Universal Dataset Constructor & Preprocessing Framework for Earth Observation AI Tasks in Digital Agriculture

Viktoria Sorokina[1], Ilya Klinkov[1], Dzmitry Yablonski[1], Stefan Henkler[2], Andrew Zakhary[2]
[1] GeoPard Agriculture, Neusser Str. 472, 50733 Cologne, Germany
[2] Hamm-Lippstadt University of Applied Sciences, Hochschule Hamm-Lippstadt Marker Allee 76-78 59063 Hamm, Germany

A paper from the Proceedings of the
17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture
13-15 July 2026
Porto Alegre, Brazil

Abstract.

High-performing AI models for yield prediction, land management, and environmental risk monitoring depend on the integration of heterogeneous Earth Observation (EO) data with field-level agronomic and meteorological sources. This paper presents a Universal Dataset Constructor and Preprocessing Framework that automates the creation of multimodal, leakage-safe, and task-ready datasets for precision agriculture applications. The framework unifies optical and SAR satellite imagery, commercial high-resolution and drone data, weather observations, soil parameters, digital elevation models (DEM), and as-applied management records into a sensor-agnostic pipeline with canonical data contracts and built-in quality assurance.

A core component of the framework is a task-aware dataset generation engine that dynamically structures features according to downstream application requirements. The system supports phenology-driven tasks such as crop detection and yield prediction, management intervention analysis including fertilization and crop protection monitoring, land-use and crop-type classification, and disturbance detection tasks such as drought, flood, fire, erosion, and pest outbreak monitoring. Built-in quality assurance modules address temporal leakage, satellite coverage gaps, weather alignment, label consistency, and class imbalance, enabling reliable model development and evaluation.

The framework generates both pixel-level and zone-level representations compatible with convolutional neural networks, transformer architectures, and physics-informed machine learning approaches. Validation on a multi-region harvest moisture prediction use case demonstrated that a physics-aware sequence model outperformed grouped statistical baselines, confirming the effectiveness of the generated datasets. By standardizing EO preprocessing and multimodal data integration, the proposed framework reduces dataset preparation effort while improving reproducibility, scalability, and cross-region model transferability for production-grade agricultural AI systems.

Keywords.
GeoAI, Earth Observation, Precision Agriculture, Remote Sensing, Dataset Engineering.

Introduction

Recent advances in Earth Observation (EO) systems such as Sentinel-1/2, Landsat 8/9, and commercial high-resolution constellations have significantly improved the availability of multi-temporal and multi-sensor data for agricultural monitoring (Drusch et al., 2012). Combined with weather and field-level data, these sources enable data-driven approaches for crop monitoring, yield prediction, and environmental risk assessment (Khaki and Wang, 2019).

However, state-of-the-art agricultural AI systems remain heavily constrained by data engineering complexity. Prior studies in precision agriculture and EO-based machine learning have highlighted that dataset construction – rather than model design – is often the dominant bottleneck in production pipelines, e.g., EO time-series modeling, multi-modal fusion frameworks, and crop monitoring systems (Barbedo, 2019). Key challenges include temporal misalignment, spatial heterogeneity, leakage risks in time-series construction, and inconsistent integration of operational farm data.

Existing approaches typically address these issues in task-specific pipelines, limiting reproducibility and scalability across regions and crops. This motivates the need for a unified infrastructure layer that standardizes EO ingestion, harmonization, and task-aware dataset generation under a single reusable framework.

Framework Architecture

The framework consists of three components: ingestion and harmonization, a canonical data model, and a task-aware dataset generation engine. The ingestion layer integrates optical and SAR satellite imagery, drone imagery, weather data, soil measurements, topographic information, operational records, and ground-truth observations through dedicated adapters. Atmospheric correction, spatial resampling, temporal alignment, and feature normalization are applied to produce consistent inputs.

All sources are transformed into five standardized data structures (*season_events_df, measurements_df, satellite_images_df, weather_df*, and *model_targets_df*) that serve as a universal contract between data engineering, agronomy and machine learning workflows. A task registry then dynamically defines feature construction logic, temporal windows, spatial granularity, and label contracts according to the target application.

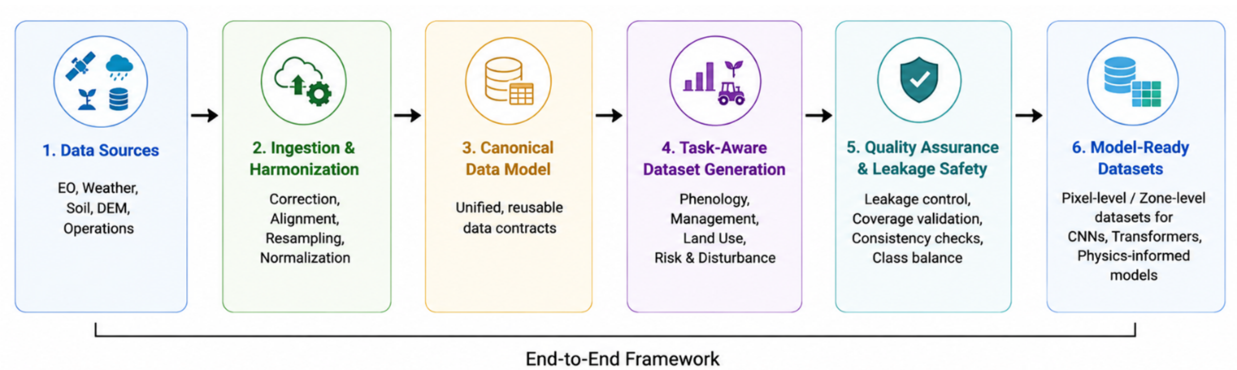

Fig 1. Architecture of framework

Task-Aware Dataset Generation

The framework supports four categories of agricultural AI tasks. Phenology-driven applications, including crop detection and yield prediction, utilize vegetation index trajectories, weather accumulations, and leakage-safe temporal windows. Management intervention tasks such as fertilization, tillage, and crop protection analysis combine pre- and post-intervention EO observations with machinery and application records. Land-use classification tasks leverage multi-season composites, SAR-optical fusion, and spatial texture features to improve transferability across regions. Disturbance and risk monitoring tasks incorporate anomaly detection, change detection, and weather-driven exposure metrics for droughts, floods, fires, pest outbreaks, and erosion events.

Quality Assurance and Leakage Safety

Quality assurance is integrated throughout the pipeline. Automated checks verify satellite and weather coverage, temporal alignment, label plausibility, and class balance for rare events. Leakage prevention mechanisms ensure that no feature information post-dates the target observation. Dataset evaluation uses grouped splitting strategies based on season and field identifiers, preventing contamination between training and validation sets and providing realistic estimates of model performance.

Validation Case Study

The framework was validated using a harvest moisture prediction use case combining Sentinel-2, commercial provider, weather observations, and multi-region field data from different countries. A physics-informed dry-down model was compared against standard machine learning baselines.

Table 1. Performance of the models built and trained with the help of universal constructor.

Model	MAE (%)	RMSE
Train Mean	4.13	4.71
Ridge Regression	3.33	3.98
Random Forest	2.79	3.45
Physics-Informed GeoPard model	2.73	3.24

The superior performance of the GeoPard model demonstrates that the generated datasets preserve meaningful temporal and agronomic information beyond static field-level correlations.

Conclusion

The proposed framework provides a universal, sensor-agnostic infrastructure for constructing multimodal EO datasets for precision agriculture. By standardizing data ingestion, harmonization, quality assurance, and task-specific feature generation, it substantially reduces dataset preparation effort while improving reproducibility and model transferability. Future work includes advanced SAR feature engineering, tighter drone-satellite integration, automated event discovery from management logs, and support for edge-based inference, further strengthening the framework as a foundation for production-grade agricultural AI systems.

References

1. Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment, 120, 25–36.
2. Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in Plant Science, 10, 621.
3. Barbedo, J. G. A. (2019). A review on the use of unmanned aerial vehicles and imaging sensors for monitoring and assessing plant stresses. Drones, 3(2), 40.