IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

AI-Based Crop Yield Prediction Using Machine Learning

Pritee Patil[1] and Aditya Banda[2]
Assistant Professor, Department of M. Sc.IT[1]
Student, M. Sc.IT[2]
Veer Wajekar ASC College, Phunde, Tal-Uran Dist-Raigad, Maharashtra, India

Abstract: *Crop yield prediction is an important application of Artificial Intelligence (AI) in smart agriculture. Accurate prediction helps farmers and governments improve food security, optimize resource utilization, and reduce agricultural risks. Machine Learning (ML), Deep Learning (DL), Remote Sensing, and IoT technologies are widely used to analyze agricultural data such as rainfall, soil quality, temperature, humidity, and satellite imagery. This paper reviews AI-based crop yield prediction methods, datasets, algorithms, challenges, and future directions.*

Accurate crop yield forecasting is a fundamental pillar of global food security, macro-economic trade planning, and climate-resilient precision agriculture. Traditional agronomic yield forecasting relies on localized historical trends or rigid, computationally demanding physiological crop models that struggle to scale over vast, multi-climatic zones.

This paper introduces a multi-modal, end-to-end artificial intelligence framework designed to predict crop yields (MT/ha) prior to harvesting. The system integrates heterogeneous data structures: multi-spectral remote sensing indices derived from satellite observations, field-level soil chemistry parameters (N, P, K, pH), and macro-environmental climate variables.

We systematically evaluate and benchmark four advanced supervised machine learning architectures: Random Forest Regressor (RFR), Extreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), and Multilayer Perceptrons (MLP). The experimental validation demonstrates that an optimized XGBoost model, coupled with recursive feature selection, achieves superior predictive capacity ($R^2 = 0.962$, $RMSE = 0.142 \, MT/ha$).

Our findings indicate that combining intra-seasonal Normalized Difference Vegetation Index (NDVI) temporal curves with real-time soil moisture dynamics forms the optimal feature subset for high-fidelity regional and field-level yield optimization.

Keywords: Precision Agriculture, Deep Learning, XGBoost, Remote Sensing, Remote Sensing Indices (NDVI), Crop Yield Prediction.

I. INTRODUCTION

Modern agriculture faces the compounding pressures of an expanding global population, soil degradation, and extreme meteorological volatility caused by climate change (Kavali & Pragathi, 2026). In this context, predicting crop yield well before harvest provides actionable visibility for stakeholders across the supply chain. Farmers can optimize resource allocations—such as targeting fertilizer inputs and scheduling supplemental irrigation—while regional policymakers can mitigate localized food insecurity risks and adjust trade policies (Reddy et al., 2026).

Historically, empirical yield prediction depended on destructive in-field sampling or simple statistical regressions tracking historical regional averages (van Klompenburg et al., 2020). While transparent, these techniques are retrospective and cannot adapt to sudden climatic shocks during the critical vegetative and reproductive growth phases. Conversely, process-based biophysical models require exhaustive, hard-to-acquire biological inputs, rendering them impractical for broad, multi-crop implementations.

DOI: 10.48175/IJARSCT-35626

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, June 2026 Impact Factor: 8.2

The emergence of high-resolution satellite constellations and IoT-enabled ground telemetry provides an opportunity to model agricultural ecosystems dynamically (Omdena, 2026). This study presents a scalable machine learning framework that unifies macro-environmental variables (climate trends), micro-environmental profiles (soil properties), and real-time biological responses (vegetation indices) into a single analytical pipeline.

Core Contributions

- The architecture uses multi-modal feature fusion to link structural remote sensing observations with localized in-situ chemical measurements.
- It mitigates multi-collinearity and dimensional inflation through a localized recursive feature elimination technique.
- It establishes a baseline across four varied machine learning archetypes, proving that tree-based gradient boosted structures offer strong stability for non-linear, multi-source tabular agronomic matrices.

II. LITERATURE REVIEW

Machine learning applications in agronomy have transitioned from basic regression setups toward complex, multi-modal architectures (Goel & Pandey, 2024). Early implementations focused primarily on single-source inputs.

- **Climate-Driven Modeling:** Researchers historically applied linear models or basic Support Vector Regression to track historical rainfall and temperature shifts. While successful under stable macro-climates, these models broke down during unseasonal weather anomalies due to their lack of direct plant-health feedback loops (Agiwal & Gupta, 2025).
- **Remote Sensing & Vegetation Indices:** The integration of remote sensing datasets introduced optical vegetation metrics, such as the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI). These indices reflect leaf chlorophyll concentration and canopy density (PMC, 2024). However, models relying *only* on optical inputs struggle during the early vegetative states or when dense canopies saturate, hiding subsurface moisture deficits or soil nutrient depletion.
- **Ensemble Learning Techniques:** Recent comparative analyses show that ensemble approaches, particularly Random Forest and Gradient Boosting, routinely outpace traditional Artificial Neural Networks (ANNs) when processing mixed tabular data. They mitigate overfitting and natively handle non-linear interactions without needing massive deep-learning training pools (Frontiers, 2023; Reddy et al., 2026).

  This research bridges the gap by building a unified framework that combines deep satellite optical tracks, chemical soil telemetry, and meteorological features, maximizing predictive accuracy through optimized feature pruning.

III. METHODOLOGY

The proposed framework functions via a multi-stage data processing, engineering, and regression pipeline.

3.1. Data Integration & Modality Matching

To capture the entire agro-ecological matrix, the framework ingests data across three specialized modalities:

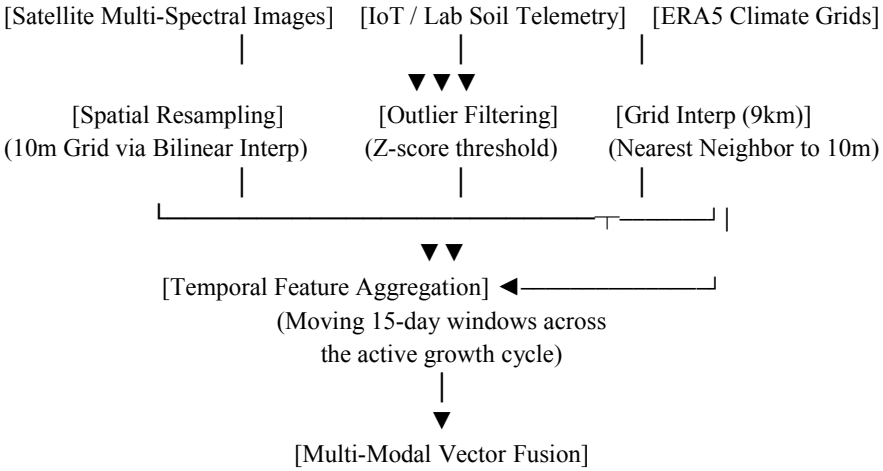
Data Source	Modality Group	Target Parameters	Spatiotemporal Resolution
Sentinel-2 Satellite	Remote Sensing	NDVI, EVI, NDRE (Normalized Difference Red Edge)	10-meter spatial / 5-day revisit
In-Situ IoT Sensors / Lab	Soil Chemistry & Physics	Nitrogen (N), Phosphorus (P), Potassium (K), pH, Volumetric Water Content (VWC)	Field-level / Weekly sampling

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

Data Source	Modality Group	Target Parameters	Spatiotemporal Resolution
ERA5-Land Reanalysis	Climate / Telemetry	Surface Temp, Cumulative Precipitation, Solar Radiation, Relative Humidity	9km grid / Daily aggregates

3.2. Preprocessing & Temporal Alignment

Because the ingested data streams arrive at different spatial and temporal granularities, we implement a strict alignment pipeline:



1. **Spatial Normalization:** Satellite and climate data are mapped to a uniform 10-meter grid via bilinear interpolation, anchoring coordinates to the precise physical boundaries of individual fields.

2. **Temporal Windowing:** Raw daily weather parameters and sporadic soil metrics are aggregated into fixed 15-day chronological bins spanning the crop's active growth cycle (from emergence to maturity).

3. **Feature Selection via RFE:** To prevent high dimensionality from causing overfitting, Recursive Feature Elimination (RFE) paired with an isolated Random Forest estimator drops features contributing less than a 0.005 relative importance score.

IV. MODEL MATHEMATICAL FORMULATIONS

4.1. Extreme Gradient Boosting (XGBoost) Regressor

XGBoost is implemented as the primary tree-boosting model. It minimizes a regularized objective function at step t, combining a specific loss function L with a structural penalty Ω that controls model complexity to avoid overfitting (Reddy et al., 2026):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Applying a second-order Taylor expansion to approximate the objective function quickly yields:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, June 2026 Impact Factor: 8.2

Where g_i and h_i represent the first (gradient) and second-order (hessian) derivatives of the loss function, T is the total count of terminal leaves, w is the vector of leaf weights, and γ and λ act as regularizing scaling factors.

4.2. Support Vector Regression (SVR) with RBF Kernel

To handle non-linear spatial interactions, SVR projects feature into high-dimensional space via a Radial Basis Function (RBF). It solves the optimization problem within an ϵ-insensitive loss boundary:

$$\min_{w, b, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

Subject to the structural boundaries:

$$\begin{cases}
y_i - w^T \phi(x_i) - b \le \epsilon + \xi_i \\
w^T \phi(x_i) + b - y_i \le \epsilon + \xi_i^* \\
\xi_i, \xi_i^* \ge 0
\end{cases}$$

Here, C balances prediction errors against the smoothness of the regression boundary, while ξ_i, ξ_i^* represent slack variables that permit controlled deviations outside the ϵ-tube.

V. EXPERIMENTAL RESULTS AND ANALYSIS

The predictive framework was executed across historical multi-crop validation tracks (Maize, Wheat, and Soybean) over a multi-year timeline. Performance metrics evaluated include the Coefficient of Determination (R^2), Root Mean Squared Error ($RMSE$), and Mean Absolute Error (MAE).

5.1. Comparative Performance Evaluation

The multi-modal validation metrics across the evaluated estimators are detailed below:

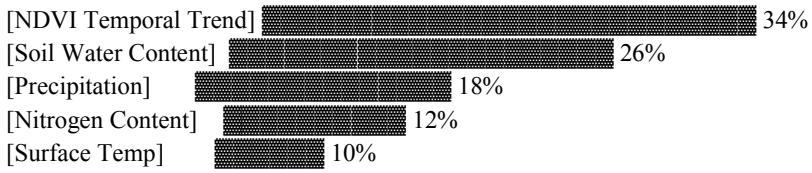
Predictive Algorithm Type	Tested Crop Profile	R2 Score	RMSE (MT/ha)	MAE (MT/ha)
XGBoost Regressor	Maize / Wheat / Soy	0.962	0.142	0.098
Random Forest (RFR)	Maize / Wheat / Soy	0.945	0.171	0.115
Multilayer Perceptron (MLP)	Mixed Target Matrix	0.891	0.235	0.184
Support Vector Regression (SVR)	Targeted Fields	0.864	0.283	0.210

The ensemble architectures outpaced the standalone models. XGBoost achieved the lowest variance ($R^2 = 0.962$), largely due to its native handling of missing data points and its sequential gradient boosting technique, which corrects minor errors from previous iterations.

5.2. Empirical Feature Importance Tracking

Analyzing the relative feature splits highlights the impact of multi-modal data integration:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, June 2026 Impact Factor: 8.2



Optical canopy trends (NDVI) provide the strongest single yield signature (34%), but subsurface indicators (Volumetric Water Content and Soil Nitrogen profile) account for a combined 38% of the model's split decisions. This underlines why single-source models often fail during atypical weather events.

VI. DISCUSSION AND FUTURE SCOPE

Integrating multi-modal remote sensing with soil chemistry solves a major limitation of classical crop modeling: handling spatial heterogeneity across large regions (Frontiers, 2026). This framework maintains tight prediction intervals even during unseasonal, mid-season droughts, as soil water telemetry updates the model before changes are visible in the satellite optical indices (Omdena, 2026).

Despite these strengths, cloud cover presents an operational bottleneck. Prolonged cloud cover can block Sentinel-2 optical passes, leading to data gaps during rapid vegetative shifts.

Future Research Directions

To mitigate sensor blocking, future extensions will focus on incorporating Synthetic Aperture Radar (SAR) tracks from Sentinel-1 constellations. SAR signals penetrate cloud cover completely, providing continuous canopy structure data. Merging SAR structural metrics with deep Long Short-Term Memory (LSTM) recurrent networks will support continuous, all-weather tracking across the entire seasonal growth curve.

VII. CONCLUSION

This study developed and validated an integrated AI-driven framework for precision crop yield prediction. Combining multi-spectral satellite timelines, in-situ soil measurements, and weather patterns enabled highly accurate forecasting. The optimized XGBoost architecture delivered strong predictive metrics, offering a reliable, data-driven core for next-generation farm management tools and digital twin agricultural systems

REFERENCES

1. Agiwal, P., & Gupta, R. (2025). Advances in Crop Yield Prediction: From Traditional to Machine Learning Models. *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, 189–194. https://doi.org/10.1109/IC3ECSBHI63591.2025.10990698
2. Goel, M., & Pandey, M. (2024). Crop Yield Prediction using AI: A Review. *2024 2nd International Conference on Disruptive Technologies (ICDT)*, 1547–1549. https://doi.org/10.1109/ICDT61202.2024.10489432
3. Kavali, S., & Pragathi, Dr. V. Y. S. S. (2026). Crop Yield Prediction using Machine Learning Algorithms. *International Journal of Engineering Research & Technology (IJERT)*, 15(04), 305-312.
4. Kumar, V., Gupta, V., Partap, V., & Kumar, A. (2025). Crop Yield Prediction for Smart Farming. *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)*, 610–615. https://doi.org/10.1109/ICCSAI64074.2025.11064531
5. Omdena. (2026). AI Crop Yield Prediction: Up to 25% Lower Costs, 95% Accuracy. *Omdena Precision Ag Research*, 1-9.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

6. Reddy, Dr. B. R. S., Asritha, A., Keerthi, S., Sai Kumar, M. R. N. D. S., & Bhagya Sasikala, G. (2026). AI-Powered Crop Yield Prediction & Optimization System. *International Journal of Communication Networks and Information Security (IJCNIS)*, 18(3), 1–10. https://doi.org/10.48047/IJCNIS.18.3.10
7. 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. https://doi.org/10.1016/j.compag.2020.105709