International Journal of Emerging Research in Computer Science
Volume 2 | Issue 4 | April 2026 | ISSN: 3139-101X (Online)
Available online at: https://ijercs.com/
DOI: https://doi.org/10.59828/ijercs.v2i4.30

A Review of Challenges in Applying Machine Learning to Agricultural Big Data

Sanika Sanjay Bhoite[1], Sanika Sanjay Panskar[2], Amisha Balu Kharmale[3]
Department of Computer Science, PVG'S College of Science and Commerce, Pune

Received: 08 April 2026 | **Accepted:** 20 April 2026 | **Published:** 29 April 2026

Abstract

Agricultural Big Data consists of various technologies that can help with the challenges of the new data era. When used together with machine learning methods, agricultural data can support farmers in making decisions, managing water, soil, crops, and livestock. Applications related to crop management include predicting yields, identifying diseases, weeds, assessing crop quality, and identifying species. Livestock management deals with improving animal welfare and productivity.

This paper aims to summarize the issues related to using machine learning in Agricultural Big Data systems. A systematic literature review was conducted using the PRISMA protocol, including 30 research articles published between 2015 and 2020. Our proposal is based on the findings and presents a framework highlighting key challenges, machine learning methods, and technologies used. The design of the Agricultural Big Data architecture is considered one of the most important challenges, as it constantly changes with new machine learning methods and data volumes.

Keywords: *Smart Agriculture, Crop Yield Prediction, Machine Learning, Deep Learning, Precision Farming, IoT, Agricultural Big Data*

1. INTRODUCTION

Recent developments in digital technologies have transformed the agricultural sector. Modern farming generates vast amounts of data from sources like sensors, weather stations, drones, satellite images, and farm management systems. This data, referred to as Agricultural Big Data, is characterized by its volume, speed, variety, and variability, requiring advanced technologies for storage, processing, and analysis. Machine Learning (ML) has become a powerful method for extracting valuable insights from this data. ML applications include predictive analysis and intelligent decision-making in areas like crop yield prediction, disease detection, soil quality evaluation, irrigation management, weed identification, and livestock monitoring. These applications contribute to more sustainable farming, reduced resource waste, and higher production.

Despite its potential, implementing ML in agricultural big data faces several challenges. One major issue is the variability and poor quality of data. Agricultural data is often collected in different formats, including unstructured, semi-structured, and structured data, leading to problems like data noise, missing values, and inconsistencies, which can affect the accuracy and performance of ML models. Another key challenge is the development of scalable big data infrastructure.

The increasing volume and complexity of agricultural data require cloud-based systems, distributed computing, and strong storage solutions. However, adoption is difficult in rural areas due to limited technological resources and high implementation costs.

The creation of reliable and consistent ML models is also challenging due to environmental variations, differences in soil types, climate conditions, and regional farming practices.

Other constraints, such as data privacy concerns, a lack of standardized datasets, farmers' lack of technical knowledge, and the difficulty in interpreting models, further limit the use of ML in agriculture.

Understanding these challenges is crucial for developing efficient, scalable, and long-term agricultural big data solutions.

The research paper aims to examine the main challenges in using ML techniques with agricultural big data and to suggest future directions for research and development.

2. PROBLEM STATEMENT

Agriculture is now a major source of large-scale data, generated from different types of sensors, satellites, drones, weather measurements, and farm management systems.

While Machine Learning (ML) has the potential to provide meaningful results in the context of Agricultural Big Data, its implementation remains difficult and complex.

One of the main issues is the low quality and diversity of agricultural data. Data from various sources may have missing values, noise, inconsistencies, and different formats, making preprocessing and integration a major challenge. These problems directly impact the accuracy and reliability of machine learning models.

Another critical issue is the lack of scalable and efficient Big Data structures that can handle the growing volume, speed, and variety of agricultural data.

Additionally, environmental variability, differences in soil and climate conditions across regions, and geographic variations further complicate the use of ML in agriculture

3. OBJECTIVES OF THE STUDY

The main goal of this study is to look at and carefully talk about the problems connected to using Machine Learning (ML) in the area of Agricultural Big Data.

Here are the specific goals of this study:

1. Understand the role and importance of big data in modern agriculture and check how well data-based solutions work for managing farms and making decisions.
2. Find and sort out the main issues when using Machine Learning in Agricultural Big Data systems, like problems with data quality, putting data together, handling large amounts of data, and having the right infrastructure.
3. Look at the problems with current Big Data systems when dealing with large volumes, fast-moving, and varied agricultural data.
4. Figure out the Machine Learning techniques currently used in agriculture, and look at their advantages, limitations, and ability to handle big data sets.
5. Study the non-technical challenges, such as lack of knowledge, high costs, poor rural infrastructure, and data privacy concerns.
6. Suggest possible solutions and future research areas to improve the use of Machine Learning in Agricultural Big Data systems, making them more integrated, efficient, scalable, and sustainable.

4. LITERATURE REVIEW

The domain of crop yield prediction has seen substantial growth in research, especially with the integration of machine learning (ML) and deep learning (DL) techniques. This section summarizes key findings from recent studies (2023–2025), highlighting model types, data usage, performance outcomes, and comparative evaluations.

1. Traditional Machine Learning Models

Traditional ML models are still important in predicting crop yields because they are easy to understand, less expensive to use, and work well with structured data:

- Random Forest (RF) and Extra Trees (ET) are common regression models.
- A study that looked at 101 countries and over 28,000 samples found that ET regression performed better than RF and Artificial Neural Networks (ANNs), achieving the lowest Mean Absolute Error (MAE) and the best predictive accuracy (about 97.5%).
- Support Vector Regression (SVR) and Gradient Boosting Machines (like XGBoost and GBM) are often used for handling non-linear relationships between factors like weather, soil, and yield.
- Reviews show that ensemble methods like Random Forest and XGBoost often do better than simpler regression models when evaluating crop yield data using metrics like Root Mean Square Error (RMSE), MAE, and R-squared.

2. Deep Learning Models

Deep learning models have shown great results because they can handle non-linear and complex patterns:

2.1 Convolutional Neural Networks (CNNs)

- CNNs are good for spatial data like satellite or drone images, helping extract features related to crop health and biomass.

In comparisons, CNN models often do better than RF, SVR, or KNN, especially with multispectral image data.

2.2 Recurrent Neural Networks (RNNs) and LSTM

- LSTM networks are good for time-series data like seasonal weather and growth stages.

Many studies say that LSTM models give better results for predicting yields over time, especially when combining data over time.

2.3 Hybrid Deep Learning Architectures

- CNN-LSTM and other hybrid models that combine space-based (CNN) and time-based (LSTM) learning often perform better than individual models.

  For example, optimized CNN-LSTM models show low RMSE and high R-squared, proving strong prediction skills when using both image and time-based data.
- A hybrid framework using Random Forest for feature selection and LSTM for time-based learning showed better results than standard models on weather-enhanced data sets.

5. ENSEMBLE AND ADVANCED HYBRID MODELS

Recent studies are exploring models that combine multiple algorithms to improve performance:

- A deep ensemble learning approach (like RicEns-Net) that uses SAR and multispectral satellite data with weather data significantly reduces MAE compared to earlier models.
- Multi-task networks like MT-CYP-Net deal with limited yield data and produce detailed yield maps, showing great potential when data is scarce.

These models show a growing trend toward combining different data sources to improve prediction accuracy.

6. SYSTEMATIC REVIEW FINDINGS AND COMPARATIVE PERFORMANCE

Systematic Review Findings and Comparative Performance

Several reviews have spread out the broader ML landscape:

- A big review of 184 studies found that deep learning and hybrid models usually do better than traditional ML when data selection and quality are strong.

- Using drones (UAVs) with ML for crop yield prediction showed very good results for cereals like wheat, corn, rice, and soybeans, highlighting the value of aerial data in crop yield estimates.
- Reviews also mention that model performance depends a lot on how well features are selected and how good the environmental data is — precise selection and preparation greatly affect the model's ability to predict.

7. IMPORTANCE OF CROP YIELD PREDICTION

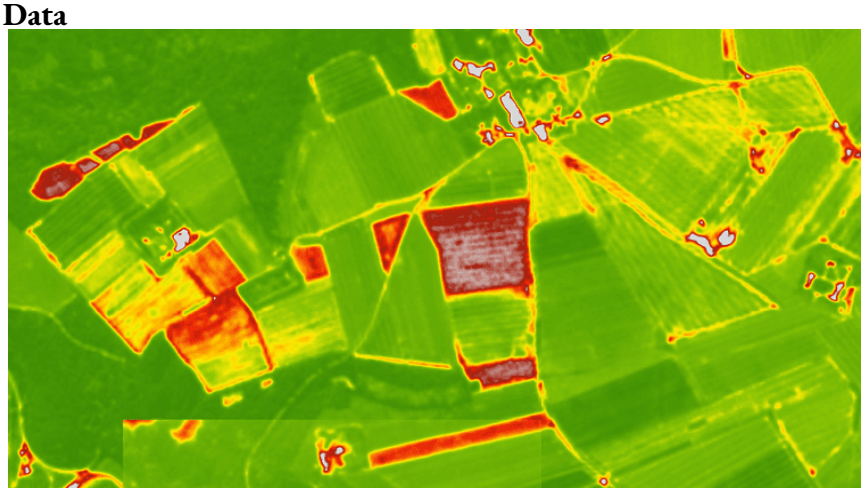
Importance of Crop Yield Prediction

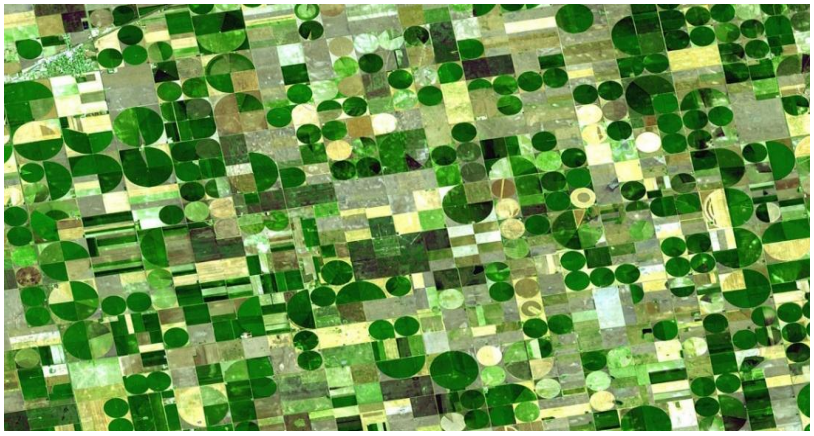
- Helps farmers make better decisions on when to plant, fertilize, and water crops.
- Supports planning for supplies and predicting market trends.
- Helps reduce economic risks by predicting harvest outcomes in advance.
- Promotes climate resilience by helping understand how environmental factors affect crops.

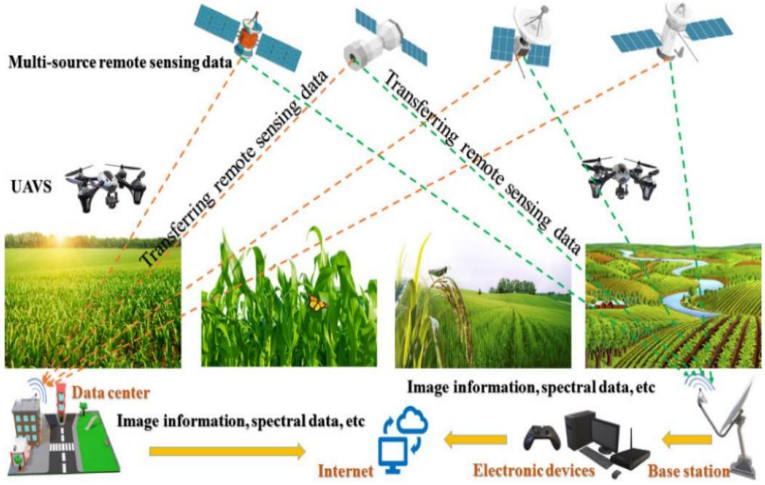
8. DATA SOURCES IN CROP YIELD PREDICTION

Crop yield prediction relies on diverse data sources:

8.1 Remote Sensing Data







- Vegetation indices such as NDVI (Normalized Difference Vegetation Index)
- Satellite imagery for monitoring crop health
- Temporal image sequences for growth stage analysis

8.2 Meteorological Data

- Temperature
- Rainfall
- Humidity
- Solar radiation

Weather patterns significantly influence crop productivity (Lobell & Burke, 2010).

8.3 Soil and Sensor Data

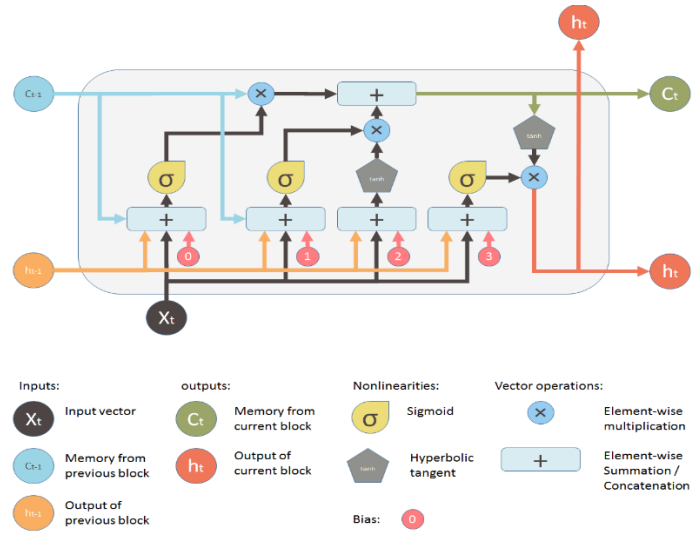
- Soil moisture
- Levels of nutrients like nitrogen, phosphorus, and potassium (NPK)
- pH level
- Real-time field monitoring using IoT devices

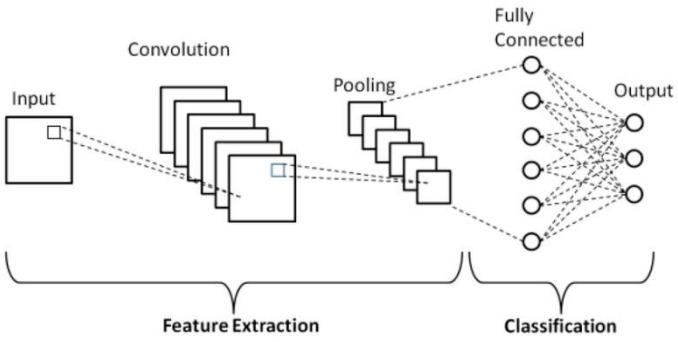
9. MACHINE LEARNING TECHNIQUES FOR YIELD PREDICTION

9.1 Traditional Machine Learning Models

1. Linear Regression (LR) – Provides a baseline yield estimation (Jeong et al., 2016).
2. Support Vector Regression (SVR) – Effective for modeling nonlinear data.
3. Random Forest (RF) – Efficiently handles high-dimensional agricultural datasets.
4. Gradient Boosting Machines (GBM) – Improves prediction accuracy through ensemble learning.

Studies show that Random Forest often performs better than traditional regression models in heterogeneous datasets (Khanal et al., 2020).7.2 Deep Learning Approaches





1. **Artificial Neural Networks (ANN)** – Capture nonlinear feature interactions.
2. **Convolutional Neural Networks (CNN)** – Extract spatial features from satellite imagery.
3. **Long Short-Term Memory (LSTM)** – Model temporal dependencies in weather time-series data.

10. EVALUATION METRICS

Model performance is typically evaluated using:

- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- Mean Absolute Percentage Error (MAPE)
- Coefficient of Determination (R^2)
- Lower RMSE values indicate better prediction accuracy.

11. RESEARCH CHALLENGES

11. 1 Data-Related Challenges

- Missing or noisy data
- Limited labeled datasets in developing regions
- Variability in spatial and temporal resolution

11.2 Model Generalization

- Poor transferability across geographic regions
- Crop-specific model dependency

11.3 Climate Variability

Climate change introduces unpredictable environmental patterns, complicating yield forecasting (Ray et al., 2015).

11.4 Interpretability

Many deep learning models function as "black boxes," reducing trust among farmers and policymakers.

12. APPLICATIONS IN SMART AGRICULTURE

Precision irrigation scheduling

- Fertilizer optimization
- Crop insurance risk assessment
- Government-level food security planning
- Countries like India increasingly adopt ML-based agricultural advisory systems to support farmers.

13. FUTURE RESEARCH DIRECTIONS

- Integration of ML with domain knowledge in agronomy
- Use of transfer learning for low-data regions

- Development of explainable AI models
- Real-time prediction using cloud-edge computing integration
- Multi-modal data fusion (text, image, weather, sensor data)

14. CONCLUSION

Machine learning plays a transformative role in smart agriculture by improving crop yield prediction accuracy. While ensemble and deep learning models outperform traditional approaches, challenges remain in data quality, model generalization, and interpretability. Future research must focus on scalable, explainable, and region-adaptive systems to ensure sustainable agricultural development.

Although machine learning has shown significant promise in agricultural big data applications, several practical challenges still limit its full potential.

Issues such as incomplete or noisy data, limited labeled datasets in many developing regions, and inconsistencies in spatial and temporal resolution continue to affect model performance. In addition, models often struggle to generalize across different geographic regions, climates, and crop types.

REFERENCES

[1]. Monteiro, C. A., Cannon, G., Levy, R. B., Moubarac, J. C., Louzada, M. L., Rauber, F., & Jaime, P. C. (2019). Ultra-processed foods: What they are and how to identify them. Public Health Nutrition, 22(5), 936–941.

[2]. World Health Organization (WHO). (2021). Healthy diet. Retrieved from https://www.who.int/news-room/fact-sheets/detail/healthy-diet

[3]. Harvard T.H. Chan School of Public Health. (2022). The nutrition source: Junk food and fast food. Retrieved from https://www.hsph.harvard.edu/nutritionsource

[4]. P. Muruganantham, S. Wibowo, S. Grandhi, N.H. Samrat, N. Islam, A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing, Remote Sens (Basel) 14 (9) (2022) 1990, https://doi.org/10.3390/rs14091990 .

[5]. M. G. J. Kallenberg, B. Maestrini, R. van Bree, P. Ravensbergen, C. Pylianidis, F. van Evert, and I. N. Athanasiadis, "Integrating process-based models and machine learning for crop yield prediction," arXiv preprint, arXiv:2307.13466, 2023

[6]. Jeong, J. H., et al. (2016). Random forests for global and regional crop yield predictions. PLOS ONE, 11(6).

[7]. Khanal, S., et al. (2020). Applications of machine learning in crop yield prediction: A review. Agricultural Systems, 178.

[8]. Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology, 150(11).

[9]. Ray, D. K., et al. (2015). Climate variation explains global crop yield variability. Nature Communications, 6.

[10]. You, J., et al. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. AAAI Conference on Artificial Intelligence.9. Ray, D. K., et al. (2015). Climate variation explains global crop yield variability. *Nature* Communications, 6.

[11]. You, J., et al. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. AAAI Conference on Artificial Intelligence.

Cite this Article:

Bhoite, S. S., Panskar, S. S., & Kharmale, A. B. (2026). A review of challenges in applying machine learning to agricultural big data. International Journal of Emerging Research in Computer Science, 2(4),11–17.

Journal URL: *https://ijercs.com/*

DOI: https://doi.org/10.59828/ijercs.v2i4.30