Cyberinfrastructure for Machine Learning Applications in Agriculture: Experiences, Analysis, and Vision

Lucas Waltz[1], Sushma Katari[1], Chaeun Hong[2], Adit Anup[2], Julian Colbert[2], Anirudh Potlapally[2], Taylor Dill[3], Canaan Porter[2], John Engle[1], Christopher Stewart[2], Hari Subramoni[2], Raghu Machiraju[2], Osler Ortez[3], Laura Lindsey[3], Arnab Nandi[2], Sami Khanal[1]

[1]Department of Food, Agricultural, and Biological Engineering, The Ohio State University, Columbus, Ohio, USA

[2]Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA

[3]Department of Horticulture and Crop Science, The Ohio State University, Columbus, Ohio, USA

A paper from the Proceedings of the
16th International Conference on Precision Agriculture
21-24 July 2024
Manhattan, Kansas, United States

Abstract.

Advancements in machine learning algorithms and GPU computational speeds over the last decade have led to remarkable progress in the capabilities of machine learning. This progress has been so much that, in many domains, including agriculture, access to sufficiently diverse and high-quality datasets has become a limiting factor. While many agricultural use cases appear feasible with current compute resources and machine learning algorithms, the lack of software infrastructure for collecting, transmitting, cleaning, labeling, and training datasets is a major hindrance towards developing solutions to address agricultural use cases.

This work aims to share the learnings from collecting a 1 terabyte (TB) multimodal dataset from three agricultural research locations across Ohio during the 2023 growing season. The dataset includes Unmanned Aerial System (UAS) imagery (RGB and multispectral), and soil and climate sensors for the state's two largest crops: corn and soybeans.

Keywords.

Precision Agriculture, Multimodal Data, Machine Learning, Unmanned Aerial Systems, Crop

1. Introduction

In recent years, there has been a surge in interest across various domains in leveraging machine learning techniques to tackle complex, long-standing challenges. While technically a subfield of artificial intelligence, machine learning is often used interchangeably. Machine learning techniques have been successfully applied across multiple domains, achieving remarkable milestones since 2012, such as AlexNet's victory in the ImageNet competition (Krizhevsky et al., 2012) and the introduction of the transformer architecture in 2017 (Vaswani et al., 2017). These milestones have propelled machine learning into unprecedented popularity.

This growing recognition of machine learning's potential has led experts in various domains to explore its applicability to their most daunting challenges. However, while the latest machine learning approaches are powerful, they perform best with extensive, high-quality datasets which are often expensive and labor-intensive to collect. Existing public datasets in agriculture, though useful, are often not adequate to harness the latest advances in model complexity and compute resources. The process of collecting and processing agricultural data for machine learning faces numerous challenges, including sensor failures, data pipelines, and data privacy concerns.

The notion that data harnessed from agriculture, coupled with the latest advancements in machine learning, can significantly enhance both the profitability and sustainability of farming practices is not novel. The agricultural industry's dominant players in seed, chemicals, fertilizer, and equipment have invested heavily in farm management information systems (FMIS). While much of these systems are focused on providing accurate records of past events, falling into the realm of descriptive analytics, there are increasing efforts to include predictive and prescriptive analytics into these software platforms. For example, Microsoft's Farmbeats project (Kapetanovic et al. 2004), launched in 2014, focuses on data-driven farming by integrating various data sources, like field sensors and UAS, to provide insightful analytics through computer vision and machine learning algorithms. It establishes an end-to-end IoT infrastructure for efficient data collection and utilizes TV white spaces for transmitting data to computing centers, thus enabling advanced data analytics, and, in turn, empowering farmers to enhance productivity and sustainability (Chandra et al. 2022). Another example is Mineral, originating from Google/Alphabet's X facility, which claims to have surveyed 10% of the world's farmland and developed 80 machine-learning models to boost production and mitigate agriculture's impact on the environment (Burwood-Taylor 2023).

The creation of large-scale, high-quality multimodal datasets, carefully curated and made ready for machine learning applications, can significantly advance predictive and prescriptive analytics in agriculture. These datasets encompass spatial, spectral, and temporal dimensions. Spatial intensity refers to ground sampling distance (GSD) measured in centimeters or meters per pixel. Spectral resolution refers to the number of wavelength intervals, while temporal denotes the frequency of data collection. Gadiraju et al. (2020) demonstrated a 60% reduction in prediction error by using a multimodal deep-learning approach that leveraged spatial, spectral, and temporal data characteristics to identify crop types. This involved integrating a Convolutional Neural Network (CNN), often used for analyzing images, with spatially intensive data and a Long Short-Term Memory network (LSTM), often used to analyze text corpora, with temporally intensive data. Presently, there is a growing research focus on data-driven agriculture systems that involve deploying a diverse array of sensors and the Internet of Things (IoT) for vast data generation and Big Data Analytics on these datasets (BDA) (Ur Rehman et al. 2019). This trend holds promise for automating farming decisions. Furthermore, edge-cloud architectures (Taheri et al. 2023) can enhance real-time decision-making by hastening data processing.

In addition to the importance of data quantity, it is crucial to consider data quality prior to processing and incorporating data into model pipelines. The utilization of data quality indicators, such as data source, collection time, and environmental conditions, can serve to flag datasets with undesirable traits (Wang et al. 1993). These considerations underpin the critical role of data quality in agriculture's data-intensive domains.

This paper outlines our journey in constructing a dataset tailored for specific agricultural use cases and outlines a vision for the necessary software and hardware infrastructure or the cyberinfrastructure (CI). This CI aims to facilitate the collection and processing of agricultural data at scale, enabling the training of Artificial Intelligence (AI) models that are ultimately used for the benefit of farmers and other stakeholders in agriculture.

2. Vision

It appears that many agricultural use cases now appear to be within the capabilities of current compute resources and AI models. However, the lack of CI dedicated to the collection, transmission, cleaning, exploration, labeling, and training of the datasets, along with the challenges of deploying these solutions onto edge and intelligent sensing devices for inference are a major hindrance towards the development of solutions to address these use cases.

Given the ongoing advancements in the AI community at large and the focused efforts within both agricultural industry and academia, we advocate a vision to build publicly available agricultural datasets and the development of associated open-source AI-centric CI. This CI would support the tools and resources necessary for the collection, transmission, cleaning, exploration, labeling, training, and inference of these datasets.

A vibrant open-source community focused on cyberinfrastructure and datasets for AI applications in agriculture has many positive benefits including:

1. Amplifies the efforts of agricultural researchers through reducing the time needed for building and debugging data pipelines, ultimately increasing the quality and quantity of their output and their extension efforts to farmers.
2. Connects computer science researchers with meaningful prevailing problems in the agricultural domain.
3. Lowers the capital requirements for startups to get to product market fit for AI based products and services in agriculture by leveraging open-source software and datasets.

While there are increasing numbers of companies that provide CI to support AI initiatives in general, the needs of agriculture are unique and can benefit from CI and datasets that are focused on salient agricultural use cases. There are several reasons for this assertion:

1. There are very few publicly available datasets of sufficient size and quality focused on agricultural use cases. However, there are many universities worldwide that collect volumes of agricultural data which if put in the right form and format, could form a tremendously valuable resource for AI model training.
2. The pipelines for collecting, transmitting, cleaning, and transforming agricultural data into formats ready for artificial intelligence are labor-intensive and error prone. Furthermore, agricultural researchers in many instances may not possess the data management and software development skills to effectively and efficiently perform these necessary tasks.
3. On-farm and small plot research can be a rich source for training data. However, the approach for splitting the dataset into training, testing, and validation needs to consider the replications in the dataset. Failure to understand this could lead to overfitted models.
4. Commonly used AI models may benefit from modifications to be better suited to agricultural data. For example, while image-based AI models typically use a softmax layer as the final layer for classification, in agriculture it can often be more appropriate to set the last layer of a neural network to a regression output (growth stage and disease severity are two examples).

For the reasons stated above, we believe that AI-amenable infrastructure that leverages the capabilities in the AI community at large while adapting it for common use cases in agriculture has the potential to accelerate the benefits of AI in agriculture. With these benefits in mind, here are several core principles that guide our efforts:

1. Data collection and CI efforts need to co-inform each other and should happen concurrently.

2. The speed for both training and inference are critical measures of value. Speed represents a holistic view that includes latency starting from the point at which data is collected in the field to the point where actionable insights are generated.
3. The CI must incorporate the latest approaches and models from the broader AI community. Vision Transformers (ViT) and semi- and weak supervised labeling techniques are examples.
4. The CI needs to be easy to use, trustworthy, and consider the range of technical proficiencies of various stakeholders in agriculture. It also needs to include interfaces that provide transparency into the "black box" of AI and build confidence in its results. Current efforts are focused on technical and advanced users as shown in **Figure 1**.
5. Existing models and data formats from the AI community at large should be used where appropriate to avoid recreating existing CI components.

Current focus of this initiative

Platform Components	Farmers	Agronomists	Technical Users	Advanced Users
Platform Interaction	Graphical User Interface, minimal setup	Graphical User Interface, more advanced setup	Command Line	Command Line
Data Engineering	NA	Load datasets	Load and manipulate datasets	Build APIs, manipulate and move data
Software Engineering	NA	NA	Beginner / Intermediate	Advanced
AI Models	View Inference Results	View Inference Results, Correct Errors	Training and Inference of AI Models	Modification of model architectures

Figure 1: Mapping of platform components to stakeholders

In this paper, we delineate our experiences in data collection, data processing, model training and user interface.

4. Initial Data Sources, Types, and Use Cases

Figure 2 is a summary diagram that shows initial data types collected and initial use cases.

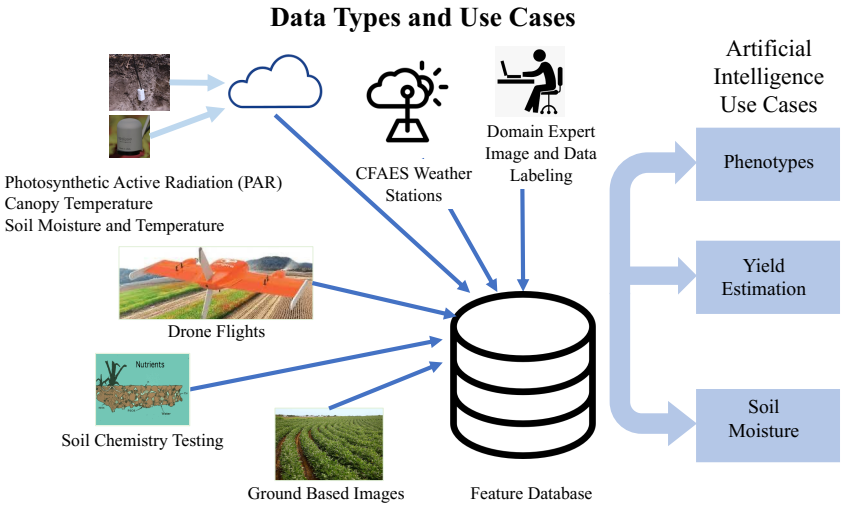

Figure 2: Summary of Initial Data Sources and Use Cases

4.1 Initial Data Sources

The initial data sources underpinning this effort originate from three agricultural research stations geographically dispersed across Ohio and operated by The Ohio State University (OSU). They include Western Agricultural Research Station in Clark County, Northwest Agricultural Research Station in Wood County, and Wooster Campus in Wayne County. Each site included 80 plots for corn and 80 plots for soybean. The experiment was a split-plot randomized complete block design with four replications of each treatment. Main plot factor included five planting dates spaced approximately every two weeks from mid-April to mid-June. The subplot factor for corn was four different hybrids of varying relative maturities (H1 – 100; H2 – 107; H3 – 111; H4 – 115 days) while the subplot factor for soybean included four different seeding rates (S1 - 100,000; S2 - 140,000; S3 - 180,000; S4 - 210,000 seeds per acre). Each replicate included a border plot on both ends of the block to reduce any edge-of-field effects on the measured plots. Furthermore, yield measurements were based on the center two rows (out of four) for corn and the center five rows (out of seven or eight) for soybeans. The research plots were managed according to agronomic best management practices for soybean (Lindsey et al. 2017) and corn (Thomison et al. 2017) outside of the main plot and subplot factors.

Rep 4	PD3 B	PD3 H2 401	PD3 H1 402	PD3 H4 403	PD3 H3 404	PD3 B	PD2 B	PD2 H2 405	PD2 H1 406	PD2 H3 407	PD2 H4 408	PD2 B
Rep 3	PD2 B	PD2 H3 301	PD2 H4 302	PD2 H2 303	PD2 H1 304	PD2 B	PD4 B	PD4 H2 305	PD4 H1 306	PD4 H3 307	PD4 H4 308	PD4 B
Rep2	PD4 B	PD4 H1 201	PD4 H3 202	PD4 H2 203	PD4 H4 204	PD4 B	PD5 B	PD5 H2 205	PD5 H1 206	PD5 H3 207	PD5 H4 208	PD5 B
Rep1	PD1 B	PD1 H1 101	PD1 H2 102	PD1 H3 103	PD1 H4 104	PD1 B	PD2 B	PD2 H1 105	PD2 H2 106	PD2 H3 107	PD2 H4 108	PD2 B

Figure 3: Plot Layout insert from the Western Research site, Corn (PD = Planting Date, H = Hybrid, B = Border Plot).

Each plot was 10 feet wide, configured as either four rows of corn at 30-inch spacing and seven or eight rows of soybeans at 15-inch spacing, spanning approximately 30 feet long or longer at each location. The plots were systematically designated using a 3-digit numbering system: 101-120, 201-220, 301-320, and 401-420. A visual representation of the plot layout for Western Corn is shown in **Figure 3**.

4.2 Initial Data Types

This section summarizes the various data types that have been collected from the 2023 growing season. While a dataset size of less than 1 terabyte (TB) may not be considered extensive according to contemporary standards, it signifies a substantial investment in terms of time and labor in the agricultural domain. The subsequent sections will offer further elaboration on the data.

4.2.1 Unmanned Aerial Systems (UAS) Imagery

The aerial image collection was facilitated using a Wingtra One Unmanned Aerial System (UAS), equipped with both a 42MP RGB camera, the Sony RX1R II, and a Micasense Altum Multi-spectral camera featuring six spectral bands: Red, Green, Blue, Red-edge, Near Infrared, and Thermal Infrared. Flight missions were executed at approximately weekly intervals throughout the entire growing season, culminating with the final flights in mid-October shortly before harvest. This strategy resulted in a total of between 13 and 16 flights per site for each camera. Each flight mission generated hundreds of images covering the corn and soybean plots at each research location.

4.2.2 Structured Soil and Climate Data

In-Situ Soil and Climate Sensing Data

An array of soil sensors was deployed at two depths, specifically at 30 cm and 60 cm, within the

corn and soybean plots for both Planting Date 2 (26-27 April 2023) and Planting Date 4 (25-30 May 2023) at all three research locations. Additionally, one Apogee SQ-521 photosynthetic active radiation (PAR) sensor and one Meter ATMOS 14 weather station were installed at each of these research sites. The weather station collected temperature, relative humidity, vapor pressure, and barometric pressure in the crop canopy.

The data collected by these sensors was aggregated by a total of six data loggers, with two loggers allocated at each research site. These loggers were connected to the Meter Group's Zentra Cloud, a data management and visualization platform. Data visualization was available through user-configurable dashboards on the website and data was also accessible via an application programming interface (API).

Weather Station Data

At each of the research locations, an OSU managed weather station collects precipitation, wind speed, and air temperature at multiple heights, which is accessible at weather.cfaes.osu.edu. In addition, the website also provides calculated daily values such as Growing Degree Days (GDD), a measure of the degrees above 10C of the average temperature each day. The accumulation of GDD over the growing season is widely used in predicting corn growth and development.

Soil Testing Data

On a weekly basis, soil samples were taken from each plot corresponding to the locations of the in-situ soil and climate sensors. These samples were submitted to a soil testing laboratory to measure plant-available nitrogen content, consisting of nitrate and ammonium, as well as CO_2 respiration reported in parts per million (ppm) as an indication of the rate of nitrogen mineralization of organic matter.

Manually Labeled Data

On a weekly basis, site visits were conducted at all three research locations by personnel from the OSU's Department of Horticulture and Crop Science (HCS). These individuals possessed expertise in the classification of corn and soybean growth stages as well as proficiency in assessing disease incidence and quantifying disease severity. Furthermore, ears of corn and soybean plants were collected at harvest for detailed measurements of the components of yield such as kernel rows, kernels per row and kernel weight in corn and seeds per pod, pods per plant, and seed weight in soybeans. The data generated from these site visits will be the labels for several machine learning use cases derived from this data set.

4.3 Initial Use Case – Yield Estimation

Figure 4 shows one agricultural use case for collecting multimodal data. In this example, growth stage is an important part of the dataset for predicting field scale yield along with precipitation, growing degree days, and photosynthetic active radiation during the growing season. UAS imagery is used to predict growth stage and subsequently combined with time-series climate data.

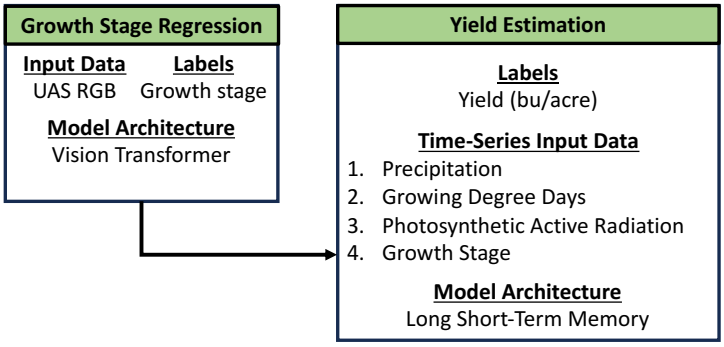

Figure 4. Interconnected AI models for Yield Estimation.

By combining the various data types into interconnected models, a field-scale yield estimate can be obtained during the growing season, which can inform farmers' grain marketing decisions. Furthermore, estimates of yield potential during the growing season can inform the profitability of field treatments such as nitrogen and fungicide applications.

5. Data Readiness Pipelines

The goal is to create data pipelines that will uniformly process the various forms of data in a consistent manner resulting in high data quality. Below, we describe the pipelines for each of the data collected in this effort.

5.1 UAS Imagery Pipeline

The UAS-based data acquisition relied on the use of Secure Digital (SD) cards as the medium for storing captured images during UAV missions and subsequently transferring those images to a OneDrive repository on a laptop. Once the images were transferred, proprietary software was utilized to geotag images from the nearest Continuously Operating Reference Station (CORS) to correct for GPS error during the flight. The pipeline also includes generating a georeferenced orthomosaic for each flight using the geotagged images. **Figure 5** illustrates this pipeline.

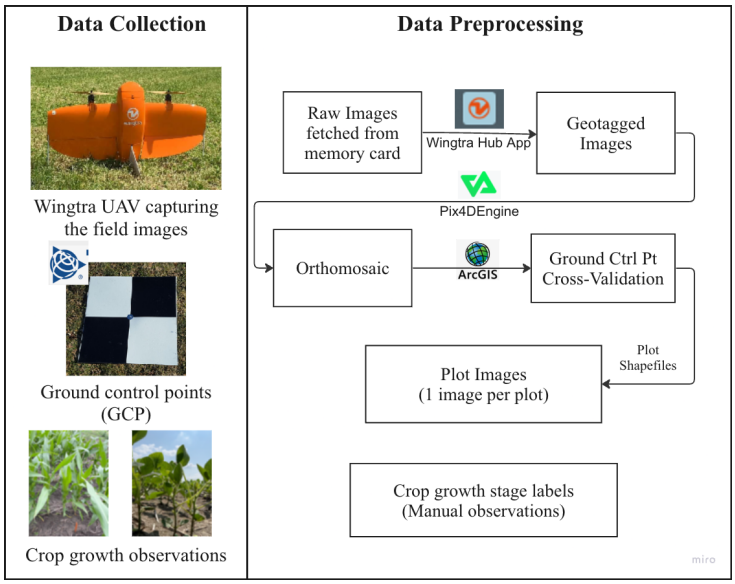

Figure 5. Version 1 of UAS image pipeline

In our study, a systematic arrangement of five Ground Control Points (GCPs) was implemented at each site. Specifically, GCPs were strategically placed at the four vertices of the field and a central point, consistent across every aerial survey. For ensuring accurate spatial referencing, these GCPs were positioned at identical coordinates during consecutive weekly surveys. A high-precision Trimble R8 GNSS receiver, in conjunction with the Trimble TSC3 Controller Data Collector, was employed to obtain precise geographical coordinates of each GCP.

Over the course of summer, a total of 85 flight missions were conducted. These missions included flights across three research locations utilizing two payload sensors, namely RGB and multi-spectral.

Our initial approach was to use Pix4D, a commercial photogrammetry software provider, to generate orthomosaics from each flight. Specifically, Pix4Dengine, a set of programming modules, facilitated the automation of the orthomosaic creation through a Python script. The parameters used in creating the orthomosaics can be tailored to the specific requirements of the end application. In cases where the images are used for plant counting and plant spacing purposes, it creates a demand for high spatial resolution imagery. Conversely, when the aim is detecting the crop type, lower resolution suffices. The careful selection of Pix4D parameters is

imperative as they exert a significant influence on data quality and processing time. To validate the accuracy and consistency of our data, we cross-validated the GCPs' positions within each orthomosaic against their surveyed locations.

Our data pipeline also involved the creation of plot boundaries in the form of polygon shapefiles (.shp) corresponding to the geographic coordinates of each plot. These shapefiles were used as a mask to create images for each plot from each flight. However, we experienced several drawbacks to this approach. First was that the orthomosaic creation was a lengthy process, generally taking 4 to 6 hours to complete. Secondly was that the orthomosaic stitching process left aliasing artifacts in the resulting orthomosaic, resulting in degraded image quality relative to the original image. Lastly, in our first attempt to create orthomosaics, we experienced roughly 10% of the orthomosaics were incomplete and did not cover the entire plot area. We were able to get these orthomosaics to cover the entire area by adjusting the settings on Pix4D which also increased processing time from 4-6 hours to 1-2 days. **Figure 6** illustrates the degradation in image quality that can occur from the creation of an orthomosaic.

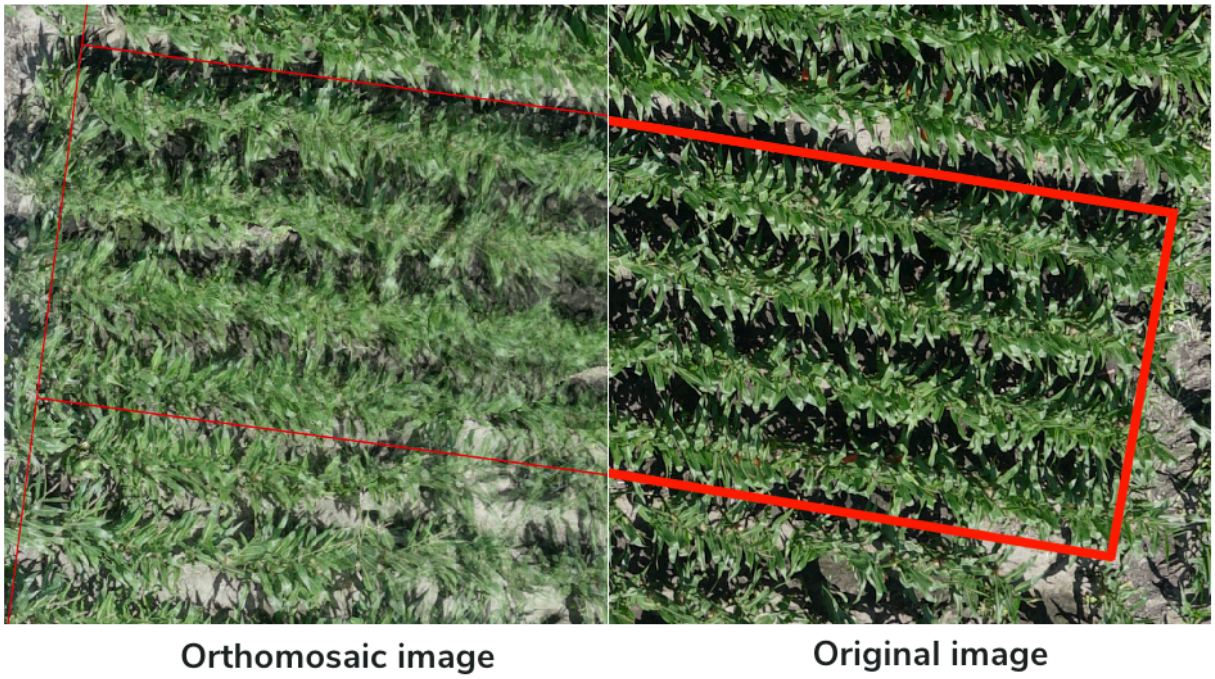

Figure 6. Orthomosaic image quality degradation

Since the settings required to generate complete orthomosaics involved processing times of 1-2 days, and yet the resulting image quality was often significantly degraded from the underlying raw image, an alternative approach to improve processing time without compromising the image quality from the underlying raw image was developed.

This approach involved using calibrated camera extrinsic parameters that were an output from the first stage of the Pix4D image processing pipeline and using the parameters to generate a georeferenced GeoTIFF image directly from the underlying raw image. **Figure 7** illustrates the changes to the sUAS image pipeline that were implemented to both reduce processing time and maintain the original image quality.

The new pipeline designed for the automated creation of tiles for each plot represents an advancement over the conventional approach of creating plot tiles from Pix4D orthomosaics. This method not only expedites the tile creation process but also yields higher quality plot tile images. The reduced processing time makes the data readiness pipeline more efficient while the improved image quality is expected to yield improvements to the accuracy of machine learning models trained from these images. Furthermore, the reduced processing time is expected to be particularly valuable for generating inferences very quickly after image capture.

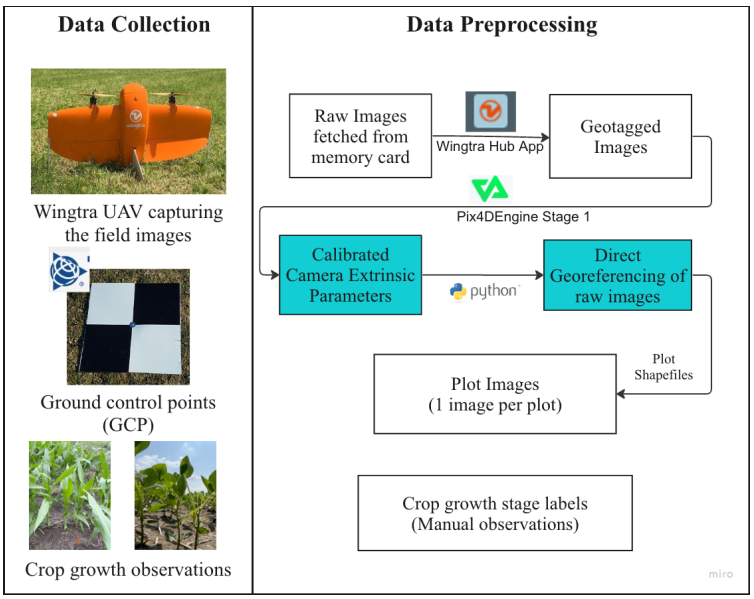

Figure 7. Version 2 of sUAS image pipeline

This process is facilitated through the utilization of Python libraries such as Rasterio, which is dedicated to the handling of geospatial data, and OpenCV for general image processing tasks. The initial phase encompasses the conversion of geo-tagged image files into GeoTIFF format. This conversion utilizes the camera extrinsic parameters generated from Stage 1 of the Pix4D including latitude, longitude, elevation, phi, omega, and kappa for each image, in addition to pertinent camera metadata like focal length. The process entails aligning the GIS locations accurately using the Ground Sample Distance (GSD) alongside rotation and translation matrices (Golparvar & Wang, 2021; Marco et al., 2018). The next phase entails the extraction of tiles from the GeoTIFF formatted images using the polygon shapefiles that represent the individual rectangular plots as a mask. The accuracy of the tile extraction process is influenced by two principal factors: the omega (roll) value and the distance between the image's center and the center of each plot. Images characterized by lower omega values and closer proximities to the plot centers can extract more precise tiles.

The accuracy of GeoTIFF files generated directly as compared to orthomosaics generated from Pix4D are compared by evaluating the plot tiles extracted from both approaches. The accuracy is assessed through the application of the Scale-Invariant Feature Transform (SIFT) algorithm (Nevins, 2017). This algorithm's utility lies in its ability to identify identical pixel points across disparate images. This new pipeline demonstrated variable error distances across three different locations—Western, Northwest, and Wooster. Data from five dates in Western, nine in Northwest, and nine in Wooster were utilized. The overall dataset exhibited a mean error distance of 0.68 meters, with the 25th percentile at 0.30 meters and the 75th percentile at 1.03 meters. Specifically, the mean error distances for each site were 0.33 meters for Western, 0.62 meters for Northwest, and 0.95 meters for Wooster. The larger errors in Northwest compared to the Western can be attributed to significantly biased omega (roll) values in the imagery. Meanwhile, in Wooster, images were captured horizontally relative to the plots, resulting in kappa (yaw) values around 0 degrees or ±180 degrees. This orientation minimally affects the rotation matrix, reducing its efficacy in image calibration adjustments. It demonstrated an average error distance of 0.57 meters, with recorded minimum and maximum error distances of 0.15 meters and 1.2 meters, respectively. Furthermore, the process boasts a processing time ranging from 4 to 25 minutes. This experiment was run on a machine with a processor of Apple M3 Pro APL1203 SoC and 18GB RAM.

There are four metrics that are important for sUAS image processing for machine learning applications. As we evaluate these metrics in **Table** 1, the version 2 pipeline shows improvements on 3 of the 4 metrics.

Table I. Comparison of metrics of interest for Version 1 (Orthomosaic) and Version 2 (Direct Georeferencing) pipelines

Metric	Version 1 pipeline	Version 2 pipeline
Completion percentage of flights that generate all plot tiles from raw images	86% (63/73)	100% (expected)
Processing time from raw images to plot tiles	4 hours on average	25 minutes on average
Geospatial accuracy	.05 meters on average	0.68 meters on average
Image quality	Aliasing artifacts	Best

Further work is planned to improve the version 2 pipeline to further reduce the processing time as well as improve geospatial accuracy. Initial work indicated that switching from *SciPy* library to *CuPy* library could reduce processing time significantly. Additional improvements will also be evaluated by exploiting parallel computing.

5.2 Soil and Climate Structured Data Pipeline

The Soil and Climate Structured Data Pipeline aggregates structured data from three separate sources into a database as shown in **Figure 8**:

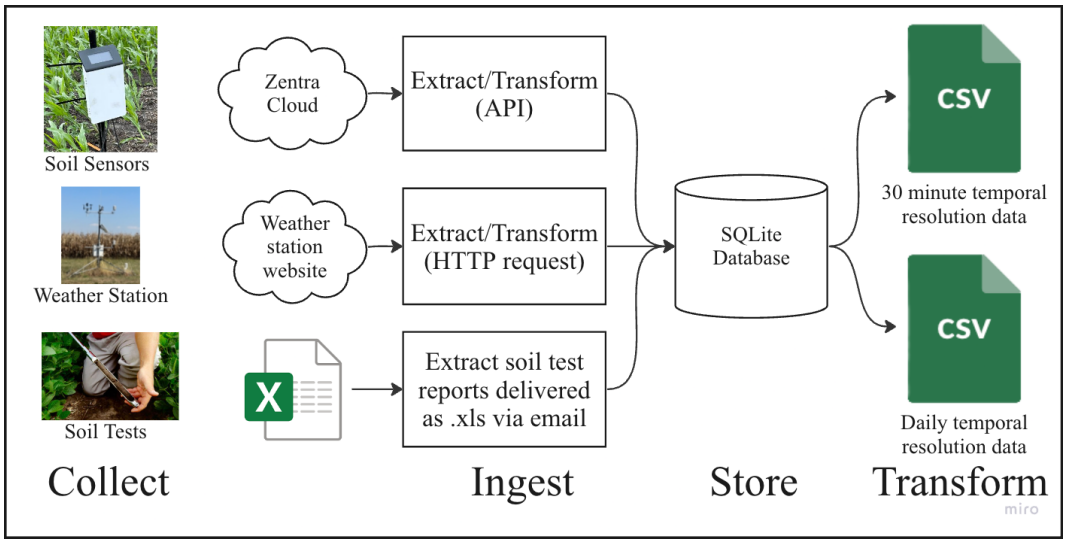

Figure 8: Soil and Climate Structured Data Pipeline

5.2.1 In-Situ Soil and Climate Sensors

The pipeline harnessed data from in-situ soil and climate sensors, which were distributed across the agricultural research sites. These sensors were connected to Meter ZL6 loggers, which recorded data at 30-minute intervals. Subsequently, the collected data was transmitted to the Meter Zentra Cloud through cellular connections. A Python script was employed to interface with the Zentra Cloud application programming interface (API) to retrieve the data and aggregate it into a local database.

5.2.2 Ohio State Weather Stations

Additionally, the pipeline incorporated data from OSU weather stations, which were located at each research site near the field plots. The data generated by these weather stations was accessible via a web interface, allowing for convenient querying and retrieval.

5.2.3 Soil Lab Testing Results

Soil lab testing results were received regularly as spreadsheets sent over email. The data in these

spreadsheets was also incorporated into the database.

5.3 Time-Series Data Alignment

The current data cleaning process is heavily reliant on the use of Jupyter notebooks to manually handle CSV and Excel files, involving unique scripts for each type of data transformation required such as mapping growth stage descriptions to numeric values, converting irregular time-series data into a standardized daily format, and averaging hourly sensor readings to daily values.

In the future, we propose a data pipeline designed to replace these manual Jupyter notebook operations with integrated, automated tasks enabling real-time data processing capabilities. This architecture would incorporate an Extract, Transform, and Load (ETL) scheme scheduled to operate continuously. As new data arrives, it is extracted and then transformed by applying various cleaning such as interpolating missing values, correcting errors, and aligning disparate data formats into a unified format. Following transformation, the data would be loaded into a database system which will act as the central repository from which the Plotly Dash application retrieves data. The Plotly Dash app can then dynamically query the database, pulling fresh data as updates occur (see Section 7.2 more details).

6. Model Training

After acquiring and preprocessing the various data using data pipelines, they are used to train AI models. Machine learning (ML) models such as Support Vector Machines (SVM), decision trees, regression networks, Convolutional Neural Networks (CNN), etc., are popularly selected for various agricultural use cases (Khanal et al. 2020). In one of our studies, we chose the Vision Transformer (ViT) model to identify corn and soybean crop growth stages using UAS RGB images. We selected the ViT model since it can capture spatial relationships, such as the development of leaves and the presence of flowers, in the images, which can be crucial to identify crop growth stages. In order to compare the classification and regression approaches for estimating crop development, the ViT architecture was modified to perform these tasks (**Figure 9**). During the training of the classification model, each crop growth stage is treated as an independent, discrete observation, whereas in a regression model, crop growth is considered a continuous observation.

After the model selection, the input data can either be segmented or resampled to match the model specifications and requirements. For the ViT model, the UAS images from each of the plots were divided into blocks of 224 x 224 x 3 to meet the ViT input requirement and then passed to the position embedding layer of ViT architecture. The embedded data is then passed to the transformer encoder and then to either the Multi-Layer Perceptron (MLP) head (classification) or Linear module (regression).

Each of the 224 x 224 x 3 blocks were annotated with ground-observed crop growth labels from their respective plot. For a more straightforward annotation, we converted crop growth labels to numerical values with VC as 1, VE as 2, V1 as 3, V2 as 4, ..., V16 as 19, R1 as 20, R2 as 21, ..., R8 as 27. These labels represent specific stages in the crop growth cycle, with VC indicating the start of the Vegetative stage and R1 indicating the start of the Reproductive stage. These labels can be considered as independent, discrete labels (classification) or as a sequence of continuous labels (regression).

In the annotated dataset, 60% was allocated for training, 20% for validation, and 20% for rigorous testing of the model performance. Here, the validation data accuracy was used for hyperparameter tuning and early stopping of the model training. The selected model parameters were carefully chosen between 100 to 150 epochs with a batch size of 16 to 64 and a learning rate of 0.001. The model demonstrated no further improvement in the accuracy after the 100 epochs, and hence, the training was stopped at that epoch. These models were trained on a robust 94GB RAM, Intel Xeon Silver @2.2GHz processor machine, with each model using a training time from 20 to 30 mins. We utilized the TensorFlow and Keras libraries in the Python

platform to execute the model training and testing.

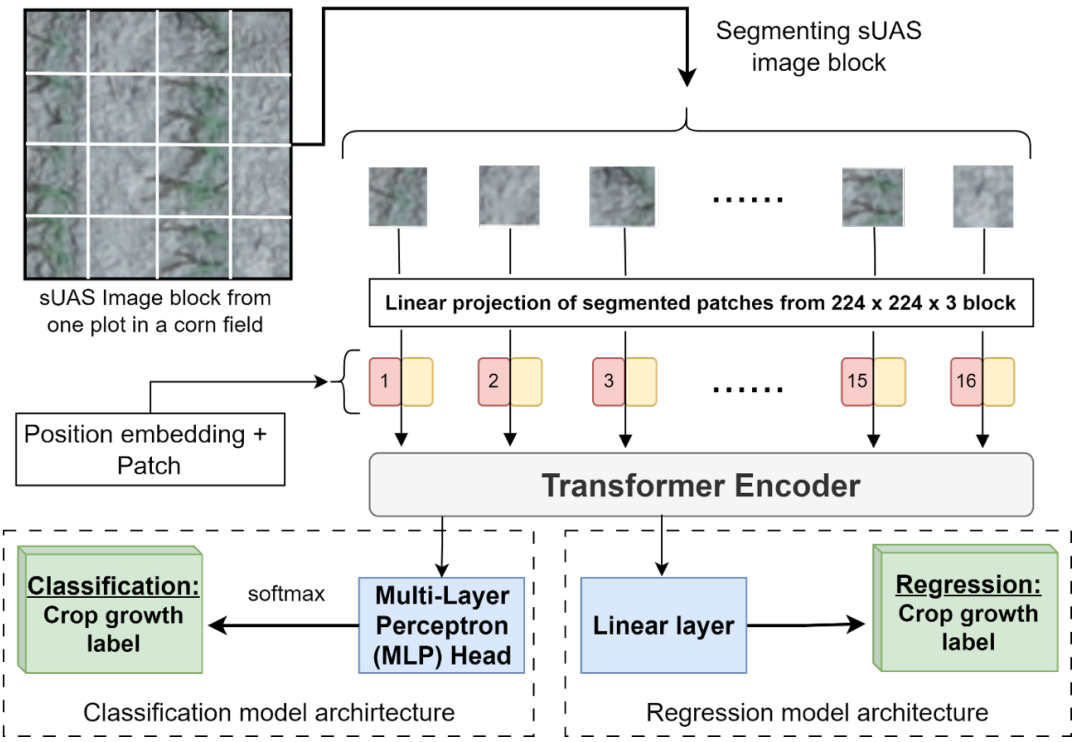

Figure 9: Illustrating steps in training the Vision Transformer (ViT) model using sUAS images to identify crop growth stages as classification and regression tasks.

We achieved an overall accuracy of 67% for the classification model and 87% for the regression model (**Figure 10**). These results demonstrate the significant impact of the model selection process on achieving better results. As continuous data better represent crop growth stages, we observed that the regression model produced superior results.

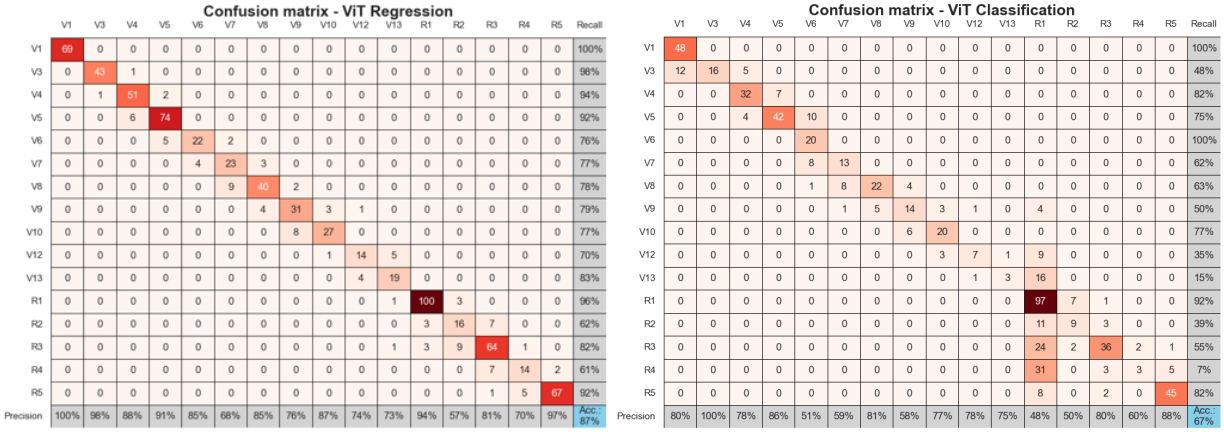

Figure 10: Confusion matrices representing ViT classification and regression model results with precision, recall, and overall accuracy values for each of the crop growth stages for a corn field in Northwest Ohio.

7. Data Visualization

In the past decade, we have seen AI and machine learning make an impact on a myriad of data rich application areas. As this trend continues, there is a growing need for tools that help practitioners gain a better understanding and trust of the data and insights presented by these technologies (Beauxis-Aussalet et al. 2021). Cultivating trust is critical for success in data-driven

and sustainable agriculture (Raturi et al. 2022). Farmers, especially, need validation that data-driven AI and ML tools will be able to achieve their envisioned goals of environmental and economic sustainability (Gardezi et al. 2024). One such tool to improve trust and provide validation is the creation of interactive data visualizations (Beauxis-Aussalet et al. 2021).

7.1 Dashboard

An interactive data visualization dashboard was created with the cleaned, wrangled data from this study. Visualization methods at this stage in the ML pipeline are generally used to explore interesting subgroups and pinpoint particular outliers. The purpose of this dashboard is to visualize the collected data at the plot level, providing specific insights for each plot and comparing them to other plots in the field. This provides the user with a reference to see if a particular plot has characteristics that are significantly different from the norm (i.e., out of distribution), enabling better understanding of the data. **Figure 11** shows a screenshot of the dashboard.

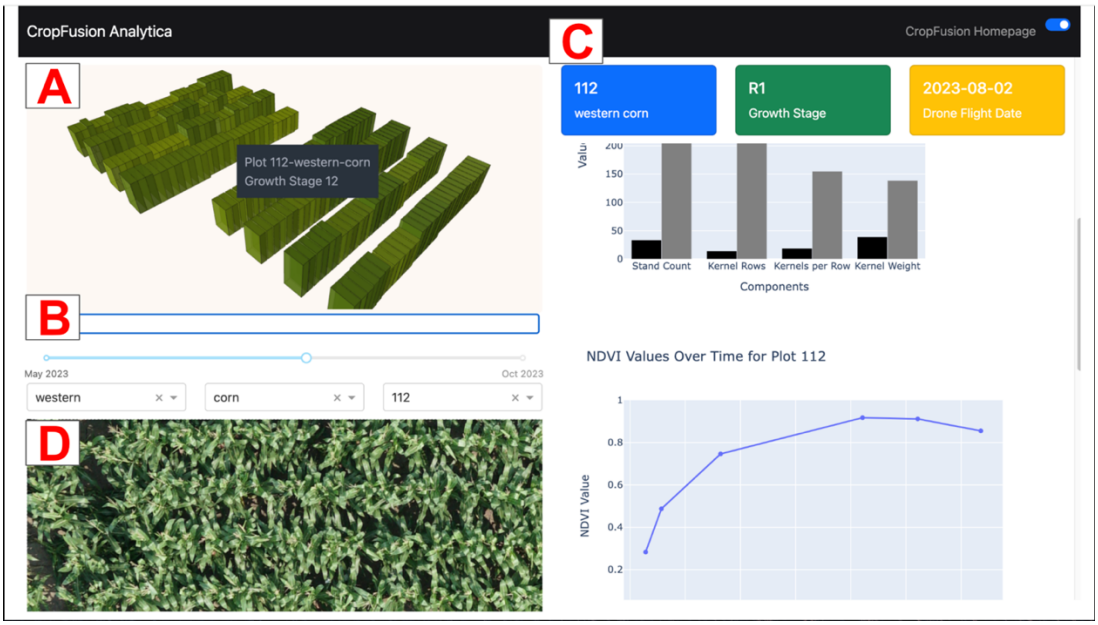

Figure 11. Dashboard Screenshot

The dashboard was created with *Plotly Dash*, an open-source Python framework that enables the creation of interactive, data-driven dashboards. The top left of the dashboard contains a pane with a map focused on a specific field [A]. A 3D geospatial layer is rendered on top of the field using the plot boundaries from the created orthomosaic files (.geojson files). This rendering is created with DeckGL, a WebGL-powered visualization framework. Each plot is outlined and has its own 3D layer. The height and color of the cross section represents the current growth stage of the plot. To view the exact growth stage value of a specific plot, the user may hover over the specific plot.

Below the map, there is a time slider and a series of dropdown menus to select field, crop type, and plot number, respectively [B]. This collection of inputs can be used to update the dashboard figures [C] on the right side of the screen that provide information based on the different modalities of tabular data collected in the study. **Table II** provides a summary of the figures in the dashboard. The position of the time slider can be altered to update the geospatial visualization and figures across different time periods across the growing season. A drone image of the selected plot at the specific time selected by the time slider is populated at the bottom left of the screen [D].

Perhaps the most significant feature in the dashboard is the interactivity provided in the geospatial visualization. The user may click on a specific plot on the 3D geospatial layer to populate the figures with information about the selected plot, enabling the user to derive plot-specific insights in an interactive, intuitive manner. This interactivity, combined with the time-slider

enables the user to explore multimodal data across both spatial and temporal dimensions.

Table II: Summary of Dashboard Elements

Chart	Description
NDVI Values over Time	NDVI values for a specific plot are reported over the growing season.
Yield Estimation	The yield component information for a specific plot is displayed.
Actual vs Predicted Growth Stage	Tracks crop growth stage over the growing season alongside a predicted line using the machine learning models.
Precipitation and Soil Moisture	Soil moisture, precipitation, and temperature at the crop location over the growing season.

Overall, these figures, driven by plot-clicks and the time-slider, provide a valuable view for users that enable further understanding of the collected data.

7.2 Early User Feedback

A user feedback session was conducted shortly after an initial prototype of the dashboard was developed. Participants were faculty and students from OSU's HCS who interacted with the dashboard to evaluate its features and data presentation. Users highlighted the necessity for clearer visual representations of crop growth stages on the geomap, and requested more robust capabilities for comparing crop yields not only by location but also by planting date and crop hybrid in the case of corn. Additionally, there was a demand for integrating growth stage data with weather data on the dashboard, with options for toggling different data layers. The feedback also highlighted the importance of certified crop advisors as a potentially important user group for the dashboard.

8. Conclusion

In summary, we articulate the importance of AI applications in agriculture and highlight a data-centric approach to building AI-centric CI. We provide an example of how multimodal data can be leveraged for yield estimation that combines UAS imagery with climate sensing. Given that one drawback of AI is that it can be a black box to users, we highlight the importance of visual interfaces that can build understanding and trust in the system.

While this paper seeks to present a broad view of AI applications in agriculture, each component represented is at a relatively early stage of research. Furthermore, there is integration work needed for each of the components to function in a unified system.

Direct georeferencing techniques show promise to retain original UAS image quality while drastically reducing compute requirements. Further work will be aimed at parallel computing to reduce processing time and improvements to geospatial accuracy. Additional work is also needed for error detection to ensure plot tile images are correctly generated.

The user interface shows promise, but more user research is needed to tailor the interface to the needs of unique stakeholders. The prototype was built using prior year's data. A valuable next step will be to enable it to show data from the current growing season within a few days after data is collected.

Additional areas of improvement include making use of standardized AI dataset formats that can promote aggregation of larger datasets across institutions. Additionally, the various data types

currently exist in an ad hoc folder structure. Organizing the data into a database schema is likely an important next step to improve the system.

The work outlined in this paper is intended to spark discussion and collaboration among various stakeholders (e.g., researchers, crop consultants, farmers) such that the promise of AI in agriculture can be more fully realized.

Acknowledgments

The authors would like to thank the farm managers, Lynn Ault, Joe Davlin, and Matt Davis, and their teams at each research site for executing these small-plot research trials and their support in collecting the data that went into this study. This study was funded by the Nationwide AgTech Innovation Hub and ICICLE (icicle.osu.edu). The Nationwide AgTech Innovation Hub is a collaboration between Nationwide Mutual Insurance Company, Ohio Farm Bureau and The Ohio State University College of Food, Agricultural, and Environmental Sciences. ICICLE is an NSF funded AI institute focused on establishing a national cyberinfrastructure for AI with Digital Agriculture as one of its three use-inspired science cases.

References

Beauxis-Aussalet, E., Behrisch, M., Borgo, R., Chau, D. H., Collins, C., Ebert, D., et al. (2021). The Role of Interactive Visualization in Fostering Trust in AI. IEEE Computer Graphics and Applications, 41(6), 7–12. https://doi.org/10.1109/MCG.2021.3107875

Burwood-Taylor, L. (2023, January 10). BREAKING: Alphabet brings agtech startup out of stealth with data from 10% of world's farmland, 3 major customers. AgFunderNews. https://agfundernews.com/breaking-alphabet-brings-agtech-startup-out-of-stealth-with-data-from-10-of-worlds-farmland-3-major-customers. Accessed 17 October 2023

Duflock, W. (2023, April 25). A Free Image Library, Now Expanded, for AgTech Startups. Western Growers Association. https://www.wga.com/news/a-free-image-library-now-expanded-for-agtech-startups/. Accessed 2 May 2024

Chandra, R., Swaminathan, M., Chakraborty, T., Ding, J., Kapetanovic, Z., Kumar, P., & Vasisht, D. (2022). Democratizing Data-Driven Agriculture Using Affordable Hardware. IEEE Micro, 42(1), 69–77. https://doi.org/10.1109/MM.2021.3134743

Gadiraju, K. K., Ramachandra, B., Chen, Z., & Vatsavai, R. R. (2020). Multimodal Deep Learning Based Crop Classification Using Multispectral and Multitemporal Satellite Imagery. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3234–3242). Presented at the KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event CA USA: ACM. https://doi.org/10.1145/3394486.3403375

Gardezi, M., Joshi, B., Rizzo, D. M., Ryan, M., Prutzer, E., Brugler, S., & Dadkhah, A. (2024). Artificial intelligence in farming: Challenges and opportunities for building trust. Agronomy Journal, 116(3), 1217–1228. https://doi.org/10.1002/agj2.21353

Golparvar, B., & Wang, R.-Q. (2021, November 27). AI-supported Framework of Semi-Automatic Monoplotting for Monocular Oblique Visual Data Analysis. arXiv. http://arxiv.org/abs/2111.14021. Accessed 3 May 2024

Lindsey, L. E., Tilmon, K., Michel, A., & Dorrance, A. (2017). Soybean production. In L. E. Lindsey & P. R. Thomison (Eds.), Ohio Agronomy Guide (15th ed., Bulletin no. 472, pp. 56–68). The Ohio State University Extension.

Kapetanovic, Z., Chandra R., Chakraborty T., & Nelson A. (2019). FarmBeats: Improving Farm Productivity Using Data-Driven Agriculture. SIAM News. https://sinews.siam.org/Details-Page/farmbeats-improving-farm-productivity-using-data-driven-agriculture. Accessed 17 October 2023

Khanal, S., Kc, K., Fulton, J. P., Shearer, S., & Ozkan, E. (2020). Remote Sensing in Agriculture—Accomplishments, Limitations, and Opportunities. Remote Sensing, 12(22), 3783. https://doi.org/10.3390/rs12223783

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386

Marco, C., Claudio, B., Ueli, R., Thalia, B., & Patrik, K. (2018, December 12). Using the Monoplotting Technique for Documenting and Analyzing Natural Hazard Events. In J. Simão Antunes Do Carmo (Ed.), Natural Hazards—Risk Assessment and Vulnerability Reduction. IntechOpen. https://doi.org/10.5772/intechopen.77321

Nevins, R. P. (2017). Georeferencing Unmanned Aerial Systems Imagery via Registration with Geobrowser Reference Imagery [Master's thesis, Ohio State University]. OhioLINK Electronic Theses and Dissertations Center. http://rave.ohiolink.edu/etdc/view?acc_num=osu1500378454106286

Raturi, A., Thompson, J. J., Ackroyd, V., Chase, C. A., Davis, B. W., Myers, R., et al. (2022). Cultivating trust in technology-mediated sustainable agricultural research. Agronomy Journal, 114(5), 2669–2680. https://doi.org/10.1002/agj2.20974

Taheri, J., Dustdar, S., Zomaya, A., & Deng, S. (2023). Edge Intelligence: From Theory to Practice. Cham: Springer

International Publishing. https://doi.org/10.1007/978-3-031-22155-2

Thomison, P., Michel, A., Tilmon, K., Culman, S., Paul, P. (2017). Corn Production. In L. E. Lindsey & P. R. Thomison (Eds.), Ohio Agronomy Guide (15th ed., Bulletin no. 472, pp. 32–55). The Ohio State University Extension.

Ur Rehman, M. H., Yaqoob, I., Salah, K., Imran, M., Jayaraman, P. P., & Perera, C. (2019). The role of big data analytics in industrial Internet of Things. Future Generation Computer Systems, 99, 247–259. https://doi.org/10.1016/j.future.2019.04.020

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention Is All You Need. arXiv. http://arxiv.org/abs/1706.03762. Accessed 17 October 2023

Wang, R. Y., Kon, H. B., & Madnick, S. E. (1993). Data quality requirements analysis and modeling. In Proceedings of IEEE 9th International Conference on Data Engineering (pp. 670–677). Presented at the IEEE 9th International Conference on Data Engineering, Vienna, Austria: IEEE Comput. Soc. Press. https://doi.org/10.1109/ICDE.1993.344012