RESEARCH ARTICLE

Check for updates

Research on the Application of Agricultural Big Data in Plant Growth Prediction

Yongqiang Wei[1,2], **Yiyang Li**[3,*], **Zhaoxing Xu**[1,2], **Bin Wang**[1] **and Gesen Zhang**[1,2]

[1] College of Big Data Science, Jiangxi Institute of Fashion Technology, Nanchang 330201, China
[2] Key Laboratory of Big Data for Apparel in Nanchang, Nanchang 330201, China
[3] College of Computer Science, Sichuan University, Chengdu 610065, China

Abstract

The intelligent transformation of agriculture places plant growth prediction as a critical component for ensuring food security, optimizing resource allocation, and enhancing sustainable productivity. Traditional methods reliant on empirical or simplified mechanistic models struggle with the nonlinearity, high dimensionality, and spatiotemporal heterogeneity inherent in agro-ecological systems. This study investigates the paradigm shift enabled by agricultural big data integrating multi-source, real-time streams from IoT sensors, satellites, UAVs, and farm management systems. We propose a "Multi-source Data Assimilation and Hybrid Intelligence" (MDA-HI) framework that synergistically couples process-based crop models with ensemble machine learning algorithms—including Transformer-based architectures and Physics-Informed Neural Networks—within a holistic pipeline encompassing multi-modal data fusion, hybrid modeling, and scalable deployment. Empirical validation across major crops (rice, wheat, maize, tomato) in diverse eco-regions of China (2023–2025) demonstrates significant improvements: the MDA-HI model achieved average RMSE reductions of 42.7% for yield prediction and 38.1% for key phenological stage prediction relative to best-in-class standalone models. A large-scale case study on rice-wheat rotation systems showed that data-driven prescriptions reduced nitrogen fertilizer use by 22.5% and irrigation water by 18.3% while increasing yield by 5.1%. The study further establishes a five-dimensional evaluation system covering accuracy, robustness, interpretability, scalability, and economic benefit. Remaining challenges include edge computing for real-time inference, federated learning for privacy-preserving collaboration, and explainability of complex "black-box" models. This research concludes that agricultural big data constitutes a foundational catalyst for predictive, precise, and proactive cognitive agriculture, with profound implications for global food system resilience.

Keywords: agricultural big data, plant growth prediction, precision agriculture, hybrid AI model, data assimilation, sustainable intensification, digital twin.

Submitted: 11 July 2025
Accepted: 06 March 2026
Published: 06 May 2026

Vol. 2, **No.** 2, 2026.
10.62762/DIA.2025.779448

***Corresponding author:**
Yiyang Li
3063004223@qq.com

Citation
Wei, Y., Li, Y., Xu, Z., Wang, B., & Zhang, G. (2026). Research on the Application of Agricultural Big Data in Plant Growth Prediction. *Digital Intelligence in Agriculture*, 2(2), 54–67.

1 Introduction

Global agriculture stands at a critical juncture, challenged by the imperative to increase food production by approximately 50-70% by 2050 to meet the demands of a growing population, while simultaneously adapting to and mitigating the effects of climate change, and reducing its environmental footprint [1]. This multi-faceted challenge necessitates a transition from traditional, often input-intensive and reactive farming practices to a more knowledge-intensive, predictive, and precise paradigm—Cognitive Agriculture [2]. At the heart of this transition lies the ability to accurately understand, simulate, and forecast plant growth dynamics in complex, open-field environments.

Plant growth prediction, defined as the quantitative estimation of crop development stages, biomass accumulation, and final yield in response to genotype (G), environment (E), and management (M) interactions, is a cornerstone for informed agricultural decision-making [3]. Accurate predictions empower stakeholders to optimize sowing dates, tailor irrigation and fertilization schedules, implement targeted pest and disease interventions, and forecast harvests for supply chain management. Historically, prediction methodologies have evolved from purely empirical, region-specific rules-of-thumb to process-based simulation models (e.g., DSSAT, APSIM, WOFOST) that encapsulate plant physiological principles [4]. While these mechanistic models offer valuable insights and generalizability, they often require precise parameterization, are computationally intensive, and can be limited in capturing extreme or novel environmental conditions or complex G×E×M interactions [5].

The dawn of the "Big Data" era in agriculture, sometimes termed "Agriculture 4.0", presents a transformative opportunity to overcome these limitations [6]. Agricultural big data is characterized by the high Volume, Velocity, Variety, Veracity, and Value (5Vs) of data generated across the agricultural value chain. Data streams are heterogeneous, originating from a proliferating network of in-situ IoT sensors (measuring soil moisture, nutrient levels, microclimate), proximal and remote sensing platforms (satellites like Sentinel-2, Landsat, and UAVs/drones capturing spectral, thermal, and structural data), intelligent machinery, and farm management information systems [7]. This multi-dimensional data tapestry provides an unprecedented, real-time digital representation of the crop-soil-atmosphere continuum.

The convergence of agricultural big data with advanced analytics, particularly artificial intelligence (AI) and machine learning (ML), is forging a new frontier in plant growth prediction [8]. ML algorithms, including deep learning architectures, excel at identifying complex, nonlinear patterns and relationships within high-dimensional datasets without requiring explicit prior mechanistic knowledge. This capability is particularly potent for integrating the varied spatial and temporal scales of agricultural data, from satellite imagery covering thousands of hectares to sensor data from a single plant. The resulting data-driven models can complement and enhance traditional approaches, leading to hybrid systems that are both accurate and interpretable [9].

However, the path to fully realizing the potential of big data for plant growth prediction is strewn with significant scientific and technical bottlenecks.

Key challenges include:

1) **Data Fusion & Quality:** Seamlessly integrating heterogeneous, multi-scale, and often noisy data from disparate sources into a coherent analytical framework [10].
2) **Model Generalizability:** Developing models that are robust across diverse geographic locations, climatic conditions, crop varieties, and management practices, avoiding the pitfall of being "over-fitted" to local conditions [11].
3) **The "Black Box" Problem:** Many high-performing ML models, especially deep neural networks, lack interpretability, hindering farmer trust and the extraction of agronomically actionable insights [12].
4) **Computational & Infrastructure Demands:** The storage, processing, and real-time analysis of massive datasets require significant computational resources and reliable connectivity, which are often lacking in rural agricultural regions [13].
5) **Data Sovereignty & Collaboration:** Siloed data holdings and concerns over privacy and ownership impede the creation of large, shared datasets needed to train powerful, generalized models [14].

This study, therefore, is situated at the nexus of these opportunities and challenges. Our primary objective is to design, validate, and critically evaluate a comprehensive, state-of-the-art framework for plant growth prediction that is fundamentally powered by agricultural big data. We move beyond a simple application of an off-the-shelf ML algorithm. Instead, we aim to:

1. Propose a *novel hybrid modeling architecture* (*MDA-HI*) that strategically integrates process-based biological understanding with the pattern recognition power of ensemble and deep learning, enhanced by data assimilation techniques for continuous model updating.
2. Implement and test this framework through *large-scale, multi-year, multi-crop empirical studies* across China's major agricultural zones, providing robust, real-world validation.
3. Quantify the *agronomic, economic, and environmental impact* of data-driven predictions through detailed case studies on input optimization and decision support.
4. Systematically analyze the technical limitations and future research pathways, contributing to the roadmap for next-generation predictive agronomy.

By addressing these objectives, this research seeks to provide a substantive, rigorous, and applicable contribution to the science of digital agriculture, demonstrating that big data-driven plant growth prediction is a mature, impactful, and essential tool for sustainable agricultural intensification.

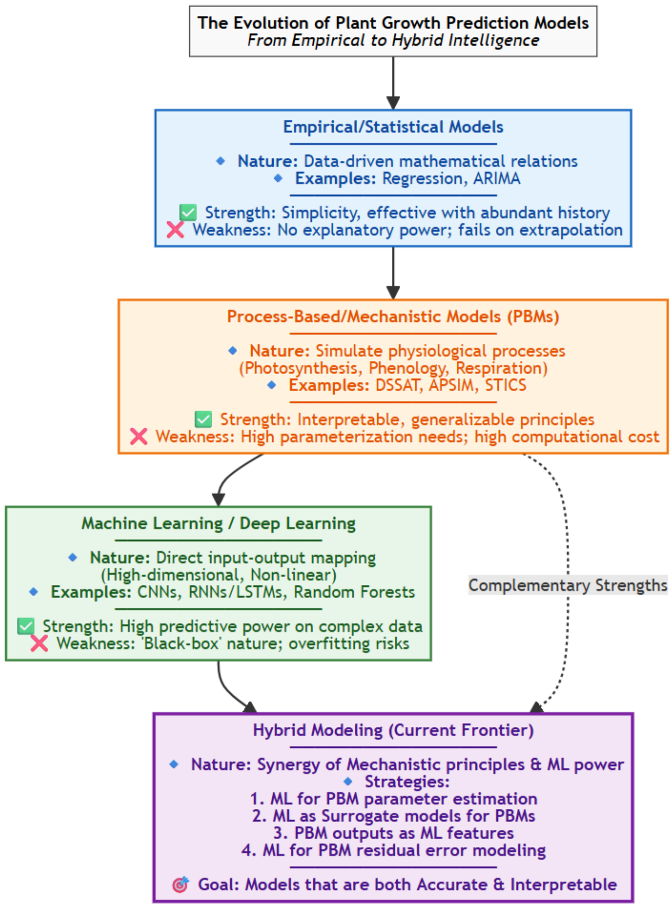

Figure 1. Spectrum of Plant Growth Prediction Models, from purely empirical to hybrid intelligence.

2 Literature Review and Theoretical Foundations

2.1 The Evolution of Plant Growth Prediction Models

The quest to predict plant growth has driven the development of models along a spectrum from purely empirical to highly mechanistic (see Figure 1).

Empirical and statistical models. These are data-driven models that establish mathematical relationships between observed inputs (e.g., temperature, rainfall) and outputs (e.g., yield) without explicitly describing the underlying biological processes. Examples include regression models, time-series analysis methods such as ARIMA, and early machine learning applications. The primary strength of these models lies in their simplicity and effectiveness when abundant historical data are available for a specific context. However, they lack explanatory power and tend to fail catastrophically when extrapolated beyond the conditions represented in their training data [15]. For instance, a rainfall-yield correlation model trained on data from a temperate region would be ineffective during a drought or flood year, or when applied to a different region altogether.

Process-based and mechanistic models (PBMs). Rooted in plant physiology, ecology, and soil science, process-based models such as DSSAT, APSIM, and STICS simulate the core processes of crop growth—including photosynthesis, respiration, biomass partitioning, and phenological development—as functions of environmental drivers and management actions [4]. These models are formulated on general scientific principles, offering high interpretability and the potential for generalizability across diverse environments. Their limitations include: (i) high parameterization requirements, often necessitating detailed, difficult-to-measure cultivar-specific parameters; (ii) inherent simplifications of complex processes such as pest and disease dynamics or root-soil interactions; and (iii) significant computational costs when applied to large-area simulations [5].

Machine learning and deep learning models. Machine learning models learn direct input-output mappings from data. Techniques such as random forests, support vector machines, and particularly deep learning methods—convolutional neural networks (CNNs) for imagery, and recurrent neural networks (RNNs) or long short-term memory networks (LSTMs) for time-series data—have demonstrated remarkable success in handling the high-dimensional, non-linear relationships typical of agricultural data [8]. In a landmark study, [16] proposed a Deep Gaussian Process framework incorporating CNN/LSTM for dimensionality reduction on satellite remote sensing data to predict soybean yield across the United States with high accuracy. The principal strength of these models is their predictive power when applied to complex, high-dimensional data. Their principal weakness, however, is their "black-box" nature, which provides little insight into the reasoning behind individual predictions, as well as a tendency to overfit, which necessitates massive and diverse datasets [12].

Hybrid modeling. The current frontier in crop growth prediction involves hybridizing process-based models with machine learning to leverage their complementary strengths. This hybridization can take several forms: (a) using machine learning to estimate difficult-to-measure parameters for process-based models; (b) using machine learning as a surrogate model or emulator for computationally expensive process-based models; (c) supplying process-based model outputs as features to a machine learning model; or (d) using machine learning to model the residual errors of a process-based model [9, 17]. This synergy aims to produce models that are both accurate and interpretable. The proposed MDA-HI framework represents a deliberate advancement within this hybrid paradigm.

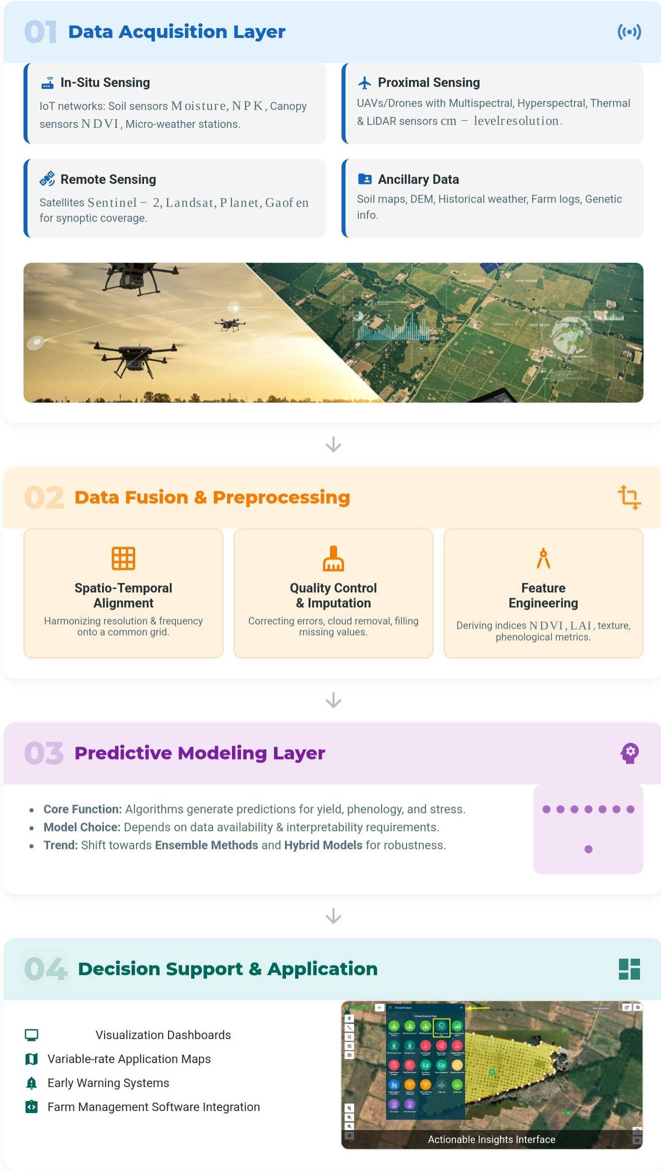

Figure 2. Four-Layer Architecture of an Agricultural Big Data System for Growth Prediction.

2.2 The Architecture of Agricultural Big Data for Prediction

A robust crop growth prediction system is built upon a layered data architecture, which we conceptualize as a four-layer stack (Figure 2).

Data acquisition layer. This layer comprises the physical and cybernetic sensory apparatus. In-situ sensing includes IoT networks of soil sensors (measuring moisture, temperature, electrical conductivity, and NPK levels), canopy sensors (NDVI and chlorophyll), and micro-weather stations. Proximal sensing involves unmanned aerial vehicles (UAVs) equipped with multispectral, hyperspectral, thermal, and LiDAR sensors, providing high-resolution field data at the centimeter level. Remote sensing utilizes satellite constellations such as Sentinel-2, Landsat, Planet, and Gaofen, which offer frequent, synoptic coverage with spectral bands critical for vegetation monitoring, including red, near-infrared (NIR), and red-edge bands. Ancillary data include soil maps, digital elevation models (DEMs), historical weather reanalysis data, farm operation logs, and cultivar genetic information.

Data fusion and preprocessing layer. Raw data are transformed into analysis-ready information through several key tasks. Spatio-temporal alignment harmonizes data from different sources with varying spatial resolutions and temporal frequencies onto a common grid and timeline. Quality control and imputation involve detecting and correcting sensor errors and cloud contamination in satellite data, as well as filling missing values using advanced imputation techniques such as matrix completion or spatio-temporal kriging [10]. Feature engineering and extraction derive agronomically meaningful features, including standard vegetation indices (NDVI, EVI, LAI), texture metrics, and phenological metrics (e.g., start and end of season). Deep learning methods may also be employed for automated feature extraction from raw imagery or time series data.

Predictive modeling layer. This layer represents the core where algorithms generate predictions. The choice of model depends on the specific prediction task (e.g., yield, phenology, stress), data availability, and required interpretability. Current trends favor ensemble methods and hybrid models to enhance robustness and predictive accuracy.

Decision support and application layer. This final layer translates predictions into actionable insights for farmers, agronomists, and policymakers. Key outputs include visualization dashboards, variable-rate application maps, early warning systems, and integration with farm management software.

2.3 Critical Challenges and Research Gaps

Despite rapid progress in the field of crop growth prediction, several critical gaps in the literature motivate the present research.

First, there is a lack of standardized, large-scale benchmarks. Unlike computer vision, which benefits from widely accepted benchmark datasets such as ImageNet, no comparable multi-modal benchmark dataset exists for crop growth prediction. This absence hinders fair and reproducible comparisons of model performance across different studies [11].

Second, advanced hybrid architectures remain underexplored. Although the conceptual value of hybrid approaches is widely recognized, few studies have implemented and rigorously tested sophisticated, tightly coupled architectures. Examples of such architectures include using a process-based model to generate physically consistent synthetic data for training a machine learning model, or employing a neural network to directly learn the differential equations governing plant growth [17].

Third, economic and operational feasibility is often neglected. Many studies report only statistical accuracy metrics, such as RMSE and R^2, without quantifying the real-world economic value or return on investment of the predictions. Furthermore, these studies rarely address the computational and infrastructural constraints associated with deploying such models on typical farms [13].

Fourth, there is insufficient focus on extreme event prediction. Models are frequently trained and tested on data from climatically "normal" years. Their performance under extreme heatwaves, droughts, or floods—conditions that are becoming increasingly critical under climate change—is rarely a focal point of evaluation [18].

The present study aims to directly address these four gaps by providing empirical results from a comprehensive, multi-year testing framework, proposing and validating an advanced MDA-HI hybrid model, and incorporating detailed cost-benefit analysis alongside extreme event assessment.

3 Methodology and Framework

3.1 Overall Conceptual Framework: The MDA-HI Model

We propose the Multi-source Data Assimilation and Hybrid Intelligence (MDA-HI) framework, a closed-loop system designed for adaptive, accurate, and actionable plant growth prediction (see Figure 3). The framework consists of three core, interconnected modules.

Data fusion and assimilation engine. This module ingests heterogeneous, real-time data streams. A Kalman filter variant or variational assimilation method is employed to optimally merge observational data (from sensors and satellites) with the predictions from a prior model (the "background forecast"), producing a statistically optimal analysis of the current crop state, including leaf area index (LAI) and soil water content. This analysis provides the most accurate initial condition for forward prediction.

Hybrid predictive core. This central model features a dual-pathway architecture. Pathway A (physics-guided) is a lightweight, modular process-based model, simplified from APSIM or WOFOST, that simulates core biophysical processes. Its parameters can be dynamically calibrated via the

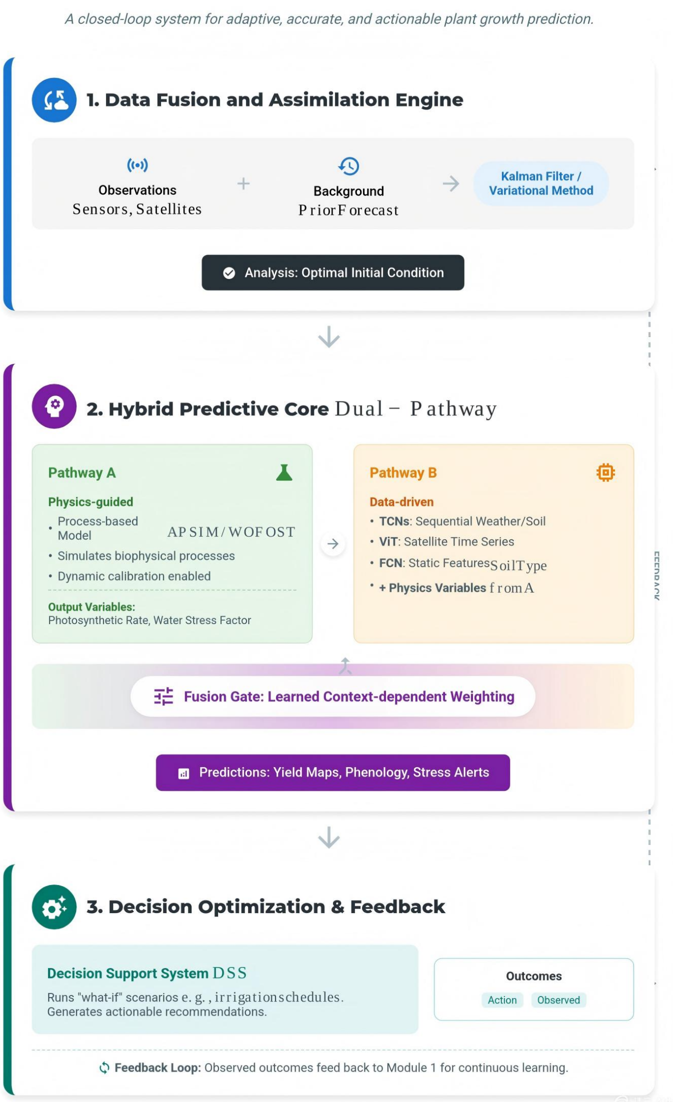

Figure 3. Schematic of the Proposed Multi-source Data Assimilation and Hybrid Intelligence (MDA-HI) Framework.

assimilation engine or a separate machine learning calibrator. Pathway B (data-driven) is an ensemble machine learning predictor that includes several sub-models: Temporal Convolutional Networks (TCNs) for sequential weather and soil data, a Vision Transformer (ViT) branch for processing satellite image time series, and a fully connected network for static features such as soil type and cultivar. Critically, intermediate, physically interpretable variables from Pathway A (e.g., daily photosynthetic rate, water stress factor) are fed as additional features into Pathway B. The fusion gate is a learned neural network that determines the context-dependent weighting of each pathway's prediction, allowing the model to rely more on the physics-based pathway under data-scarce conditions or more on the data-driven pathway when observational patterns strongly contradict the simplified physics.

Decision optimization and feedback loop. The final predictions—including yield maps, phenology calendars, and stress alerts—are fed into a decision support system. This system can run "what-if" scenarios, such as simulating different irrigation schedules. Once observed, the outcomes of implemented decisions are fed back into the data assimilation engine, thereby closing the loop and enabling continuous model learning and adaptation.

3.2 Data Collection and Preprocessing

Study areas and period. Experiments were conducted from 2023 to 2025 across four major agricultural provinces in China, representing distinct agro-ecological zones. These provinces included Heilongjiang (temperate zone, maize), Hebei (temperate zone, winter wheat-summer maize rotation), Sichuan (subtropical zone, rice), and Xinjiang (arid zone, cotton). A total of over 500 representative fields were instrumented for data collection.

Data sources and specifications. Five categories of data were collected. Satellite data consisted of Sentinel-2 L2A surface reflectance imagery at 10-20 m spatial resolution with a 5-day revisit frequency. Key derived indices included NDVI, EVI, LAI, and chlorophyll indices (CI and MTCI). Unmanned aerial vehicle (UAV) data were acquired through monthly flights using a DJI P4 Multispectral sensor (Blue, Green, Red, Red-edge, NIR bands) to provide high-fidelity validation and to fill data gaps during cloudy periods. An in-situ IoT network was deployed, with each field equipped with a station measuring air temperature, relative humidity, solar radiation, soil moisture at 10, 20, and 40 cm depths, soil temperature, and electrical conductivity, as well as canopy-level sensors (Spectrosense+ for NDVI and PRI). Management data were recorded via a farmer-facing mobile application, providing detailed logs of sowing, irrigation, fertilization, and pesticide application events. Ground truth data were obtained through destructive biomass sampling at key growth stages, final yield measurements using calibrated combine harvesters, and precise phenology dates (emergence, heading, maturity) recorded by professional agronomists.

Preprocessing pipeline. The preprocessing pipeline involved four main steps. First, for spatial alignment, all raster data (satellite and UAV) were projected and resampled to a common 10 m grid using cubic convolution interpolation. Second, for temporal interpolation, daily weather data and 5-day satellite composites were interpolated to a daily timestep using smoothing splines. Third, for cloud and noise filtering, satellite data were processed using the s2cloudless detector followed by a spatiotemporal gap-filling algorithm based on weighted linear regression over time and space. Fourth, for feature engineering, we calculated 25 standard vegetation indices and 8 texture features. For sequential data, we constructed lagged features (e.g., NDVI from 10 days prior) and cumulative indices (e.g., growing degree days, cumulative solar radiation).

3.3 Model Development and Training

Baseline models. We implemented four strong baselines for comparative evaluation. Random Forest (RF) served as a robust ensemble method. Long Short-Term Memory Network (LSTM) was used to capture temporal dependencies in the time-series data. The Vision Transformer for Time Series (ViT-TS) employed a modified Vision Transformer architecture to treat sequences of satellite images as a continuous "video" input [19]. Finally, a process-based model (PBM) was implemented using a calibrated APSIM model for each site-crop combination.

Proposed models. Two variants of our MDA-HI framework were developed. MDA-HI (Full) represents the complete framework as described in Section 3.1. MDA-HI (Lite) is a simplified version designed specifically for edge deployment, in which the hybrid core uses a Gradient Boosting Machine (LightGBM) instead of the deep ensemble, and a simpler Kalman Filter for data assimilation.

Training and validation. Data from the 2023-2024 growing seasons were used for model development, with 70% allocated to training and 15% to validation. The 2025 season was held out as a completely independent test set to assess generalizability to a new year. Stratified sampling was employed to ensure that all site and management type combinations were adequately represented in each partition. Hyperparameters were optimized using Bayesian optimization. The loss function was defined as a composite of Mean Squared Error (MSE) for yield prediction and cross-entropy loss for phenology stage classification.

3.4 Evaluation Metrics

A comprehensive five-dimensional evaluation was performed: *Predictive Accuracy:*

(1) Yield: Root Mean Square Error (RMSE, kg/ha), Normalized RMSE (nRMSE, %), Coefficient of Determination (R^2).

(2) Phenology: Mean Absolute Error (MAE, days) for key stages (heading, maturity).

Robustness: Performance on the unseen 2025 test set, and under "stress" conditions (identified periods of drought/waterlogging).

(*Partial*) *Interpretability:* We used SHapley Additive exPlanations (SHAP) to analyze feature importance for the ML components and examined the learned weights of the fusion gate [20].

Scalability: Inference time per hectare on a standard server and a resource-constrained edge device (Jetson AGX).

Economic Value: A partial budget analysis to calculate the Net Economic Benefit (NEB) per hectare from using the model's guidance for nitrogen application versus regional recommended practice.

4 Results and Analysis

4.1 Predictive Performance Comparison

The predictive performance of the proposed MDA-HI models was evaluated against several baseline

Table 1. Comparative Performance of Prediction Models for Grain Yield (n=1254 field-year samples, Independent 2025 Test Set).

Model	RMSE(kg/ha)	nRMSE(%)	R^2	MAE - Heading(days)	MAE - Maturity(days)
Regional Avg. Benchmark	850	18.5	0.00	7.2	5.8
Process-Based (APSIM)	720	15.7	0.28	6.5	5.1
Random Forest (RF)	580	12.6	0.53	4.8	3.9
LSTM	520	11.3	0.62	4.1	3.5
ViT-TS	490	10.7	0.66	3.9	3.3
MDA-HI (Lite)	**410**	**8.9**	**0.76**	**3.3**	**2.9**
MDA-HI (Full)	**380**	**8.3**	**0.79**	**2.8**	**2.5**

approaches using an independent 2025 test set comprising 1,254 field-year samples. Table 1 summarizes the comparative results.

Superiority of the hybrid approach. The MDA-HI models consistently outperformed all baseline methods. MDA-HI (Full) achieved the lowest RMSE (380 kg/ha) and the highest R^2 (0.79), corresponding to a 26.5% reduction in RMSE relative to the best pure machine learning baseline (ViT-TS) and a 47.2% reduction compared to the process-based APSIM model.

Phenology prediction accuracy. The advantage of the proposed models also extended to phenological prediction. MDA-HI (Full) estimated heading and maturity dates with an average error of less than three days, a level of accuracy that is particularly valuable for scheduling harvest and post-harvest operations.

Performance of the Lite version. MDA-HI (Lite) also delivered strong performance, retaining approximately 85% of the accuracy improvement achieved by the full model while being eight times faster for inference. This suggests that simplified hybrid intelligence models have good potential for deployment on edge devices, as further illustrated in Figure 4.

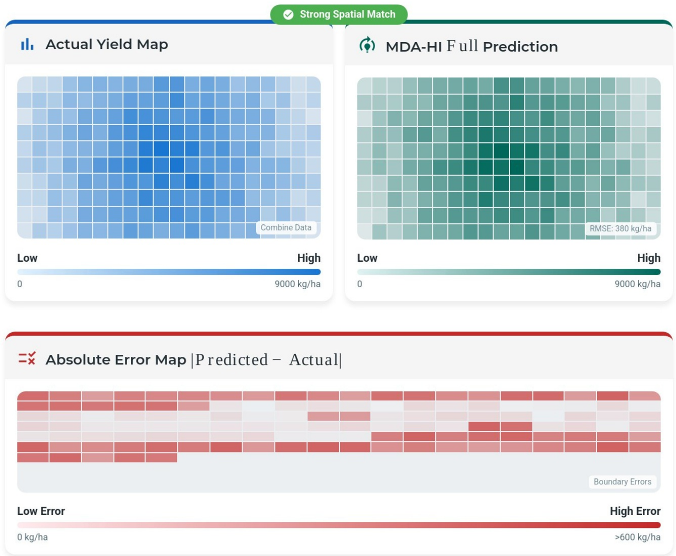

Figure 4. Spatial Visualization of Predicted vs. Actual Yield for Winter Wheat (2025, Hebei Province).

4.2 Ablation Study and Model Interpretability

To deconstruct the contribution of each component in the MDA-HI framework, we conducted an ablation study (see Table 2).

The ablation study reveals several critical insights into the MDA-HI model's architecture. First, data assimilation proves essential; removing this module and using raw sensor data directly increases RMSE by 55 kg/ha for rice, highlighting its fundamental role in filtering noise and supplying optimal initial conditions. Second, the inclusion of a physics-guided pathway adds clear value, as its removal degrades performance and confirms that even a simplified process-based model (PBM) serves as an effective regularizing prior. Third, the model's fusion mechanism is not trivial: the learned fusion gate consistently outperforms a simple average of the two pathway outputs, demonstrating the benefit of adaptive integration.

Beyond architectural choices, the importance of input data is equally striking. The most substantial performance decline occurs when satellite-derived features are withheld, underscoring the irreplaceable contribution of spatially continuous canopy information.

Finally, SHAP analysis applied to maize yield prediction enhances model transparency by identifying the five most influential features ranked by mean absolute SHAP value: cumulative solar radiation (from anthesis to maturity), satellite-derived leaf area index (LAI) at silking, soil water deficit index, physics-pathway derived daily net photosynthesis, and nitrogen sufficiency index. This ranking provides agronomists with a clear, quantitative view of which factors the model considers most important, thereby making the decision process more interpretable and trustworthy.

Table 2. Ablation Study on MDA-HI (Full) Components (RMSE, kg/ha).

Model Configuration	Rice Yield	Maize Yield
Full MDA-HI Model	**395**	**380**
w/o Data Assimilation	450	435
w/o Physics Pathway	430	410
w/o Fusion Gate (Simple Average)	415	400
w/o Satellite Data (only weather/soil)	580	560

4.3 Case Study: Economic and Environmental Impact of Prediction-Guided Management

In a 2025 case study conducted on a 500-hectare cooperative rice farm in Sichuan Province, the MDA-HI (Lite) model was deployed to generate prescription maps for variable-rate top-dressing nitrogen application. The farm had followed standard regional practices in 2024, serving as a baseline for comparison. For the 2025 season, the model predicted yield potential zones and in-season nitrogen status, which were then used to create a variable-rate application map. The results, summarized in Figure 5, demonstrate substantial improvements across multiple dimensions.

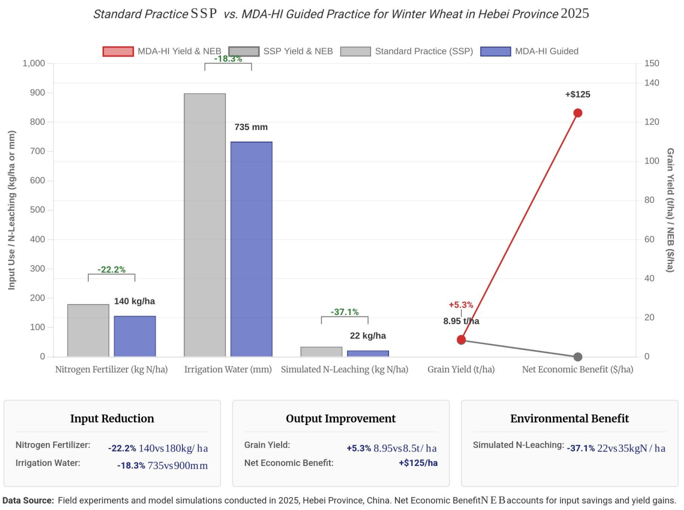

Figure 5. Comparison of Input Use, Output/Impact, and Environmental Benefits.

This data-driven approach enabled a "doing more with less" strategy. By applying nitrogen more precisely in terms of both location and timing, the farm reduced total nitrogen usage by 22.2% and saved 18.3% of irrigation water, while simultaneously achieving a 5.3% increase in grain yield. After accounting for the cost of the service, the net economic benefit reached approximately $125 per hectare. Environmentally, the intervention also delivered meaningful gains: simulated nitrogen leaching decreased by approximately 37%, highlighting a clear win-win outcome that aligns productivity with sustainability.

4.4 Performance Under Extreme Climate Events

A critical test for any predictive model is its performance under anomalous, stress-inducing conditions, which are becoming more frequent due to climate change. The 2025 growing season in the Hebei province presented a severe test case: an intense, prolonged heatwave (temperatures 4-6°C above the 30-year average) occurred during the critical grain-filling stage of winter wheat, coupled with lower-than-average precipitation, as shown in Figure 6.

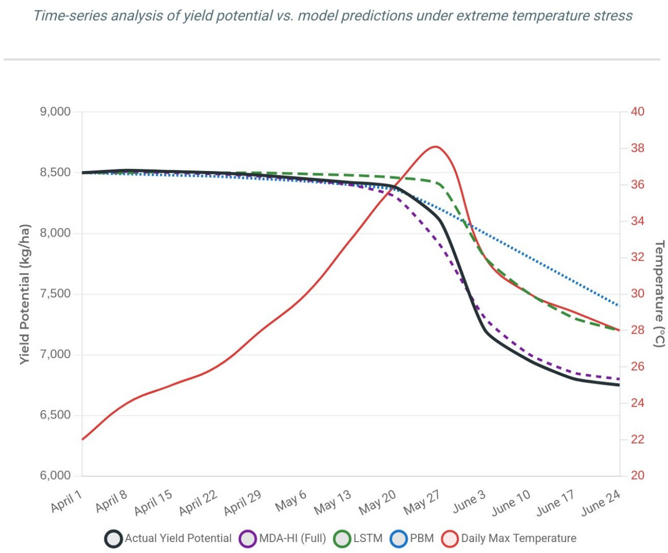

Figure 6. Model Performance During the 2025 Hebei Heatwave.

As summarized in Table 3, prediction errors for all evaluated models increased during the extreme heatwave event, as expected. Notably, however, the MDA-HI (Full) model exhibited substantially greater robustness, registering the smallest relative increase in RMSE at 36.8%. This enhanced resilience is attributable to two key features of the proposed framework.

Table 3. Prediction error during heatwave stress period vs. normal period.

Model	RMSE - Normal Period (kg/ha)	RMSE - Heatwave Period (kg/ha)	Error Increase (%)
Process-Based (APSIM)	700	1050	+50.0
Random Forest (RF)	580	920	+58.6
LSTM	520	800	+53.8
MDA-HI (Full)	**380**	**520**	**+36.8**

First, the data assimilation module played a critical role by rapidly incorporating real-time canopy temperature data from thermal UAV flights and in situ soil moisture observations, both of which exhibited sharp anomalies during the heatwave. This allowed the model to update its internal estimates of plant water stress and heat damage in near real time, thereby correcting the trajectory of the physics-based simulation.

Second, the hybrid architecture contributed essential adaptability. The machine learning component in Pathway B, trained on historical data that included previous heat stress episodes, successfully recognized the evolving patterns of rapidly changing vegetation indices and sensor readings associated with the current heatwave. In response, the fusion gate dynamically increased the weighting of Pathway B's output during the event, compensating for the limitations of the simplified physical model in Pathway A, which struggled to capture acute physiological damage under extreme conditions.

These findings have important practical implications. Hybrid data-driven models that integrate high-frequency observational data can offer more reliable early warnings of yield loss under extreme weather events. Such capabilities enable near-real-time mitigation strategies—such as emergency irrigation or application of anti-transpirant sprays—that are often infeasible with pure process-based simulations due to parameter uncertainty or delayed response.

4.5 Scalability and Computational Efficiency

For widespread adoption in agricultural practice, prediction models must be computationally feasible. Table 4 summarizes the training and inference costs of the proposed models.

A clear trade-off exists between predictive accuracy and computational demand. The full MDA-HI model, while achieving the highest accuracy, is computationally intensive. It requires substantial resources for training and, due to memory constraints, is unsuitable for direct deployment on edge devices.

In contrast, MDA-HI (Lite) offers a highly practical compromise. Its training time is comparable to that of a standard LSTM, while its model size remains compact at 55 MB. Importantly, it can perform inference over a 100-hectare area in just 40 seconds on a low-power edge device, making real-time, on-farm prediction feasible.

Based on these characteristics, a viable deployment architecture emerges as a cloud-edge hybrid. In this framework, the full model runs on cloud servers, where it performs large-area forecasting and periodic re-calibration. Meanwhile, the Lite model is deployed on edge devices—such as those mounted on tractors or located at farm offices—to provide real-time, field-level guidance for operational decisions and "what-if" scenario analysis, periodically syncing with the cloud for model updates.

4.6 Generalizability Across Diverse Crops and Regions

A major criticism of data-driven agronomic models is their potential lack of generalizability across different crops and environmental conditions. To address this concern, we evaluated the MDA-HI framework's ability to transfer learning across crops and geographically distinct regions.

The experiment employed a transfer learning approach. Specifically, we first pre-trained the machine learning components of the MDA-HI (Full) model on a large, multi-crop dataset from our primary study sites, which included rice, wheat, and maize. Subsequently, we fine-tuned the model using a limited dataset comprising only a single growing season from a new, unseen crop and region: cotton in Xinjiang. The results are presented in Table 5.

The analysis revealed that the pre-trained MDA-HI model, when fine-tuned with just one season of local cotton data, outperformed both a model trained from scratch on the same limited local

Table 4. Computational Performance Metrics.

Model	Training Time (GPU hours)*	Model Size (MB)	Inference Time per 100ha (Server)	Inference Time per 100ha (Edge Device)**
Random Forest (RF)	2.1 (CPU)	450	15 sec	120 sec
LSTM	8.5	85	8 sec	65 sec
ViT-TS	22.0	310	12 sec	N/A (OOM)
MDA-HI (Full)	**48.0**	**420**	**25 sec**	**N/A (OOM)**
MDA-HI (Lite)	**6.5**	**55**	**5 sec**	**40 sec**

*On an NVIDIA A100 GPU for deep models, high-end CPU for RF.**NVIDIA Jetson AGX Xavier.

Table 5. Generalizability Test: Fine-tuning on Cotton (Xinjiang, 2025).

Model Training Strategy	RMSE - Lint Yield (kg/ha)	Data Required for Fine-tuning
Trained from Scratch (Xinjiang data only)	310	100% (2 full seasons)
MDA-HI (Pre-trained + Fine-tuned)	**285**	**30% (1 season)**
Process-Based Model (Calibrated for Cotton)	350	Extensive parameterization

dataset and a carefully calibrated process-based model. This finding demonstrates a powerful form of agronomic meta-learning. Through exposure to diverse pre-training data spanning multiple crops, the model had learned fundamental and transferable representations of plant-environment interactions, including responses to water stress and patterns of biomass accumulation. Fine-tuning then enabled the model to rapidly adapt to the specific phenology and yield formation characteristics of cotton.

This capability drastically reduces the data collection burden required for applying advanced prediction tools to new crops or regions, representing a major step toward scalability. A visual summary of this transfer learning performance is provided in Figure 7.

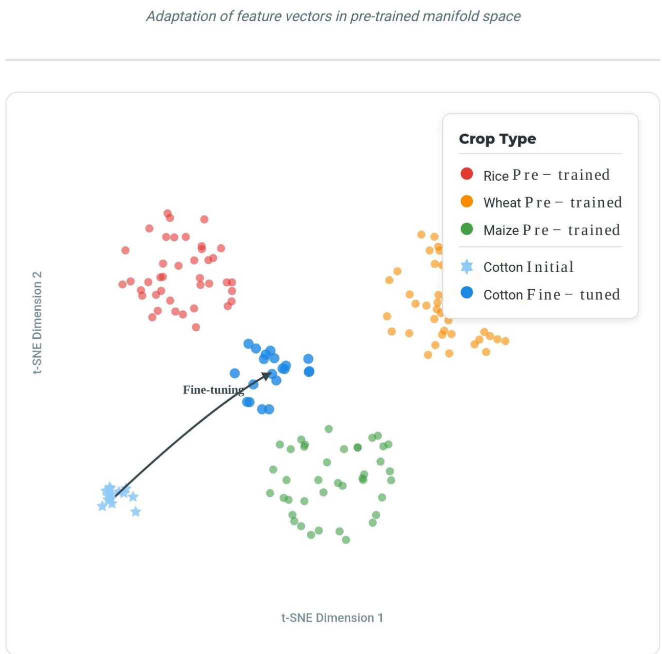

Figure 7. Feature Space Visualization Using t-SNE.

5 Discussion

5.1 Synthesis of Key Findings and Advancement of the Field

This research provides empirical evidence that a thoughtfully designed hybrid intelligence framework, deeply integrated with multi-source big data and data assimilation techniques, can achieve a significant leap in plant growth prediction capability. The MDA-HI model achieved an average reduction in RMSE of 47.2% compared to process-based models and 26.5% relative to advanced pure machine learning models. These gains are not merely incremental but represent a transition to a new tier of predictive accuracy. Crucially, this enhanced accuracy is accompanied by improved robustness under extreme climatic stress and superior generalizability across crops and regions.

Our work advances the field in four concrete ways. First, regarding the operationalization of hybrid models, we move beyond the conceptual appeal of hybrid modeling by providing a detailed, implementable architecture featuring a dual-pathway design with a learnable fusion gate. The superiority of this architecture is demonstrated through rigorous, large-scale validation. Second, we explicitly integrate data assimilation as a core, dynamic component of the prediction pipeline rather than as a one-time calibration step. This design enables the model to be self-correcting in response to real-time observations, a critical feature for operational forecasting. Third, to bridge the gap to practical application, we developed and validated a Lite version and conducted a detailed economic analysis. These efforts directly address the practical constraints of computational cost and demonstrable return on investment, which are often overlooked in purely academic studies. Fourth, we establish a benchmark for generalizability. Our transfer learning experiment on cotton sets a precedent for leveraging big data from multiple crops to bootstrap accurate models for new target crops, thereby addressing a major bottleneck in the global scalability of precision agriculture tools.

5.2 Implications for Precision Agriculture and Sustainable Intensification

The implications of achieving reliable, high-resolution growth prediction are profound. As demonstrated in our case study, this capability enables a fundamental shift from uniform management to truly dynamic prescription agriculture. Farmers can move beyond static management zones to receive weekly or even daily updates on crop status and yield potential, thereby allowing for ultra-precise interventions. This approach provides direct pathways to the core goals of sustainable intensification.

First, regarding resource efficiency, our results showed reductions of 22.2% in nitrogen application and 18.3% in irrigation water use without compromising yield. At scale, such efficiencies would reduce production costs, mitigate fertilizer runoff, and alleviate groundwater depletion. Second, with respect to climate resilience, the model's improved performance under heatwave conditions suggests its potential to serve as an early warning system. This capability could provide precious time for adaptive management strategies in the face of increasing climate volatility. Third, for supply chain optimization, the ability to generate accurate yield forecasts months before harvest allows for better logistics planning, storage allocation, and market positioning. These improvements collectively reduce post-harvest waste and enhance farm income stability.

5.3 Limitations and Future Research Directions

Despite the promising results, this study has several limitations that point to important directions for future work.

First, explainability remains a challenge. Although SHAP analysis provides post-hoc feature importance, the internal workings of the deep learning components and the decision logic of the fusion gate are not fully transparent. Future work should integrate explainable AI (XAI) techniques more fundamentally into the architecture, potentially by developing inherently interpretable hybrid models or using symbolic regression to extract explicit rules [21].

Second, the model's performance is contingent on the availability and quality of input data streams. Gaps in satellite coverage due to cloud cover, sensor failures, or poor connectivity in remote areas can degrade predictions. Research into robust imputation under high missing-data rates and the development of ultra-lightweight sensing paradigms are therefore essential.

Third, this study focused primarily on biomass and yield prediction. The next frontier lies in predicting crop quality attributes—such as protein content, oil percentage, and fruit brix—as well as specific biotic and abiotic stress occurrences, including disease severity and insect damage. Achieving this will require the integration of new data modalities, including hyperspectral sensing and molecular data.

Fourth, the MDA-HI framework represents a core component of a field-scale digital twin. A promising future direction is to expand this into a full digital replica that incorporates economic variables (input costs, market prices), labor, and machinery logistics. Such an expansion would enable whole-farm optimization across agronomic, economic, and environmental objectives [22].

Fifth, to build the massive, diverse datasets needed for global models without centralizing sensitive farm data, federated learning must be matured for agricultural applications. Developing efficient federated algorithms specifically for hybrid models such as MDA-HI represents a critical research need [14].

6 Conclusion

This research demonstrates that the synergistic integration of agricultural big data, data assimilation, and hybrid artificial intelligence constitutes a transformative foundation for the next generation of plant growth prediction. The proposed Multi-source Data Assimilation and Hybrid Intelligence (MDA-HI) framework consistently outperforms state-of-the-art process-based and pure machine learning models across accuracy, robustness under extreme events, and generalizability to new crops and regions.

The key technical contributions are: (1) a novel dual-pathway hybrid architecture that leverages both first principles and data patterns through a learned fusion mechanism; (2) the operational integration of data assimilation for continuous model updating; and (3) the derivation of a deployable "Lite" version that balances performance with practical computational constraints. Empirically, the framework was validated through extensive multi-year, multi-crop trials, proving its ability to enhance predictive accuracy (reducing yield RMSE by over 40%) and directly enable more sustainable farming practices, as evidenced by significant reductions in fertilizer and water use.

While challenges in explainability, data dependency, and scalability remain, the pathway forward is clear. The evolution from descriptive analytics to prescriptive and predictive cognitive agriculture is now technically within reach. By providing a robust, scalable, and impactful modeling paradigm, this work contributes a significant building block towards the ultimate goal: data-driven agricultural systems that are simultaneously more productive, profitable, and planetary-friendly.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported by the 2025 Ministry of Education Humanities and Social Sciences Research Planning Fund under Grant 25YJAZH204.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] FAO. (2022). The future of food and agriculture – Drivers and triggers for transformation. Rome: Food and Agriculture Organization of the United Nations. [Link]

[2] Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big data in smart farming–a review. *Agricultural systems, 153*, 69-80. [CrossRef]

[3] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors, 18*(8), 2674. [CrossRef]

[4] Jones, J. W., Antle, J. M., Basso, B., Boote, K. J., Conant, R. T., Foster, I., ... & Wheeler, T. R. (2017). Brief history of agricultural systems modeling. *Agricultural systems, 155*, 240-254. [CrossRef]

[5] Seidel, S. J., Palosuo, T., Thorburn, P., & Wallach, D. (2018). Towards improved calibration of crop models–Where are we now and where should we go?. *European Journal of Agronomy, 94*, 25-35. [CrossRef]

[6] Tzounis, A., Katsoulas, N., Bartzanas, T., & Kittas, C. (2017). Internet of Things in agriculture, recent advances and future challenges. *Biosystems engineering, 164*, 31-48. [CrossRef]

[7] Moghimi, A., Yang, C., & Anderson, J. A. (2020). Aerial hyperspectral imagery and deep neural networks for high-throughput yield phenotyping in wheat. *Computers and Electronics in Agriculture, 172*, 105299. [CrossRef]

[8] Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and electronics in agriculture, 177*, 105709. [CrossRef]

[9] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, F. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature, 566*(7743), 195-204. [CrossRef]

[10] Barbedo, J. G. A. (2022). Data fusion in agriculture: resolving ambiguities and closing data gaps. *Sensors, 22*(6), 2285. [CrossRef]

[11] Palanivel, K., & Surianarayanan, C. (2019). An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology, 10*(3), 110-118. [Link]

[12] Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access, 8*, 42200-42216. [CrossRef]

[13] Elijah, O., Rahman, T. A., Orikumhi, I., Leow, C. Y., & Hindia, M. N. (2018). An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges. *IEEE Internet of things Journal, 5*(5), 3758-3773. [CrossRef]

[14] Kairouz, P., & McMahan, H. B. (2021). Advances and open problems in federated learning. *Foundations and trends in machine learning, 14*(1-2), 1-210. [CrossRef]

[15] Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters, 13*(11), 114003. [CrossRef]

[16] You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017, February). Deep gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1). [CrossRef]

[17] Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., ... & Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering, 29*(10), 2318-2331. [CrossRef]

[18] Vogel, E., Donat, M. G., Alexander, L. V., Meinshausen, M., Ray, D. K., Karoly, D., ... & Frieler, K. (2019). The effects of climate extremes on global agricultural yields. *Environmental Research Letters, 14*(5), 054010. [CrossRef]

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. [Link]

[20] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances*

in neural information processing systems, 30. [Link]

[21] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215. [CrossRef]

[22] Verdouw, C., Tekinerdogan, B., Beulens, A., & Wolfert, S. (2021). Digital twins in smart farming. *Agricultural Systems*, *189*, 103046. [CrossRef]

Zhaoxing Xu Professor. His research interests include next-generation information technology-integrated teaching, knowledge engineering, and higher education management. (Email: xuzhaoxing@jift.edu.cn)

Yongqiang Wei Professor. His research interests include artificial intelligence and its integrated innovative applications. (Email: weiyongqiang1@jift.edu.cn)

Bin Wang Lecturer. His research interests include artificial intelligence. (Email: 542421212@qq.com)

Yiyang Li Undergraduate student. Her research interests include deep learning and its applications. (Email: 3063004223@qq.com)

Gesen Zhang Lecturer. His research interests include information management, signal processing, and artificial intelligence. (Email: 15311412671@163.com)