bottom rot
●	lettuce-damping off
●	lettuce-downy mildew
●	lettuce-healthy



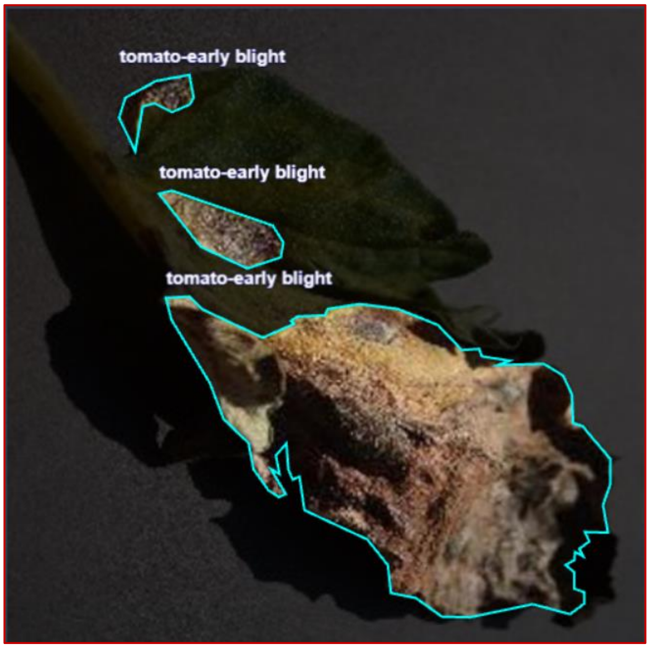
Classes

●	lettuce-mosaic-virus
●	lettuce-powdery mildew
●	lettuce-septoria
●	lettuce-unhealthy
●	lettuce-wilt and leaf blight
●	strawberry fruit-healthy
●	strawberry fruit-unhealthy
●	strawberry leaf-healthy
●	strawberry leaf-unhealthy
●	strawberry-angular-leafspot
●	strawberry-mildew
●	tomato-early blight
●	tomato-healthy
●	tomato-septoria
●	tomato-unhealthy

IMAGE LABELLING, PREPROCESSING AND AUGMENTATION

The process of annotating an image entails adding labels or metadata to the image to provide more details about its contents. In particular, for tasks like object detection, image classification, and segmentation, this procedure is crucial for training machine learning models.

The process of image annotation for healthy crops entails locating and labeling areas of an image that represent leaves and fruits or vegetables that are in good condition. Similarly, the infected areas of the crops are located and labelled according the category of disease. Both healthy and unhealthy annotations are done based on the defined classes on Roboflow platform.



For image preprocessing, various operations like image resizing, renaming, consist file formatting and normalization is done. It creates a uniform format for all filenames inside a specified directory ("img {count}.jpg"), where {count} is a sequential number that starts at 0. This feature makes file naming consistent and streamlines dataset management, making it simpler to handle and process data in later phases.

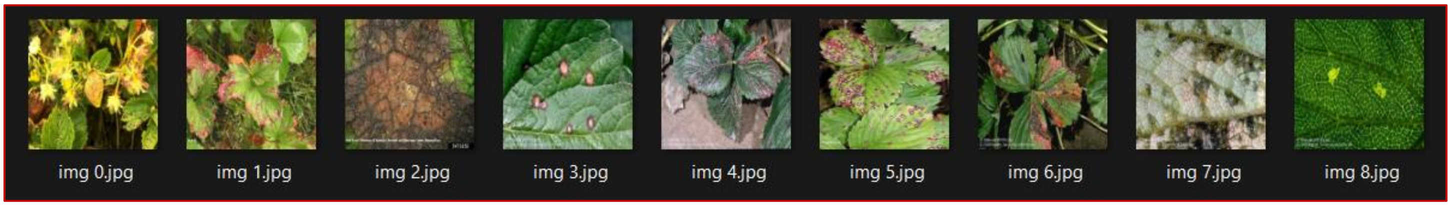
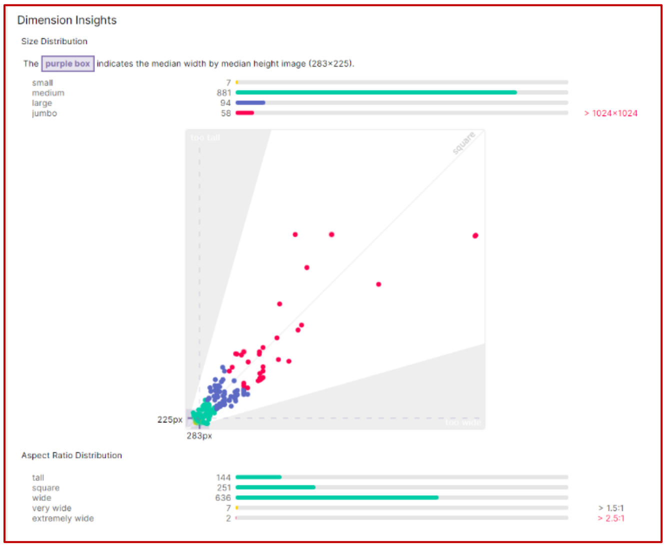


Image resize and rename output, Dimension 400*400 px, Naming convention: img [0,1,2,…so on].jpg

The user-supplied size is used by the script to resize all of the images within a folder to the given dimensions (width x height). This is especially helpful for normalizing image sizes within a dataset, which is a common need for input data to guarantee consistent dimensions in machine learning models. Numerous augmentation options are available with Roboflow, such as brightness adjustment, rotation, flipping, and zooming. We have tried various augmentation methods to boost model robustness and increase dataset variability.



Dimension Insights

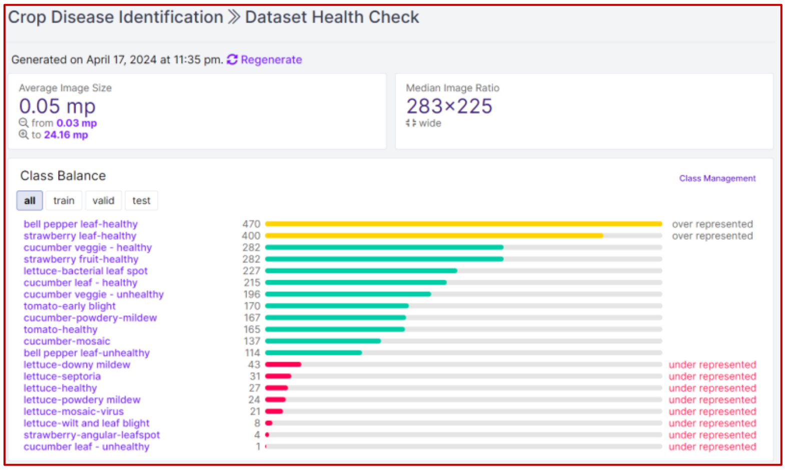
Size Distribution

The purple box indicates the median width by median height image (283×225).

small	7	
medium	881	
large	94	
jumbo	58	> 1024×1024

Aspect Ratio Distribution

tall	144	
square	251	
wide	636	
very wide	7	> 1.5:1
extremely wide	2	> 2.5:1



Crop Disease Identification » Dataset Health Check

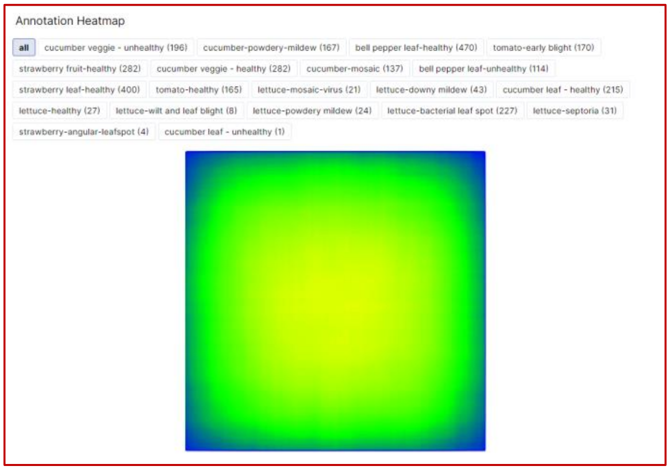
Generated on April 17, 2024 at 11:35 pm. Regenerate

Average Image Size
0.05 mp
from 0.03 mp
to 24.16 mp

Median Image Ratio
283×225
wide

Class Balance

bell pepper leaf-healthy	470	over represented
strawberry leaf-healthy	400	over represented
cucumber veggie - healthy	282	
strawberry fruit-healthy	282	
lettuce-bacterial leaf spot	227	
cucumber leaf - healthy	215	
cucumber veggie - unhealthy	196	
tomato-early blight	170	
cucumber-powdery-mildew	167	
tomato-healthy	165	
cucumber-mosaic	137	
bell pepper leaf-unhealthy	114	
lettuce-downy mildew	43	under represented
lettuce-septoria	31	under represented
lettuce-healthy	27	under represented
lettuce-powdery mildew	24	under represented
lettuce-mosaic-virus	21	under represented
lettuce-wilt and leaf blight	8	under represented
strawberry-angular-leafspot	4	under represented
cucumber leaf - unhealthy	1	under represented



Annotation Heatmap

all
cucumber veggie - unhealthy (196)
cucumber-powdery-mildew (167)
bell pepper leaf-healthy (470)
tomato-early blight (170)
strawberry fruit-healthy (282)
cucumber veggie - healthy (282)
cucumber-mosaic (137)
bell pepper leaf-unhealthy (114)
strawberry leaf-healthy (400)
tomato-healthy (165)
lettuce-mosaic-virus (21)
lettuce-downy mildew (43)
cucumber leaf - healthy (215)
lettuce-healthy (27)
lettuce-wilt and leaf blight (8)
lettuce-powdery mildew (24)
lettuce-bacterial leaf spot (227)
lettuce-septoria (31)
strawberry-angular-leafspot (4)
cucumber leaf - unhealthy (1)

MODELLING, TRAINING AND TESTING

Model training and testing is done on both ways i.e. manually through python scripting for binary classification and modelling as well as through Roboflow platform for disease detection and modelling to check and compare the accuracy

CNN, VGG16 AND VGG19

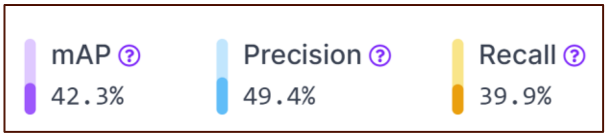
CNN model has ability to analyze spatial patterns in image data which is important for disease detection. VGG16 and VGG19 have deep architecture and proven effectiveness in image classification tasks that enhances accuracy. A python script is prepared which performs binary classification of the image dataset into healthy and unhealthy crops. Once the image preprocessing is done with image resizing, labelling and renaming, the model training is performed. We have created CNN and VGG16 models for the above processed bulk images. The model training and testing operations have been performed which gives the following outcomes:

- CNN (Binary Classification) -> Accuracy: 0.97
- VGG16 (Multi Classification) -> Accuracy: 0.33
- VGG19 (Multi Classification) -> Accuracy: 0.47

YOLO V8

Yolo v8 is the most suitable for efficient classification involving segmentation, identifying anomalies. It has enormous computational and augmentation capacity. Yolo v8 model training and testing is done on Roboflow platform. Yolov8 Instance Segmentation is done which is beyond simple image classification – not only identifies objects but also precisely labels individual instance of object within image. It distinguishes between different instances of same class. For example: not only class dogs in image also segment different dog breeds. It detailed object recognition, precise identification using enhanced visual insights. The following is the result of Yolov8 Instance Segmentation.

Model Type: Roboflow 3.0 Instance Segmentation (Fast)
Checkpoint: crop-disease-identification/8
Version: 9
Data Split Ration: 70:20:10



mAP	Precision	Recall
42.3%	49.4%	39.9%

USE CASES

1. **Early Detection:** Early detection due to automation enables timely and targeted action
   - Automation makes early detection possible, enabling prompt and accurate intervention.
   - Farmers can minimize crop losses and maximize yield by taking targeted actions, like applying pesticides or putting preventive measures in place, as soon as they notice diseases or abnormalities in their crops.
2. **Resource Management:** Better resource allocation including water, nutrients, pesticides and human resource
   - By maximizing the use of water, fertilizers, pesticides, and labor, automation in agriculture improves the distribution of resources.
   - Farmers can effectively allocate resources based on crop needs, environmental conditions, and operational requirements through real-time monitoring and data-driven insights, leading to increased productivity and sustainability.
3. **Yield Optimization:** Approximate more efficiently on crop yields and plan accordingly
   - Accurate crop yield estimation and cultivation strategy planning are made easier with the help of automation.
   - Farmers can more accurately predict crop yields, anticipate market demands, and optimize planting dates and cultivation techniques to maximize production and profitability by utilizing data analytics and machine learning algorithms.
4. **Quality Control:** Less disease and anomalies ensures better quality
   - Automation lowers the likelihood of crop diseases and anomalies, which improves quality control.
   - Farmers can reduce the risk of crop contamination, guarantee product safety, and uphold high standards of quality throughout the production process by employing early detection systems and routinely monitoring crop health.
   - This will ultimately improve consumer satisfaction and competitiveness in the market.

CHALLENGES

1. **Data Collection**
   - Required variables in a dataset crucial for crop yield prediction and/or optimization (soil nutrients, pesticides, and insecticides) were confidential.
   - Multiple attempts from multiple channels did not yield result for what was required.
   - Images collected had quality issues that had to be tackled efficiently.
2. **Computational Limitations**
   - Image dataset after augmentation resulted in huge dataset requiring proper pipeline and memory.
   - Code prepared to be run on python/google colab could not run all required epochs due to computational limit.
   - Less epochs created poor accuracy and generalization.
3. **Industry Knowledge**
   - Considerable amount of time was spent on research and meeting industry experts including site visit to navigate through best practices and requirements to navigate through the project.

CONCLUSION

In essence, our project signifies a significant advancement in merging machine learning with agriculture, particularly in the realm of greenhouse farming within Canada. With the aim of enhancing sustainability and yield amid population growth and resource constraints, we're exploring the potential of artificial intelligence for maximizing crop management and disease detection.

Modern greenhouses generate vast amounts of data through advanced systems like irrigation and climate control. Leveraging state-of-the-art machine learning techniques, our project harnesses this data abundance to develop two pivotal models: a disease detection system and a crop management optimizer.

While the technological advancements are noteworthy, our focus extends to practical applications in sustainable agriculture. By integrating machine learning into greenhouse farming practices, we contribute to resource optimization, reduce reliance on imports, and ensure a consistent supply of produce year-round. The scalability and adaptability of our models hold promise for the agricultural industry, marking a transformative stride towards a more resilient and productive farming landscape.

REFERENCES

1. https://www.sciencedirect.com/science/article/pii/S0168169920302301
2. https://www.nature.com/articles/s41598-023-42843-2
3. https://www.sciencedirect.com/science/article/pii/S0168169920302301#:~:text=Several%20machine%20learning%20algorithms%20have,in%20crop%20yield%20prediction%20studies
4. https://docs.roboflow.com/datasets/image-augmentation
5. https://www.statcan.gc.ca/en/data-science/network/greenhouse-detection
6. https://docs.ultralytics.com/#yolo-licenses-how-is-ultralytics-yolo-licensed
7. https://yolov8.com/

PROJECT LINKS

- Github link for the python script:
  https://github.com/Adivenkat28/Capstone-Project---Increasing-Crop-Yield-in-Greenhouses-using-AI

- Youtube link for the project presentation:
  https://www.youtube.com/watch?v=3QVfAPOaegI

- Roboflow link:
  https://app.roboflow.com/capstone-project-2wtea