CYIENT

GEOAI IN CROP ANALYTICS AND SUSTAINABLE AGRICULTURE

CONTENTS

Abstract	01
Introduction	01
Problem Definition	02
High-Level Solution	02
Solution Details	04
Business Benefits/Best Practices	06
Case Study	07
Recommendations and Way Forward	09
References	09
About the Author	09

Abstract

Geospatial Artificial Intelligence (GeoAI) has been at the forefront of spatial analytics with the advent of huge open source Earth Observation (EO) data and easy accessibility over cloud platforms, offering new ways of solving problems across industries such as energy, utilities, mining, supply chain, and in ESG compliance.

EO has proven its efficacy in mapping and monitoring the health/condition of crops or vegetation. A time-series EO assessment of crop phenology helps study the crop health, underlying soil, and waterbodies, can enable recommendations for suitable seeds, fertilizers, and pesticides, and measure the impact of environment and land management. This paper provides a comprehensive overview of machine learning operations (MLOps) in GeoAI in sustainable agriculture. It discusses enabling a cloud-agnostic platform with an automated pipeline built on comprehensive EO analytics using MLOps. It further describes a framework of open-source tools that help build an end-to-end GeoAI pipeline and presents select case studies done in agriculture.

Introduction

Environment and land management in most countries entail encouraging good farming practices that ensure sustainable, intensive agriculture, and periodically monitoring and assessing these farming practices. Considering field inspections are time-consuming and cost-intensive, it often calls for a collaborative approach of monitoring farmlands to understand the area of parcel cropped and assess the implementation of farming practices and farm management remotely. With remote analytics, agencies locate farm areas where vegetation and soil health is under stress, either due to poor water supply or non-compliance with best practices. They can help organize field inspections and crew for select high-risk regions and plan mitigation to restore land/environment health.

The type of crops or vegetation grown on the farm directly impacts the health of the soil and its productivity, which determines its sustainability in future agriculture. Therefore, to achieve the UN's climate change target of Net Zero by 2050, soil restoration and protection have become core issues in the 2020 Agriculture Bill. The new farming rules for water require following best practices to reduce soil and water erosion.

Problem Definition

With the humongous spatial data from earth observation and advances in analytical capabilities, there is a need to optimize and automate the DevOps pipeline for GeoAI solutions. MLOps addresses this by provisioning an automated machine learning platform for effective end-to-end processing. The Crop Variety Classification—based on phenological characteristics throughout the year for the inter-seasonal crop species as identified by their spectral profiles—has been considered a basis for classifying the crops and analyzing crop health.

High-Level Solution

Machine learning operations (MLOps) is an engineering practice that provides a seamless pipeline covering data engineering and machine learning along with DevOps. It covers aspects such as best practices and a DevOps culture in provisioning end-to-end conceptualization, implementation, monitoring, deployment, and scalability of machine learning solutions. MLOps aims to facilitate the creation of machine learning products by leveraging CI/CD automation; workflow orchestration; reproducibility; versioning of data, model, and code; collaboration; continuous ML training and evaluation; ML metadata tracking and logging; continuous monitoring; and feedback looping.

Key Objectives of MLOps

- Easy, repeatable, portable deployments on diverse infrastructure
- Deploying and managing loosely-coupled microservices
- Scaling on demand

Key Components of MLOps

Source Code Repository Commit and merge–data, model, and code	**Model Registry** Centrally stores the trained ML models together with their metadata	**Model Serving Component**
• Source code collaboration/merging • Versioning	• Model reproducibility • Model versioning	• CI/CD automation
CI/CD Component Build, test, deliver, and deploy steps along with feedback looping	**Model Training Infrastructure** Provisioning a scalable and distributed infrastructure	**Monitoring Component** Model serving performance; ML infrastructure, CI/CD, and orchestration
• CI/CD automation • Continuous ML training and evaluation • Feedback loops	• Continuous ML training and evaluation—monitoring component, a feedback loop, and an automated ML workflow pipeline	• Continuous monitoring • Feedback looping
Workflow Orchestration Task orchestration of ML workflow via Directed Acyclic Graphs representing execution order and artifact usage of single steps of workflow	**Feature Stores** Central storage of commonly used features typically configured with one offline datastore for experimentation and other online stores for predictions in production	**ML Metadata Stores** Tracking of various kinds of metadata – including model specific hyperparameters and resulting performance metrics, model lineage: data and code used; for each training job—training date and time, duration
• Workflow orchestration • Reproducibility • Continuous ML training and evaluation	• Reproducibility • Versioning	• Versioning • ML Metadata tracking/ logging

Figure 1: Solution Components

Solution Details

Based on the above components, a generalized MLOps end-to-end architecture typically features:

Problem definition, including exploratory data analysis: This step analyzes the business problem to be solved using ML, designs the overall ML solution, tools, and technology, determines what data and possible sources of data for EDA, trains models, and evaluates the distribution and quality of the data, whether annotated/labeled.

Research/experimentation and model building: This step covers training models, feedback looping, and retraining tasks.

Automated ML workflow pipeline up to the model serving: This covers the iterative process of data extraction, data preparation and validations, model refinement, validations, and finally, pushing to the model registry.

Pre-processing, cleansing, transformation, feature engineering, and data ingestion: This step covers data transformation rules (normalization, aggregations) and cleaning rules to bring the data into a usable format and defining feature engineering rules such as the calculation of new and more advanced features based on other features and using feedback looping.

Solution deployment: Deployment of the final model and pipeline on Kubernetes.

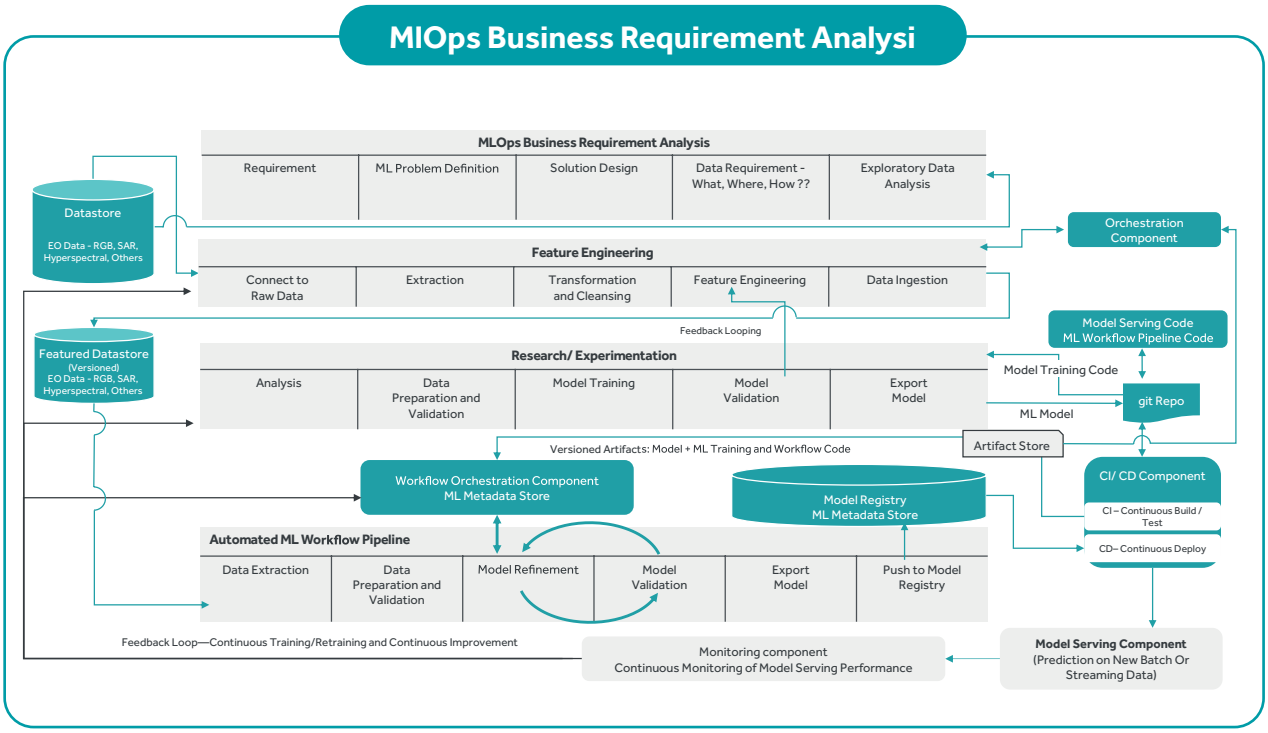


Figure 2: Solution Architecture

MLOps Framework

A generic open-source MLOps framework with Kubeflow as the core engine or platform on top of Kubernetes is depicted below. It indicates select open-source tools to execute a series of tasks for end-to-end MLOps.

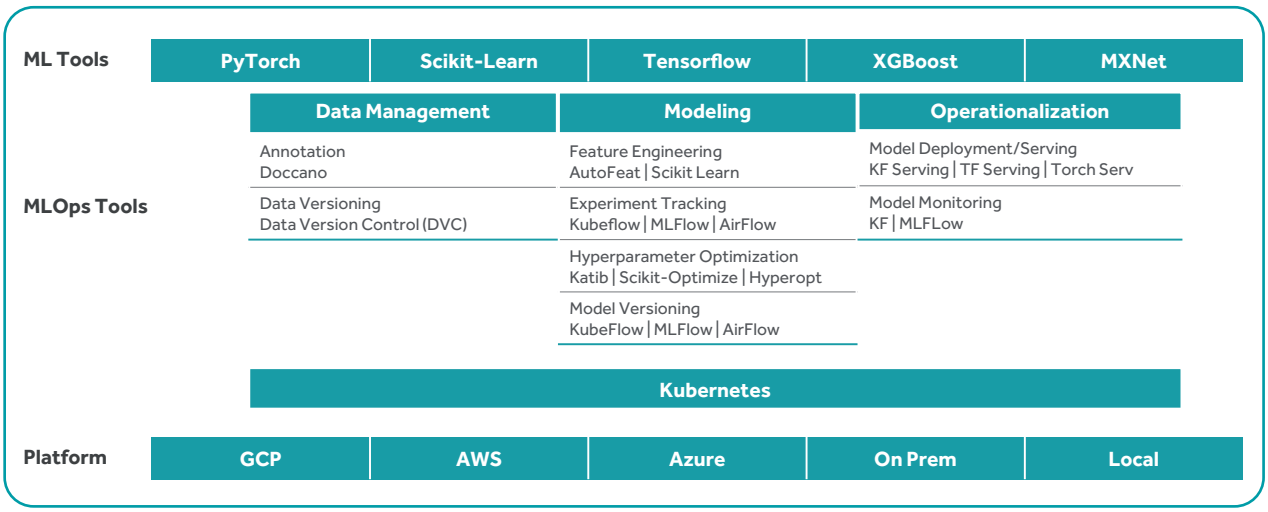


Figure 3: MLOps Framework

Kubeflow builds on Kubernetes as a system for deploying, scaling, and managing an MLOps system. The Kubeflow configuration interfaces help users specify the ML tools required for the workflow. The workflow can be deployed over various clouds, local, and on-premises platforms for experimentation and for production use.

Business Benefits/Best Practices

MLOps provides DevOps for an ML platform to effectively collaborate with data and business analysts, engineers, data scientists, and researchers in provisioning accelerated model development and its deployment with the help of monitoring, validation, and management systems for machine learning models. It enables streamlining model training and model deployment pipelines, and using continuous integration/continuous delivery to simplify retraining and integrate machine learning easily into existing release processes. It also enables advanced data bias analysis to improve model performance over time.

Key MLOps Benefits

Evaluates the importance of features and creates more advanced models with minimal bias using uniform distribution metrics.

Reduces variation in model iterations and provides resiliency for enterprise-level scenarios with reproducible learning and models.

Ease of deployment of models on production environment.

Uses dataset registries and advanced model registries to track resources.

Creates audit trails to meet regulatory requirements and automatically trace experiments.

Provides improved traceability by tracking code, data, and metrics in the execution log; keeps track of version history and model origin to enable auditing.

Packs models quickly, ensuring high quality at every step through profiling and model validation.

Uses built-in integration with Azure DevOps or AWS CodeBuild and GitHub actions to plan, automate, and manage workflows efficiently; uses automatic scaling and managed clusters of CPUs and GPUs with distributed learning in the cloud.

Case Study

Cyient used MLOps using Sentinel 2 Optical Imageries for a marquee government agency to identify crop variety and health. The work entailed satellite-based monitoring with Computer Vision to detect and locate good and healthy farming practices. Keys tasks included:

A. Problem definition, including exploratory data analysis

The Crop Variety Classification and Health Analytics were based on biophysical indicators—NDVI, LAI, and leaf chlorophyll of the plantation—estimated using high-resolution open-source satellite imageries that distinguished each crop species based on the spectral signature of crop phenology. Some of the key vegetation indices thus estimated and reviewed were:

- The Normalized Difference Red Edge (NDRE) index for chlorophyll is to assess whether a growing plant is healthy or not. A low chlorophyll index indicates problem crops —sick, infested with pests, or nutrient-deficient plants. It uses a combination of a Near-InfraRed (NIR) band and the Red Edge range between visible Red and NIR and calculated as NDRE = (NIR – RedEdge)/(NIR + RedEdge).
- The Normalized Difference Vegetation Index (NDVI) calculates photosynthetically active biomass indicating vegetation health, estimated as = (NIR-Red)/(NIR+Red).
- The Normalized Difference Moisture Index (NDMI) detects moisture levels in vegetation using a combination of near-infrared and short-wave infrared (SWIR) spectral bands. It is a reliable indicator of water stress in crops and helps monitor irrigation, especially in areas where crops require more water than nature can supply. NDMI is calculated using the near-infrared and the short-wave infrared reflectance: NDMI = (NIR – SWIR)/(NIR + SWIR).

Although the ML-driven analysis generated a noise-free, highly accurate, and consistent output, its dependency on ground truth validation was a major bottleneck. To train and validate the model, ground truths were crucial to classify and evaluate every agricultural zone and identify crop variety. The model helped identify wheat and potato with precision and accuracy.

Seasonal data of the growing period of the crop were considered to assess the vegetation sensitivity of the NIR and Red spectrums; the Red Edge difference index was applied to accentuate the species-wise phenology variance. The Red Edge is a region in the Red-NIR transition zone of the vegetation reflectance spectrum. It marks the boundary between absorption by chlorophyll in the red visible region and scattering due to the internal leaf structure in the NIR region. As a result, this part of the spectrum reflects any slight change in the bio-chemical structure and physical parameters during the plant phenology, either by biomass accumulation or impact of stress.

B. Research/experimentation, model building, and containerized deployment

The crop and stress identifier model was trained over various iterations to improve its performance. The following aspects were considered in refining the training labels:

- **Variability:** A large number of variations of each class of target crops were captured. Wheat parcels from different parts of the work area showing variations with respect to soil, climate, and physiography characteristics, were captured. Training data of the target crops were thus added from different regions at every iteration to enhance the model's robustness.
- **Size of crop parcels:** Different size filters were implemented to assess their impact on detection.
- **Features outside croplands:** Training areas were also created for features that were not croplands but were present in the Area of Interest AOI. This was to ensure that model did not wrongly classify them as croplands and could identify them as non-crop classes (e.g., parcels with grassland with similar spectral patterns as the target crop).

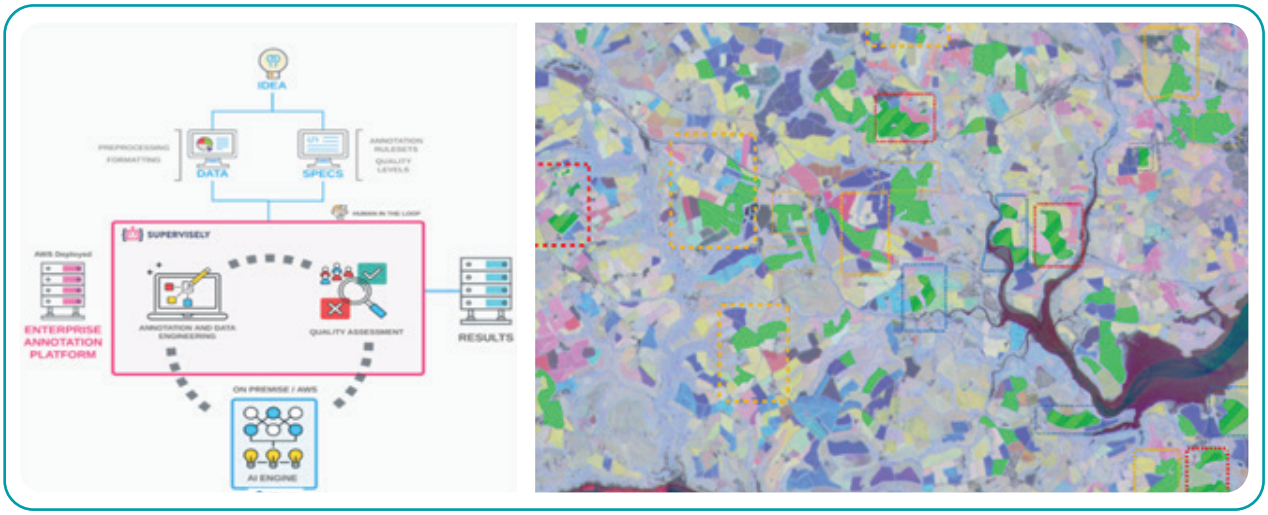


Figure 5: ML- vs. DL-based crop monitoring outcome

The crop classifier/detector was trained to identify crops automatically and predict without any ground truth input. The trained model was deployed and tested in an unknown region to assess its efficacy. The figure above (Fig 5), exhibits the results where the region highlighted in green, points to wheat cultivation. The predicted wheat parcels were validated with the farmers' claim data (highlighted in yellow) and were having a 90% match.

The model was retrained with additional region and variability, and the prediction improved.

Within the project's timeline, Cyient created an annotations library with 1000+ scenarios.

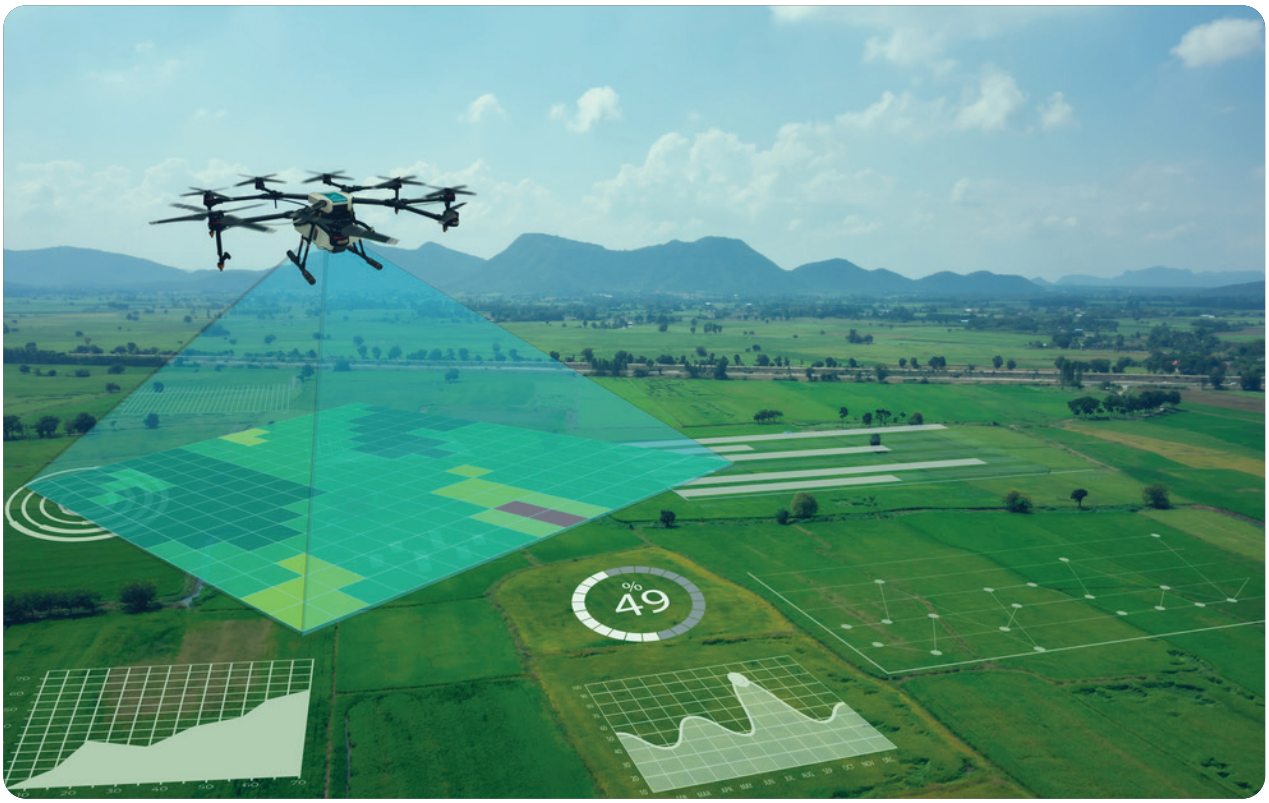
Once generated by the model (either ML or DL), the detected output goes through a visual inspection for QC to guarantee data quality. The existing ML model has accuracy confidence of >85%, and we apply visual quality checks to a certain extent to clean up unwanted polygons and false positives. As per the DL model (accuracy confidence of 86%), more training is required from different geographies to enhance the model's robustness and consistency.

Recommendations and Way Forward

The next steps would be to build a comprehensive system of precision agriculture—**crop yield, production quantity forecasting, identifying crop stress areas, supply chain for seeds, fertilizers, pesticides, and produce, estimating soil organic carbon,** and so on. The solution would cover building an expert system that uses soil, water (drainage pattern), weather conditions, and other factors to strategize for enhanced production. Our potential partnership with commercial data providers such as Planet and Wyvern and other EO data providers can help build extended capabilities.

References

- Kubeflow: An MLOps Perspective. ML Pipelines and ML Components | by Alex Punnen | Towards Data Science
- Kubeflow MLOps: Automatic pipeline deployment with CI / CD / CT | by Antoine Villatte | Towards Data Science
- https://environment.data.gov.uk/rpa/api
- Amazon SageMaker geospatial capabilities - Amazon SageMaker
- https://research.aimultiple.com/mlops-tools/



About Cyient

Cyient (Estd: 1991, NSE: CYIENT) is a leading global engineering and technology solutions company. We are a Design, Build, and Maintain partner for leading organizations worldwide. We leverage digital technologies, advanced analytics capabilities, and our domain knowledge and technical expertise, to solve complex business problems.

We partner with customers to operate as part of their extended team in ways that best suit their organization's culture and requirements. Our industry focus includes aerospace and defense, healthcare, telecommunications, rail transportation, semiconductor, geospatial, industrial, and energy. We are committed to designing tomorrow together with our stakeholders and being a culturally inclusive, socially responsible, and environmentally sustainable organization.

For more information, please visit www.cyient.com

Contact Us

North America Headquarters
Cyient, Inc.
99 East River Drive
5th Floor
East Hartford, CT 06108
USA

T: +1 860 528 5430
F: +1 860 528 5873

Europe, Middle East, and Africa Headquarters
Cyient Europe Limited
Apex, Forbury Road,
Reading
RG1 1AX
UK
T: +44 118 3043720

Asia Pacific Headquarters
Cyient Limited
Level 1, 350 Collins Street
Melbourne, Victoria, 3000
Australia

T: +61 3 8605 4815
F: +61 3 8601 1180

Global Headquarters
Cyient Limited
Plot No. 11
Software Units Layout
Infocity, Madhapur
Hyderabad - 500081
India

T: +91 40 6764 1000
F: +91 40 2311 0352

Follow us on: