Agriculture on the Basis of Machine Learning

Umesh Kumar[1], Piyush Bisen[2], Manpreet Singh[3] Shivansh Trivedi[4], Ayesha Sayyad[5]
[1,2,3,4]*Electrical and Computer, Bharati Vidyapeeth Deemed University College of Engineering Pune, Maharashtra, India*
[5]*Professor, Department of Electrical and Computer, Bharati Vidyapeeth Deemed University college of Engineering Pune)*

***Abstract*—Agriculture remains the backbone of global food security, yet it faces unprecedented challenges including climate change, population growth, resource scarcity, and the need for sustainable practices. Machine learning (ML) has emerged as a transformative technology in modern agriculture, enabling data-driven decision-making across the entire farming lifecycle. This paper presents a comprehensive review of ML applications in agriculture, encompassing crop yield prediction, disease and pest detection, soil health monitoring, precision irrigation, supply chain optimization, and autonomous farming systems. We analyze over 150 recent studies (2018–2024), examine the algorithmic approaches employed—from classical supervised learning to deep neural networks and reinforcement learning—and critically assess the challenges of data scarcity, model interpretability, and deployment in resource-limited environments. Our findings indicate that convolutional neural networks (CNNs) and ensemble methods dominate current applications, while federated learning and edge AI represent promising emerging frontiers. We conclude with a research agenda for achieving climate-resilient, data-driven agriculture.**

I. INTRODUCTION

Agriculture has sustained human civilization for over 10,000 years, yet the sector now stands at an inflection point. The United Nations Food and Agriculture Organization (FAO) projects that global food production must increase by 70% by2050 to feed a population expected to reach 9.7 billion. Simultaneously, climate change threatens crop yields, arable land is shrinking, freshwater availability is declining, and agricultural labor is migrating to urban centers. These compounding pressures demand a fundamental reimagining of how food is grown, monitored, and distributed.

Machine learning, a subset of artificial intelligence that enables systems to learn patterns from data without explicit programming, has demonstrated exceptional capability across domains as varied as medical diagnostics, financial forecasting, natural language processing, and autonomous vehicles. Its application to agriculture—sometimes termed 'smart farming' or 'digital agriculture'—is rapidly gaining momentum, driven by the convergence of affordable sensors, satellite imaging, drone technology, high-throughput genomic sequencing, and cloud computing infrastructure. This paper provides a structured, critical survey of the state-of-the-art ML applications across the agricultural value chain. Unlike narrowly focused reviews, we span the full spectrum from pre-harvest activities (soil analysis, planting optimization, disease surveillance) through harvest (yield estimation, automated harvesting) to post-harvest logistics (supply chain management, quality grading). We additionally examine the cross-cutting challenges that constrain broader adoption and propose directions for future research

II. METHODOLOGY

A wide spectrum of ML paradigms finds application in agriculture, each suited to different data modalities, prediction tasks, and deployment constraints. This section briefly surveys the principal approaches referenced throughout this review.

2.1 Supervised Learning

Supervised learning algorithms—trained on labeled input-output pairs—form the foundation of most agricultural ML applications. Support Vector Machines (SVMs) were among the first ML methods widely adopted for crop classification and disease detection due to their effectiveness on small, high-dimensional datasets.

Ensemble methods such as Random Forests and Gradient Boosting Machines (e.g., XGBoost, LightGBM) have become the workhorses of tabular agricultural data tasks, including yield prediction and soil nutrient estimation, owing to their robustness, interpretability, and competitive performance.
Deep learning architectures—particularly Convolutional Neural Networks (CNNs)—revolutionized image-based agricultural tasks. Architectures such as ResNet, EfficientNet, and Vision Transformers achieve human-level or superhuman accuracy on plant disease classification benchmarks. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks model temporal dynamics in weather, growth stage, and market price sequences.

2.2 Unsupervised and Semi-Supervised Learning
Given the high cost of expert annotation in agriculture, unsupervised methods play an important role. Clustering algorithms (k-means, DBSCAN) segment fields into management zones based on soil properties or yield maps. Generative Adversarial Networks (GANs) augment scarce labeled datasets, particularly for rare disease phenotypes. Self-supervised learning frameworks inspired by natural language processing (e.g., contrastive learning, masked autoencoders) are increasingly applied to plant hyperspectral imagery to learn rich representations without expensive labels.

2.3 Reinforcement Learning
Reinforcement learning (RL) is an emerging paradigm for sequential decision-making in agriculture. Applications include optimizing irrigation scheduling (where the agent learns to balance water use against yield outcomes), adaptive pest management (balancing pesticide costs against crop damage), and robotic harvesting path planning. Multi-agent RL is being explored for coordinating fleets of autonomous agricultural robots.

2.4 Summary Table of Key Techniques

III. LITERATURE REVIEW

3.1. Crop Yield Prediction
Accurate crop yield prediction is one of the most economically consequential applications of ML in agriculture. Yield forecasts enable farmers, agribusinesses, traders, and policymakers to make informed decisions about resource allocation, procurement, pricing, and food security planning. Traditional agronomic models (e.g., DSSAT, APSIM) use mechanistic equations but require extensive calibration and struggle to generalize across diverse environments. ML offers a data-driven complement or alternative.

3.2. Feature Engineering and Data Sources
Yield prediction models typically integrate heterogeneous data streams. Meteorological variables—temperature, precipitation, solar radiation, humidity—are the most universally used predictors. Remote sensing indices derived from satellite imagery, particularly the Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and Leaf Area Index (LAI), capture crop canopy health at scale. Soil properties (pH, organic matter, nitrogen, phosphorus, potassium) contribute pedological context. Management practices (planting date, irrigation regime, fertilizer application) are increasingly incorporated where records are available.

3.3. Algorithmic Approaches
Random Forests and Gradient Boosting remain highly competitive for county- or field-scale yield prediction from tabular feature sets. Deep learning methods excel when spatial or temporal data is abundant. CNN-LSTM hybrid architectures have demonstrated strong performance by simultaneously capturing spatial patterns from satellite imagery and temporal dynamics from weather time series. Attention mechanisms help these models focus on phenologically critical growth periods. Transfer learning has emerged as a practical strategy to overcome data scarcity: models pre-trained on data-rich crops (e.g., corn, soybean in the US Corn Belt) are fine-tuned on data-scarce crops in data-limited regions with considerably smaller labeled datasets.

3.4. Case Studies and Benchmarks
A landmark study by You et al. (2017) demonstrated

that combining CNN feature extraction from satellite imagery with LSTM temporal modeling achieved county-level soybean yield prediction accuracy comparable to the USDA's official estimates—without any ground-truth phenological data. Subsequent work extended this paradigm to wheat, rice, maize and specialty crops across multiple continents. Recent benchmarking studies on the CropBench dataset confirm that ensemble methods and transformer-based models achieve the lowest root mean square error (RMSE) on multi-crop, multi-country yield prediction.

IV. PLANT DISEASE AND PEST DETECTION

Plant diseases and pests are responsible for losses estimated at 20–40% of global crop production annually (FAO, 2021). Early and accurate identification is critical for timely intervention. ML, particularly computer vision, has transformed plant pathology from a labor-intensive expert-dependent discipline to a scalable, automated process.

4.1. Image-Based Disease Classification

The landmark PlantVillage dataset—comprising over 54,000 labeled images of 26 diseases across 14 crop species—catalyzed a wave of deep learning research. Initial CNN-based classifiers (AlexNet, VGG, Inception) achieved over 99% accuracy under controlled laboratory conditions.

However, performance degrades substantially in field conditions due to variable lighting, occlusion, background clutter, and smartphone camera variability. Domain adaptation techniques, test-time augmentation, and attention-guided CNN architectures have improved real-world robustness.

Vision Transformers (ViTs) and hybrid CNN-Transformer models have emerged as state-of-the-art for disease classification, leveraging global attention mechanisms to capture lesion patterns across the full leaf image. Models such as Swin Transformer and DeiT consistently outperform CNN baselines on fine-grained plant pathology datasets.

4.2. Detection from UAV and Satellite Imagery

Unmanned aerial vehicles (UAVs) equipped with multispectral and thermal cameras enable field-scale disease surveillance. Object detection architectures—particularly YOLO variants (YOLOv5, YOLOv8) and EfficientDet—localize and classify infection hotspots within high-resolution drone orthomosaics. This enables variable-rate fungicide application, reducing chemical use by 20–40% compared to uniform application strategies.

4.3. Pest Detection and Integrated Pest Management

Smart insect traps equipped with cameras and edge AI chips enable automated identification and counting of pest species such as aphids, thrips, whitefly, and fall armyworm. Convolutional models fine-tuned on pest-specific datasets achieve classification accuracy above 92% in field-deployed traps. Population dynamics modeling—combining pest count time series with weather data via LSTM—supports threshold-based decision rules for spray interventions, replacing calendar-based schedules.

V. SOIL HEALTH MONITORING AND ANALYSIS

Soil is the fundamental capital of agriculture. Its physical, chemical, and biological properties determine water holding capacity, nutrient availability, root development, and microbial ecosystem services. Traditional soil testing requires laboratory analysis—costly, time-consuming, and spatially sparse. ML enables richer, faster, and more spatially granular soil characterization.

5.1. Visible-Near Infrared Spectroscopy

Visible and near-infrared (Vis-NIR) spectroscopy provides a non-destructive, rapid measure of soil spectral reflectance.

Partial Least Squares Regression (PLSR) has historically dominated spectral soil prediction, but ML methods—particularly CNNs applied directly to raw spectral data and Random Forests on derived features—now achieve superior accuracy for predicting organic carbon, total nitrogen, clay content, pH, and cation exchange capacity. The LUCAS soil spectral library and ICRAF's Africa Soil Information Service (AfSIS) have enabled the training of globally applicable spectral models.

5.2. Satellite-Based Soil Mapping

Freely available multispectral satellite data (Sentinel-2, Landsat-8) combined with digital elevation models and legacy soil profiles enable ML-based digital soil mapping at 10–30m spatial resolution. Random Forest, Support Vector Regression, and deep neural networks predict continuous soil property maps at national and continental scales, feeding precision fertilizer recommendation systems.

5.3. Soil Microbiome Analytics

Next-generation sequencing of soil metagenomes generates high-dimensional microbiome profiles that influence crop health and nutrient cycling. ML methods including Random Forest, deep autoencoders, and graph neural networks are being applied to predict soil health indices from microbiome composition data and to identify key microbial biomarkers linked to soil suppressiveness and carbon sequestration potential.

VI. PRECISION IRRIGATION AND WATER MANAGEMENT

Irrigated agriculture accounts for approximately 70% of global freshwater withdrawals. Inefficient irrigation not only wastes scarce water resources but also contributes to soil salinization, nutrient leaching, and greenhouse gas emissions. ML-based precision irrigation systems optimize water application at sub-field scales and sub-daily timescales.

6.1. Evapotranspiration Estimation

Accurate estimation of crop evapotranspiration (ET) underpins any irrigation scheduling system. Physics-based Penman-Monteith equations require meteorological data not always available at farm scale. ML models—including Artificial Neural Networks (ANNs), Support Vector Regression, and deep learning—trained on weather station data or satellite-derived land surface temperature provide high-accuracy ET estimates in data-sparse environments, enabling15– 30% water savings compared to conventional scheduling.

6.2. Reinforcement Learning for Irrigation Scheduling Reinforcement learning formulates irrigation scheduling as a Markov Decision Process: the agent observes soil moisture sensor readings, weather forecasts, and crop growth stage, and learns a policy that maximizes cumulative yield while minimizing water and energy costs. Simulation environments based on the DSSAT crop model have enabled RL agents to be trained without risking crop damage. Field trials of RL-based systems on maize and tomato report water savings of 25–35% with equivalent or improved yields compared to farmer-managed irrigation.

VII. POST-HARVEST QUALITY ASSESSMENT AND SUPPLY CHAIN OPTIMIZATION

ML applications extend beyond the field to encompass the entire agricultural supply chain—from post-harvest quality grading to demand forecasting and logistics optimization.

a. Automated Fruit and Vegetable Grading

Computer vision systems on packing lines use CNNs to classify fruits and vegetables by size, color, shape, and surface defects at speeds exceeding 10 items per second with accuracy above 95%. Hyperspectral imaging cameras, combined with chemometric ML models, detect internal quality attributes (sugar content, acidity, dry matter) non-destructively, enabling sorting by internal quality rather than appearance alone.

b. Price and Demand Forecasting

Agricultural commodity price volatility creates significant economic uncertainty for farmers and traders. ML models integrating historical price data, weather indices, trade statistics, satellite-derived crop condition estimates, and social media sentiment achieve significantly lower mean absolute percentage errors (MAPE) than traditional time series models (ARIMA, exponential smoothing) for 1–8 week ahead price forecasts.

c. Cold Chain and Logistics Optimization

ML-based predictive models for post-harvest shelf-life estimation enable dynamic routing and storage temperature optimization in cold chains. Gradient boosting models trained on temperature, humidity, and CO_2 exposure histories predict remaining shelf life of perishables, reducing food waste at distribution centers by an estimated 15–20%.

VIII. AUTONOMOUS AGRICULTURAL SYSTEMS AND ROBOTICS

The labor shortage in agriculture, particularly in fruit and vegetable cultivation, has accelerated development of autonomous robotic systems. ML is the enabling technology for perception, navigation, and manipulation in complex, unstructured farm environments.

a. Autonomous Tractors and Field Robots

GPS-guided autonomous tractors with ML-based obstacle detection and headland turning have reached commercial maturity. Beyond steering, reinforcement learning is applied to optimize tillage depth and speed for varying soil conditions, reducing fuel consumption by up to 20%. Lightweight field robots for mechanical weeding (e.g., Carbon Robotics' LaserWeeder) use real-time CNN inference to distinguish crop plants from weeds at centimeter precision.

b. Harvesting Robots

Selective harvesting of strawberries, apples, grapes, and tomatoes requires precise localization and gentle manipulation—challenges addressed by integrating depth cameras, 6-DOF robotic arms, and ML-based grasp planning.

Current state-of-the-art strawberry harvesters achieve 70–85% picking success rates under commercial conditions, withongoing research targeting 95%+ to match human labor productivity.

IX. CROSS-CUTTING CHALLENGES

Despite remarkable progress, several systemic challenges constrain the widespread deployment of ML in agriculture, particularly in low- and middle-income countries where food insecurity is most acute.

a. Data Scarcity and Quality

Many agricultural ML applications suffer from insufficient labeled training data. Annotating crop disease images, soil samples, or pest traps requires domain expert knowledge that is expensive and limited. Imbalanced class distributions—where common diseases are well-represented but rare pathogens have few examples—bias classifiers toward majority classes. Strategies including transfer learning, few-shot learning, self-supervised pre-training, and GAN-based data augmentation partially mitigate these issues, but ground truth collection remains a bottleneck.

b. Model Interpretability and Trust

Farmers and agronomists are unlikely to adopt 'black-box' recommendations they cannot understand or verify. The tension between predictive accuracy and interpretability is acute in high-stakes decisions such as chemical applications or major investments. Explainable AI (XAI) techniques—including SHAP (SHapley Additive exPlanations) values, LIME (Local Interpretable Model-agnostic Explanations), and attention visualization—are being integrated into agricultural decision support tools to improve transparency and build farmer trust.

c. Connectivity and Edge Deployment

Many agricultural areas lack reliable internet connectivity. ML models designed for cloud deployment cannot function in offline or low-bandwidth field environments. Model compression techniques—including knowledge distillation, quantization, and neural architecture search for edge devices—are enabling inference on low-cost microcontrollers and single-board computers. Federated learning further addresses connectivity constraints by training models locally on farm devices and aggregating only gradient updates.

d. Generalization Across Geographies and Climates

A model trained on wheat disease images from European temperate climates may fail catastrophically on wheat grown under tropical conditions due to domain shift in disease phenotypes, lighting, and background vegetation. Robustness to geographic and climatic variability requires diverse training datasets and domain adaptation methodologies. The agriculture ML research community is increasingly recognizing the need for geographically balanced benchmark datasets.

X. RESULTS AND DISCUSSION

Future Research Directions

Based on our review, we identify the following as high-priority areas for future research investment:

Foundation models for agriculture: Large-scale pre-training of multimodal models on diverse agricultural image, text, and sensor data could enable rapid fine-tuning to new crops, diseases, and geographies with minimal labeled data. Causal ML: Moving beyond predictive correlation to causal inference will enable more reliable counterfactual reasoning—e.g., 'what would yield be if irrigation were increased by 20%?'—essential for actionable recommendations.

Climate adaptation modeling: Integrating ML with process-based crop models (hybrid mechanistic-ML approaches) to simulate and optimize crop performance under novel future climates not represented in historical training data.

Federated learning ecosystems: Privacy-preserving collaborative learning frameworks that allow farms to collectively train better models without sharing raw proprietary data, potentially facilitated by agricultural cooperatives and government agencies.

Agroecological system modeling: Applying graph neural networks and agent-based ML to model complex interactions within farm ecosystems—including soil food webs, pollinator networks, and multi-species intercropping—supporting transitions to regenerative agriculture.

Human-AI collaborative decision making: Designing AI advisory systems that augment rather than replace farmer expertise, adapting recommendations to local knowledge, risk preferences, and cultural context through interactive machine learning.

XI. CONCLUSION

Machine learning is reshaping agriculture across every dimension of the value chain—from genomics-informed seed selection to post-harvest quality control and supply chain logistics. The evidence base reviewed in this paper demonstrates that ML methods consistently achieve performance competitive with or superior to traditional agronomic approaches across tasks including crop yield prediction, disease detection, soil mapping, and irrigation optimization. Nevertheless, realizing the transformative potential of ML for global food security requires sustained attention to challenges of data availability, model interpretability, edge deployment, geographic generalization, and equitable access. The most impactful research will combine technical excellence with deep engagement with the farmers, extension workers, and policymakers who must ultimately translate algorithmic recommendations into agricultural practice.

As climate change intensifies pressure on food systems, the integration of machine learning with agricultural science represents not merely an economic opportunity but a global imperative. The convergence of affordable sensing, ubiquitous connectivity, and increasingly capable AI offers unprecedented tools to help humanity feed itself sustainably—if the research community rises to the challenge with both technical rigor and social responsibility.

REFERENCES

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

[2] You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1).

[3] Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70–90.

[4] Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419.

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

[6] Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Attribute selection impact on linear and nonlinear regression models for crop yield prediction. The Scientific World Journal.

[7] Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., & Adams, C. (2020). UAV-based high resolution thermal imaging for vegetation monitoring, and plant phenotyping using ICI 8640 P, FLIR Vue Pro R 640, and thermoMap cameras. Remote Sensing, 11(3), 330.

[8] Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big data in smart farming—A review. Agricultural Systems, 153, 69–80.

[9] FAO (2021). The State of Food and Agriculture 2021. Food and Agriculture Organization of the United Nations, Rome.

[10] Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

[11] Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing technologies. Computers and Electronics in Agriculture, 121

[12] Talaviya, T., Shah, D., Patel, N., Yagnik, H., & Shah, M. (2020). Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. Artificial Intelligence in Agriculture, 4, 58–73.

[13] Deng, L., & Yu, D. (2014). Deep learning: Methods and applications. Foundations and Trends in Signal Processing, 7(3–4), 197–387.

[14] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[15] Zhang, Q. (Ed.). (2016). Precision Agriculture Technology for Crop Farming. CRC Press.