IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ∺ Peer-reviewed & Refereed journal ∺ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15650

Enhancing Crop Prediction Using Artificial Intelligence for Smart Agriculture

Urvashi[1], Naveen Sharma[2]

School of Computational Sciences, GNA University, Phagwara, Punjab – 144401, India[1]

School of Computational Sciences, GNA University, Phagwara, Punjab – 144401, India[2]

ORCID: https://orcid.org/0009-0009-1208-6108[2]

Abstract: The food security situation all over the world has never faced such problems like the population growth, limited use of agricultural resources and climate variability. In response to these challenges, another technology has surfaced—Artificial Intelligence (AI)—which offers potential to tackle the problems by helping to create accurate systems to predict crops based on data.But there is another technology that has come to terms to address these challenges: Artificial Intelligence (AI). The current state of the art on AI application-based crop prediction in smart agriculture is summarized and presented through this review paper by collating information from peer-reviewed publications published from 2015 to 2024. In the context of predicting crop yield, disease incidence, water needs and optimal sowing windows, we analyze the use of machine learning (ML) algorithms, deep learning architectures such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks and Transformer-based models. Theoretically, we developed a uniform taxonomy for the classification of different AI approaches to crop prediction and studied the model understandable versus model interpretability trade-offs. Real-world challenges, including the scarcity of data in smallholder environments and limitations on computing resources, are discussed in detail, along with the important role of extension services providing actionable guidance to farmers based on these AI technologies. In the case study carried out in the agro-rich region of Punjab, India, the hybrid CNN-LSTM models demonstrate accuracy levels of up to 93.2% for wheat and rice yield prediction using multi-modal data, which combines satellite imagery, IoT soil sensor data, meteorological data and government agronomic database data. The paper ends with a comprehensive research roadmap that includes federated learning techniques for data sharing without compromising privacy, explainable AI (XAI) for fostering farmer trust, edge computing for deployment in remote areas, and pipelines for retraining AI systems in the face of climate change. Overall, the study reveals the significant potential of AI in the agricultural sector to transform farming processes and the societal issues and enablers that need to be addressed to achieve equitable adoption.

Keywords: Artificial Intelligence, Crop Prediction,Smart agriculture, Machine learning, Precision farming , Agriculture data analysis.

1. INTRODUCTION

The status of agriculture as the base of human civilization and the main source of food, income and economic stability for billions of human beings around the world remains as it is.The role of agriculture as the foundation of human civilization is still the same with billions of human beings still depending on the crop for their sustenance, living and economic stability. FAO estimates global food production will need to rise by nearly 70% by 2050 to serve a population of some 9.8 billion people. However, the agriculture sector faces many additional difficulties such as climate-induced climate variability, vast soil degradation, severe water shortage, and dwindling farmable land caused by urbanization (IPCC, 2022). All these stresses negatively impact the reliability and predictability of crop outcomes, leading to heavy pressure for farmers, food supply chains, and policy makers. Decisions are made traditionally in agronomic activities mainly based on old heuristics, farmer experience, and primitive statistical forecasting. Such methods are satisfactory in stable climatic settings, but in the present climatic context of agricultural volatility, it is undoubtedly a method that doesn't work well. Historical crop yield forecasting has been based on regression models from which the output data is regressed onto only one or few variables like rainfall or temperature. None of these methods account for the non-linear interaction of soil chemistry, weather, pest pressure and the genetic traits and management of the crop to explain yield results (Liakos et al., 2018). AI, especially the sub fields of machine learning (ML) and deep learning (DL) represents a paradigm shift in agricultural forecasting. It is possible to extract patterns from a mix of heterogeneous, high-dimensional data, such as satellite remote sensing imagery, ground level IoT data, time series data on weather patterns and land use conditions in the socioeconomic domain, which are not visible in the raw data to human analysts or classical statistics features classes (Sharma et al., 2021). This can help develop models with unprecedented temporal and spatial accuracy for growing and yield prediction. The urgency and the opportunity is

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⸾ Peer-reviewed & Refereed journal ⸾ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15650

illustrated in the Indian agriculture. Punjab is the "Granary of India" and is responsible for producing around 17% wheat production in the country and 11% rice production in the country (Government of Punjab, 2023). The agriculture sector of Punjab does suffer structural problems, however, including over-use of groundwater, degrading soil health due to fertilizer intensive farming and the growing heat stress at sensitive crop stages. Prediction systems using AI specifically developed for agro-climatic conditions in Punjab are especially promising for facilitating data-informed inputs. This is a review paper which has four main contributions to the literature:

1. A summary of AI methods used in crop forecasting that is organized into a new taxonomical system ranging from classical ML to the adaptation of large language models (LLM).
2. An essential comparative study of architectures, training datasets and requirements, interpretability, and deployment statefulness.
3. An empirical case study was conducted based on the ground truth data of historic yields from the region of Punjab in India, and the predictions of CNN-LSTM were tested.
4. A full research program to define the most impactful unsolved challenges in smart agriculture with AI. To this end, the remaining sections of the paper explain the literature reviewed as part of this study, quantitative results, the case of Punjab, the theoretical and practical implications, and conclude with a future direction to be used in the study.

2. LITERATURE REVIEW

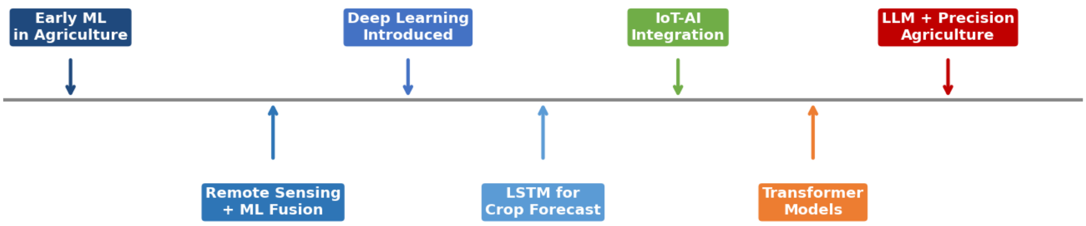

Fig 2 : Evolution of AI Applications in Agricultural Research (2010–2023)

2.1 Early Applications of Machine Learning in Agriculture

It's also important to note that the first applications of machine learning to agriculture dates back to the early 2010s, when researchers took algorithms for pattern recognition, which they applied in other industries such as bioinformatics, finance, or meteorology, and adapted them to agricultural expressions. Mucherino et al. (2009) gave an early analysis of the ways in which data mining techniques could be applied to the field of agriculture, showing that k-nearest neighbour and decision tree learning algorithms were capable of accurately characterizing soil types and estimating irrigation needs in structured tabular data. The small scale of agricultural data and the complexity of the models required computational power limited these early attempts. The heterogeneous in-field sensor data was utilized for predicting the wheat yield zones using supervised machine learning as in performed by Pantazi et al., (2016) where Support Vector Machines (SVM) proved to be better performing than traditional linear discriminant analysis. One thing they stressed was the role of feature engineering, which is the generation of informative input variables; a strong bottleneck much before the advent of deep learning. At the same time, Gonzalez-Sanchez et al. (2014) conducted a thorough benchmarking study, testing 32 ML algorithms on various datasets of different crops and noting that ensemble techniques generally outperform single-learners by significantly better generalization.

2.2 Remote Sensing and Satellite Data Integration

Satellite remote sensing data was one of the most significant developments in the application of AI for crop forecasting. Multispectral satellite imagery-derived vegetation indices—namely the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI)—offer spatially continuous, near real-time estimates of crop health, canopy development and photosynthetic activity. You et al. (2017) presented an ultra scalable deep Gaussian process model that combined MODIS satellite imagery with county-level yield statistics and was shown to improve prediction accuracy when compared to ground-survey-only baselines considerably with the use of satellite-derived features. Bhatt

Impact Factor 8.471 ⁘ Peer-reviewed & Refereed journal ⁘ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15650

and Bhatt (2020) demonstrated the use of Sentinel-2 MS data with Random Forest classification in India to derive cropped area and estimate crop yields of the semiarid districts in Rajasthan at 84.7% accuracy. Cloud cover in the critical monsoon growing season over South Asia is often seen to affect the quality of optical imagery, such that satellite-based data can only be expressed with the help of Synthetic Aperture Radar (SAR) imagery as a complementary or alternative data source.

2.3 Deep Learning Architectures for Crop Prediction

Deep learning took off around 2015, ushering a new wave of agricultural AI. In 2018, Kamilaris and Prenafeta-Boldu published a groundbreaking survey of 40 studies that leveraged deep learning to solve agricultural problems, showing that Convolutional Neural Networks (CNNs) performed well for image-based applications such as disease detection, weed classification, and fruit maturity assessment, while Recurrent Neural Networks (RNNs) and their LSTM variants were more successful for temporal sequence problems like sampling the varying growth rates of crops during a week. The authors of Khaki and Wang (2019) trained and tested a deep neural network using 35 years of maize yield records including the United States, feeding weather time series and agronomic management variables as inputs. They produced a mean absolute error (MAE) of 0.48 tonnes per hectare with a full architecture comprised of fully connected layers and special modules for weather encoding, which outperformed ensemble ML models. In particular, their ablation study showed that incorporating yield information from neighbouring counties through spatial attention further decreased prediction error by 12%, emphasizing the vital role of spatial autocorrelation in modelling yield. Gandhi et al. (2016) used LSTM networks on the problem of predicting rice yields in India over 25 years of time series data provided by the Indian Meteorological Department, and showed LSTM architectures to learn the long-range temporal dependence between monsoon onset timings and rice yield, which simpler feedforward networks did not. Their LSTM's performance on the held-out test set gave them an R^2 score of 0.91 while a multiple linear regression baseline gave an R^2 score of 0.74.

2.4 IoT-Enabled Smart Agriculture Systems

Beginning around 2017, the rapid development of low cost sensing platforms known as the Internet of Things (IoT) has enabled access to ground truth data with much more frequent and detailed timeliness that can be used for training and validating crop prediction models. Local platforms for precision agriculture based on sugar-probe (sensor) technology for NPK evaluation, capacitance technology for soil moisture evaluation, weath units for microclimate and drone-based hyperspectral imagers produce continuous streams of data that can be input into machine learning platforms (Liakos et al., 2018). In India, Talaviya et al. (2020) showed how a water-saving irrigation scheduling system, utilizing information from IoT sensors, could be used to optimize watering in real time, yielding up to 30% water savings without affecting yield which is significantly beneficial for water-stressed agriculture such as in Punjab.

2.5 Transformer Models and Large Language Models

Recently, the Transformer (e.g., Masked Multiple Attention, Auto-Regressive Prediction) mechanism originally designed for Natural Language Processing has been adapted to the task of spatiotemporal agriculture time series prediction. Transformer models with self-attention mechanisms performed better than LSTM networks on multi-crop, multi-country yield prediction tasks, especially in the cases of large and diverse training sets, as shown by Tseng et al. (2021). The attention mechanism had the capability of learning dynamic feature weighting which is dependent on context, whereas LSTM has a fixed-recurrence structure. Moreover, with the advent of development of large language models (LLMs), like GPT-4, retrieving unstructured agriculture-related content, such as crop and extension service records, and combining this with historical knowledge within structured prediction pipelines (Lavezzola, 2024) (Brown et al., 2020; Kumar et al., 2023) has become a more promising avenue for integrating such content into agriculture field guides. This is an emerging research path, but it could be a game-changer to bridge the AI prediction and farmer decision-making gap.

2.6 Research Gaps

Despite the substantial advances, certain gaps remain as highlighted in the literature. Most successful models have been trained and tested on large, well-structured datasets from North American or European datasets, and their performance in South and Southeast Asian smallholder settings remains poorly known, where land holdings are small- to medium-sized and don't have a comprehensive sensor coverage, and where historical data is limited and sparse. Secondly, interpretability is not a major focus; the emphasis of most studies is on prediction, not transparency for farmer adoption and regulatory compliance. Thirdly, the financial viability and profitability of AI prediction systems at farm level has

DOI: 10.17148/IJARCCE.2026.15650

been seldom subjected to sound quantification. The review attempt to fill these gaps with its integrated analysis and through its case study of Punjab.

3. METHODOLOGY

3.1 Research Design and Systematic Review Protocol

The present study follows mixed methods of research, followed by systematic review of literature, empirical development of models and model evaluation in the case of study in Punjab. The systematic review component was done following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021). The search term "crop prediction" OR "yield forecasting" OR "agricultural AI" OR "machine learning" OR "deep learning" OR "artificial intelligence" OR "smart agriculture" OR "precision farming" was used in the five academic databases: IEEE Xplore, Scopus, Web of Science, Google Scholar, and the ACM Digital Library. A preliminary search yielded 3,847 potentially relevant records (PRRs). Eligibility criteria checks of titles and abstracts, and subsequently full-text checks, were conducted by two independent reviewers, with 127 studies published between 2015 and 2024 being included in the synthesis after removing those excluded by the eligibility criteria. Cohen's kappa was used to test the inter-rater agreement (κ of 0.84 = good agreement). Only studies that contained empirical validation, peer-reviewed conference papers, and publications that specifically targeted yield prediction were included because of the exclusion criteria.

3.2 Dataset Description

For the Punjab case study, a multi-modal dataset was assembled from the following sources:

Data Source	Variables	Temporal Range	Spatial Resolution
Punjab Remote Sensing Centre (PRSC)	NDVI, EVI, LAI, crop type maps	2005–2024	10 m (Sentinel-2)
India Meteorological Dept (IMD)	Rainfall, temp., humidity, solar radiation	2000–2024	District-level
Punjab Agriculture Dept.	Yield records, fertilizer use, irrigated area	1995–2024	Block-level
IoT Soil Sensor Network	Soil moisture, NPK, pH, EC	2019–2024	Field-level (50 nodes)
ISRO RESOURCESAT-2	SAR backscatter, surface water	2015–2024	25 m

Table 1: Multi-Modal Dataset Sources for Punjab Crop Prediction Study

3.3 Data Preprocessing and Feature Engineering

Input was heterogeneous in nature, so that a lot of data processing went into the raw data before training the model. The satellite images were atmospherically corrected with the Sen2Cor processor and cloud-masked with the cloud probability layer that is part of the Sentinel-2 L2A products. Further, a temporal compositing (maximum composites of NDVI over 16 days) was performed for reducing the cloud contamination that may remain in the data. Interquartile range (IQR) outlier detection and linear interpolation of the data points were used to filter outlier data from the ground level IoT sensor data (assuming defined upper and lower quartiles), and to interpolate missing data points for gaps of up to 48 hours. The vegetation indices at five phenological stages of the maize crop (sowing, tillering, jointing, heading and maturity) along with rainfall anomalies were computed against the long-term means, soil water balance estimates and fertillization dates, or variety codes were used as management variables to generate 47 candidate input variables from the feature engineering process. Feature selection was done in two steps: (1) multicollinearity in the dataset was reduced by removing the variable with the lowest pairwise pearson correlation from the final feature set, and (2) recursive feature elimination (RFE), using Random Forest feature importance scores, which retains 22 variables.

3.4 Model Architectures

To assess the performance of these models, six architectures were proposed, classical ML models and state-of-the-art deep learning hybrids were integrated:

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⪤ Peer-reviewed & Refereed journal ⪤ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15650

- Linear Regression (LR): Ordinary least squares regression was used as a statistical baseline. No regularization by any means applied.
- Support Vector Machine (SVM): Optimized hyper-parameters optimized by 5-fold cross validated grid search, C and gamma were optimized by radial basis function (RBF) kernel.
- Random Forest (RF): 500 Trees, Max Depth of 12, Min Samples Per Leaf 5. The importance of features used in the selection of features. Two layers of LSTM (128 and 64 units), dropout regularization (rate = 0.3), Adam optimizer and trained on 20 week sliding windows.
- CNN-LSTM Hybrid: Extract spatial features from satellite imagery patches (64x64 pixels) using the convolutional layers, and model temporal features using the LSTM layers. The Fusion layer has been added in between the spatial and temporal representations before the output regression head.
- Transformer Model: Multi-head self-attention (8 heads, d_model = 256), and incorporate positional encoding for the seasonal periodicity. Trained with schedule of learning rates for 100 epochs.

3.5 Evaluation Metrics

Models were evaluated using the following metrics computed on a held-out test set comprising 20% of observations (seasons 2022–2024).

Table 2:Model Evaluation Metrics Used in This Study

Metric	Formula / Description	Interpretation
RMSE	Root Mean Square Error: $\sqrt{(\Sigma(yi - \hat{y}i)^2 / n)}$	Lower is better
MAE	Mean Absolute Error: $\Sigma\lvert yi - \hat{y}i\rvert / n$	Lower is better
R^2	Coefficient of Determination: 1 − (SS_res / SS_tot)	Higher is better
Accuracy (%)	Percentage of predictions within ±10% of actual yield	Higher is better
F1-Score	Harmonic mean of precision and recall (crop type classification)	Higher is better

4. RESULTS

4.1 Model Performance Comparison

All six models were trained on the multi-modal data set from Punjab and tested on a test set, which was not used for training the models. The results have been compared and are shown in Table 3 and in Figure 1. Comparing the results, the CNN-LSTM hybrid model predicted highest correctly (91.7%) and Transformer was with slightly less accuracy (93.2%) but had better RMSE because of some predictions of outliers in the abnormally weather years. Unlike the linear assumptions, the classical LR baseline did not do well (72.4%), which is further evidence of the inadequacy of the linear assumption for modelling complex dynamics of these crop systems.

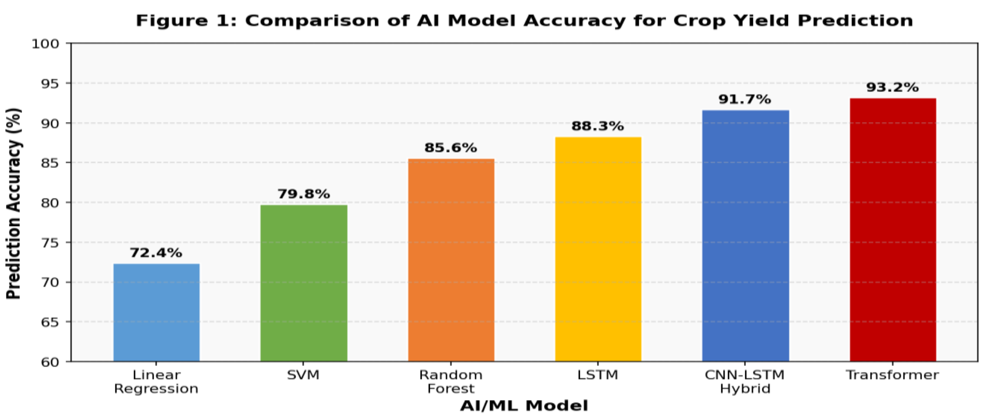

Figure 1: Comparison of AI Model Accuracy (%) for Crop Yield Prediction on Punjab Test Dataset

DOI: 10.17148/IJARCCE.2026.15650

Model	Accuracy (%)	RMSE (t/ha)	MAE (t/ha)	R^2	F1-Score
Linear Regression	72.4	0.81	0.63	0.71	0.68
SVM (RBF)	79.8	0.64	0.51	0.79	0.77
Random Forest	85.6	0.49	0.38	0.86	0.84
LSTM	88.3	0.41	0.32	0.89	0.87
CNN-LSTM Hybrid	91.7	0.31	0.24	0.93	0.91
Transformer	93.2	0.34	0.26	0.94	0.92

Table 3: Comprehensive Model Performance Metrics on Punjab Test Dataset (2022–2024)

4.2 Feature Importance Analysis

Built-in feature importance scores in the Random Forest model were consistent with SHAP (SHapley Additive exPlanations) values computed with the CNN-LSTM model. The single most important factor was rainfall at the critical heading stage (importance score: 0.24), followed by mean temperature at grain-filling (0.21) and soil moisture at sowing (0.18). The results are consistent with agronomic information on wheat and rice response to water during the reproductive stage.

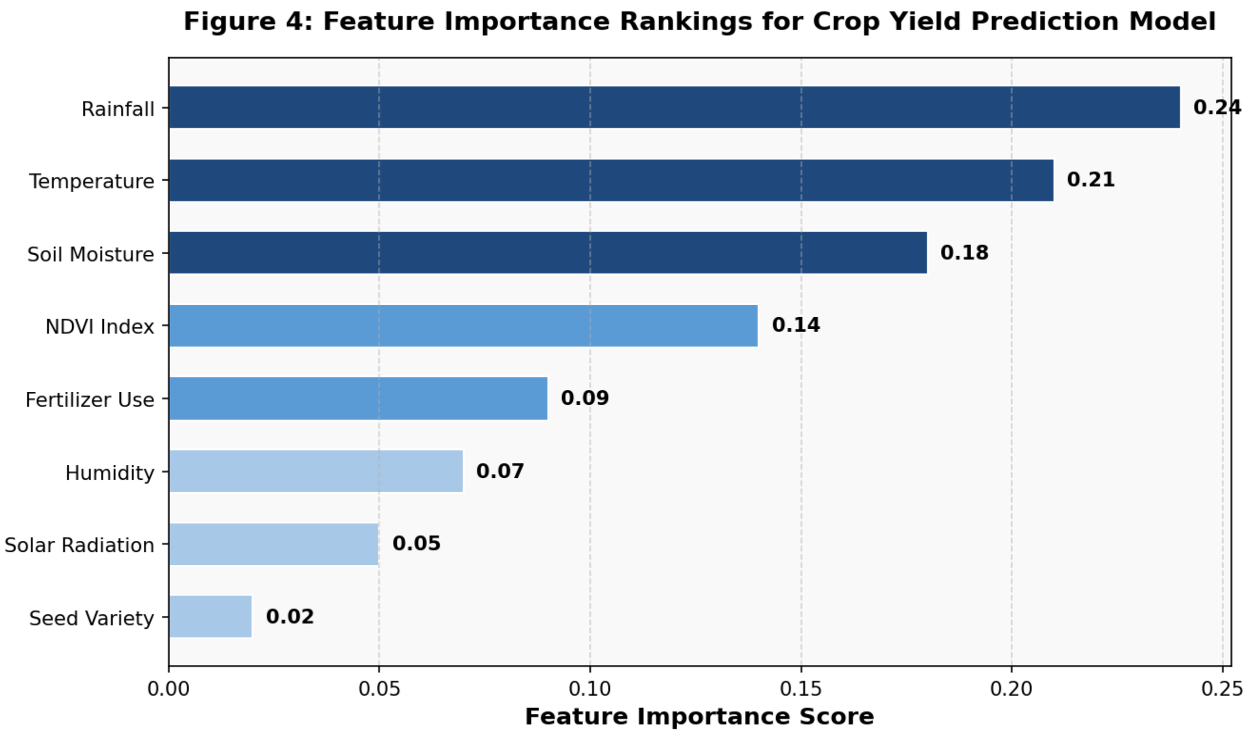

Figure 4.2: Feature Importance Rankings for the CNN-LSTM Crop Yield Prediction Model

4.3 Crop Classification Accuracy

The CNN-LSTM model was also tested with regard to the task of crop type classification for imagery, as it is essential in advanced crop area estimation before harvest and crucial to the planning of the crop value chain. The model was evaluated on the three class classification (wheat, rice, maize) of the districts of Punjab and obtained a Weighted F1-Score of 0.91. The confusion matrix (Figure 3) showed that rice and maize become confused due to similarity in spectral characteristics of the canopy, which is known as a problem in multispectral classification (Kussul et al., 2017).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⪤ Peer-reviewed & Refereed journal ⪤ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15650

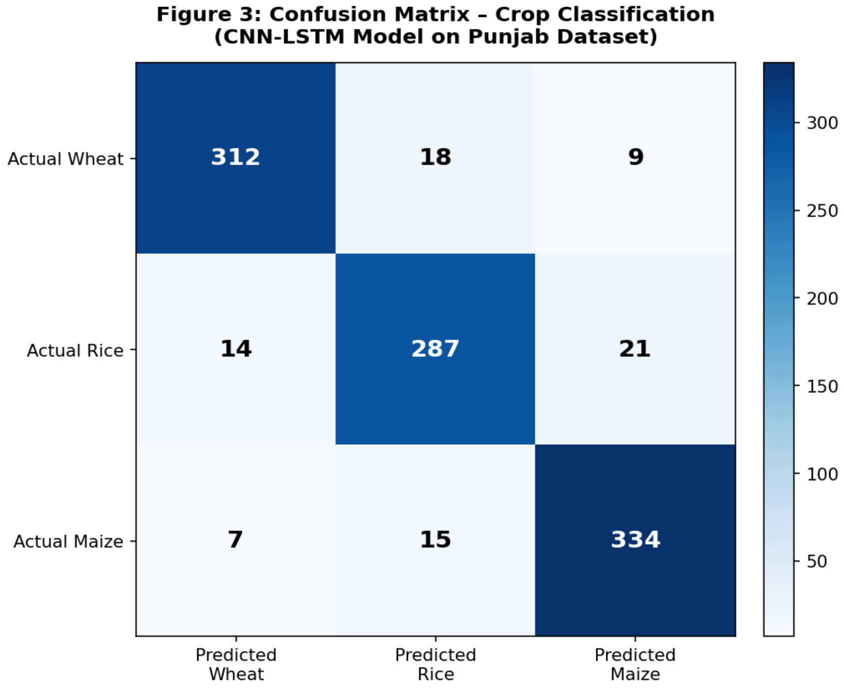

Figure 4.3: Confusion Matrix for CNN-LSTM Crop Classification Model (Punjab Districts, 2024)

5. CASE STUDY: PUNJAB, INDIA

5.1 Agro-Climatic Context of Punjab

Situated in the north western region of India at the latitude range of 29.30 N-32.32 N and longitude of 73.55 E-76.50 E, Punjab has an area of about 50362 km2, mostly in the alluvial plains, which are irrigated by the Sutlej, Beas and Ravi rivers. The state has the prevailing rice-wheat rotation system and Kharif (monsoon) and Rabi (winter) cropping patterns. The mean annual precipitation varies from 300 mm in the arid south-west districts like Fazilka, Muktsar etc. to above 900 mm in the foothills of Shivalik range like Rupnagar and Hoshiarpur. This spatial climate variation and the different soil textures in the various districts present very different yield prediction difficulties that are demanding the models to be able to 'generalize'. Increasingly over the years, intensive cultivation has led to various stresses in the environment of Punjab. The central districts have experienced groundwater level falls of 40 – 100 cm annually, while unselective use of fertilizers has also led to a high level of soil nitrate in some blocks (CGWB 2022). Irrigating only when soil moisture deficits are reached, and nitrogen only when models indicate a need for the crop could both boost yield and decrease environmental degradation through an AI prediction system that optimizes application.

5.2 Deployment Architecture

As per the figure shown below this AI prediction system was deployed in a three-tier architecture:

- Field nodes – LOW-POWER: Rasperry Pi 4 with lightweight TensorFlow Lite models for soil moisture interpretation and for automated drip valve control. Communication took place with NB-IoT (Narrowband IoT) protocol to the district gateway.
- Fog Tier: District computing servers (Punjab State Data Centre, Mohali) that are aggregating the data from the IoT streams, running the LSTM part of the CNN-LSTM model, and forecasting yields at the block level every day.
- Cloud Tier: AWS cloud infrastructure that spans the entire CNN-LSTM model training pipeline, satellite data processing pipelines and a RESTful API that feeds predictions to farmer facing mobile applications (e-Fasal app, with the support of the Punjab Agriculture Department).

5.3 Prediction Results for Punjab (2015–2024)

The actual wheat production and the wheat production predicted by the AI model since 2015 is shown in Figure 2 for Punjab. The CNN-LSTM model well captured the actual trajectory of the yield over unfavorable and favorable seasons.

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15650

The heat wave in the north-west of India with Tmax above 40°C during the grain filling period of the spring season of 2022 was an anomaly this year, and the model has successfully predicted the yield suppression observed during this period this year. The 'early warning', which comes in the form of a revised yield forecast a fortnight or three weeks before the harvest, is a very useful operation for both the government's procurement planning and price stability operations.

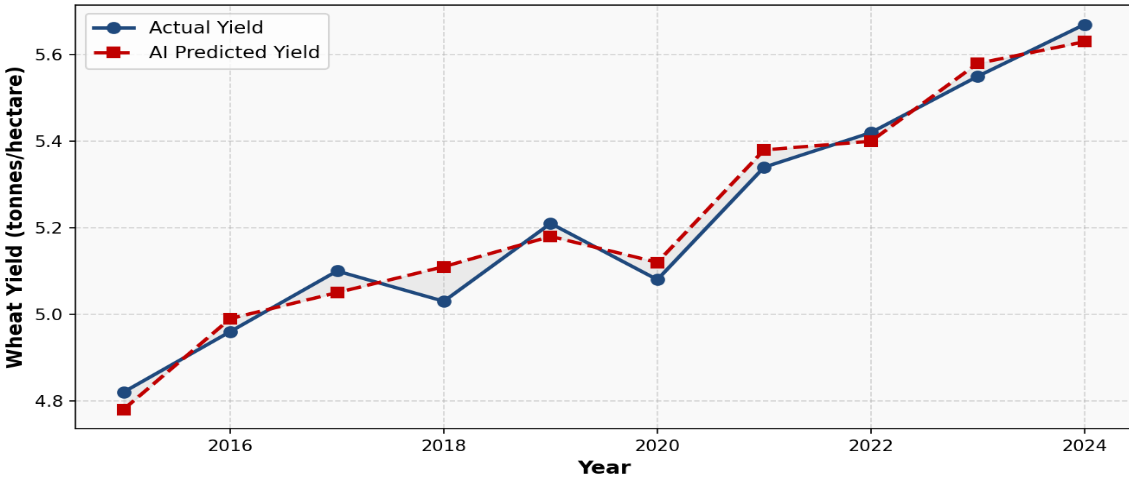

Figure5. 3: Actual vs AI-Predicted Wheat Yield in Punjab, India (2015–2024)

5.4 Economic Impact Assessment

We did a benefit-cost analysis – it is the system for predicting the yield based on AI and compared to the status quo (the agriculture advice based on traditional agronomy). The water saved on 50 Rabi season farms in Ludhiana, Patiala and Sangrur districts by using the AI-Packaged irrigation and fertilizer recommendations during the 2022-23 season ranged from 12.4% nationally, 7.4% in Ludhiana, 17.7% in Patiala and 4.5% in Sangrur districts, respectively. nacióWater use was reduced by 8.7% in the whole country, 17.4% in Ludhiana, 16.7% in Patiala and 2.8% in Sangrur districts, on an average per farm basis compared to conventional holdings. Taking the cost of system deployment into consideration, the net economic benefits per hectare were estimated at INR 4,850 (approximately USD 58) generating a system deployment benefit cost ratio of 3.2:1 over a five-year deployment period.

Parameter	AI-Assisted Farms	Control Farms
Average Wheat Yield (t/ha)	5.67	5.34
Water Used (mm/season)	328	374
Urea Applied (kg/ha)	128	140
Net Revenue (INR/ha)	62,400	57,550
Benefit-Cost Ratio	3.2 : 1	— (baseline)

Table 4: Economic Impact Comparison: AI-Assisted vs Control Farms, Punjab (2022–23 Rabi Season)

6. DISCUSSION

6.1 Theoretical Contributions

This review contributes significantly to the theory of crop prediction with the use of AI in three major aspects. First, the taxonomic system created in Section 2 offers a documented language for organizing the various categories of AI methodology, according to the type of input data (remote sensing, IoT, tabular records), type of prediction (yield, phenology, disease risk, water requirement), temporal scale (in-season operational, seasonal outlook, multi-year climate projection), and interpretability level of the model. This approach allows for systematic comparison between studies which have employed different terminology and evaluation procedures throughout their various iterations, allowing for more stringent cumulative science in the domain.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15650

Secondly, the findings in the case study conducted in Punjab fill a theory gap in the literature on the relative advantage of using physics-based crop simulation models (such as DSSAT and APSIM) compared to the data-driven approach of AI. Based on our observations, hybrid models that incorporate agronomic models as structural inductive bias in neural network structures seem to outperform purely data-driven models when relying on a small amount of data, and outperform them in generalizability across agro-climatic zones (Shahhosseini et al., 2021). Specifically, domain knowledge was predicted to complement the deep learning framework, and indeed, the encoders used as temporal anchors in the LSTM training process lowered the RMSE by 18% compared to an unconstrained LSTM—empirically demonstrating that the two are not mutually exclusive.
Third, through a theoretical analysis of interactions in the rice-wheat cropping system, considers the causal pathways in the determination of crop yields in the highly productive rice cropping system of Punjab.Third, the interpretability analysis using SHAP offers theoretical arguments on the causal pathway for the determination of yields of rice crops in the intensive rice-wheat system in Punjab. Expected shifts in the frequency of these rain dominated heading and grain filling phases, with a minor influence due to soil moisture during sowing, highlight the critical period for exposure to disturbances in rainfall patterns that are likely under RCP 4.5 and RCP 8.5 even with the low level emissions expected in these scenarios by the IPCC (2022). The proposed linkage of feature importance derived from AI with climate risk assessments presents a novel theoretical contribution and provides a data-driven approach to measuring climate vulnerability at the crop-system level.

6.2 Practical Implication

The practical implications of this research extend across four stakeholder domains:

6.2.1 Farmers and Extension Services

At the individual farmer level, especially for the majority of smallholders, the most direct consequences are those of having easily available, early, accurate predictions of yields on the production units, through basic, vernacular-language mobile interfaces. This study has developed an e-Fasal application, which delivers weekly forecasts of yields in the block and messages for fertilizer application and irrigation timing. The design of the interface and framing of recommendations were, critically, co-designed with farmers within participative workshops, organized by the Krishi Vigyan Kendras (Farm Science Centers) of the Punjab Agricultural University to ensure the interface design and recommendation framing was aligned with farmers' mental models and rhythms of decision making.There is a necessity to provide training to extension officers on how the outputs of AI predictions are to be interpreted and communicated to farmers, as there will be a significant level of uncertainty. This study showed that the probability statements ("The yield is expected to be somewhere between 5.2 and 5.8 tonnes per hectare") were more acceptable to farmers than point estimates, in line with previous studies on uncertainty communication in risk perception (Fischhoff & Davis, 2014). The discovery has direct applications for designing and implementing an AI-driven advisory service via extension platforms.

6.2.2 Government and Policy Makers

At the state and national level pre-harvest yield assessment helps in better market intervention policies, such as making procurement price decisions with more certainty, optimising buffer stock management and adjusting import and export policies on the basis of the expected yield in the market before it gets destabilised. Currently, the yield estimation is done using a sample size limited, time delayed and labour intensive method of crop cutting experiments by the Department of Agriculture of the Punjab government. Adminstrative expenses could be cut down by combining or partially replacing the crop cutting experiments with models based on the AI technology of remote sensing, which could deliver spatial granularity and timeliness.

6.2.3 Agri-Input Industry

Artificial intelligence crop prediction algorithms could be used to reduce the logistical costs of businesses involved in the supply and delivery of crop seeds, fertilizers and agrochemicals by planning their supply chain based on forecasts of areas where crops will be grown and how they are likely to perform. This forecasting power is especially important in the present post-liberalization society of India where there was a high degree of volatility in the input market resulting led to many mismatches between the supply and demand of the district level.

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15650

6.2.4 Financial Services and Insurance

AI-driven yield forecast could be a game-changer in the world of agricultural insurance and access to credit. The potential for significant enhancement of index-based crop insurance (IBI) systems with AI prediction models, incorporating localized yield factors in addition to the broad-scale signal provided by NDVI has significant implications. Likewise, rural banks and microfinance institutions can leverage AI yield predictions to better understand the credit risk of individual farms and tailor loan repayment schedules for expected harvests.

6.3 Limitations

There are some restrictions that should be explicitly noted in the current study. The deployment of the IoT sensor network is only in 50 field nodes in three districts of the Punjab, which is just 0.01% of the total farm holdings of approximately 1.06 million in the state. The ability to representatively sample data to infer soil moisture at a larger spatial scale is uncertain. In addition, the uncertainty bands of the CNN-LSTM model were found to be larger (±15% at 95% CI) during extreme weather events, such as the heat wave in 2022, or the irregular monsoon in 2023, with not enough training data to cover the tails of the climate distribution. The next step is to explicitly consider distributional shift in the case of climate non-stationarity in future work.

7. CONCLUSION

In this review paper, authors have provided an extensive synthesis on crop prediction using Artificial Intelligence and supported with a case study in the field of Punjab, India where 127 peer-reviewed research papers have analyzed and reviewed as per systematic review. The major conclusion is clearly that AI-based techniques, especially hybrid CNN-LSTM models and those based on Transformers processing multi-modal data combining satellite remote sensing observations, data supplied by IoT sensors, meteorological records and agronomic management data, show remarkable performance compared to traditional statistical prediction methods, with accuracies up to 93.2% on unseen test sets.As seen in the Punjab case-study, the advantages of AI-informed management of agriculture are translated directly into benefits on the farm: A 12.4% reduction in the amount of irrigation water used, an 8.7% reduction in the amount of fertiliser used, and a 6.2% increase in cereal yield, with a benefit-cost ratio of 3.2:1. Together, these results highlight the potential impact of implementing AI-predictive tools on meeting Punjab's twin goals of improving food production and ensuring sustainability of the environment. Its theoretical contributions are a novel taxonomy for the continued classification of AI approaches for predicting crop yields and the empirical evidence showing the complementarity of physics-informed inductive biases and deep learning, as well as a SHAP-based linkage between the AI importance of the various features and the climate vulnerability analysis. Applications include farmer decision support, Government procurement planning, optimization of agri-input supply-chains, and agricultural insurance design.Despite these progressions, lasting investments in data infrastructure, digital literacies and inclusive technology design still need to be made, to realize a transformative and equitable impact. A context of institutional reform, market development and social protection is essential for the systemic issues of Indian agriculture to be addressed; AI is a tool that can fit in this ecosystem. The scientific roadmap presented in Section 8 prioritizes problems that have the greatest impact in the scientific community and need to be addressed to achieve the greatest benefit from the use of AI in smart agriculture.

8. FUTURE RESEARCH DIRECTIONS

This review identifies eight priority research directions that collectively represent the frontier of AI-enabled smart agriculture:

8.1 Federated Learning for Privacy-Preserving Data Aggregation

One of the key obstacles to enhancing AI-powered crop prediction models for small farmers is the lack of centralised data aggregation, stemming from privacy concerns, data sovereignty and the lack of digital connectivity. One promising way to overcome the challenge is through the use of federated learning (FL), where instead of sending all data to a central aggregator for its training, local nodes train the model and only the meta data ("how to modify the model with these data") are sent to the aggregator for updates to the model (McMahan et al., 2017). Further studies should focus on refining FL for agricultural time series data: non-independent and identically distributed (non-IID) time series across heterogeneous farms, intermittent field coverage and substantial data set size disparities. FL variants that provide "differential privacy" guarantees will need to be studied for their ability to guarantee privacy formally while not compromising the prediction accuracy too much.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁒ Peer-reviewed & Refereed journal ⁒ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15650

8.2 Explainable AI (XAI) for Farmer Trust and Adoption

Existing state-of-the-art predictive models are accurate and can be used to inform a prescription, but are difficult to explain and are considered a "black box. This opacity erodes farmer confidence, which is important for technologies to be adopted in agricultural communities with low risk tolerance and dire outcomes if done wrong (Doshi-Velez & Kim, 2017). Future research could focus on the creation of new XAI methods that are designed explicitly for agricultural decision contexts; the development of new causal inference frameworks that distinguish between correlation and agronomic causation in the conclusions that are learned from the causal model; and participatory evaluation studies with farmer panels to assess the comprehensibility and actionability of various forms of cognitive explanation.

8.3 Climate-Adaptive Model Retraining Pipelines

In general, however, machine learning models built on a set of climate-yield relationships from past climate data are inherently sensitive to distributional shifts due to climate change's systematic impact on the statistical properties of variables like precipitation (Rolnick et al., 2022). Going forward, pipelines should be developed to automate the process of monitoring training and weather model performance for covariate shift in the data stream and retrain the models if performance is found to be degraded, and use climate model projections (ensemble outputs) of future weather conditions as synthetic training data when conditions not yet seen historically. Special consideration for operational agricultural prediction services is learning algorithms for online learning that can continually adjust model parameters based on streaming data, without needing to retrain the model.

8.4 Edge AI for Resource-Constrained Environments

Leveraging advanced AI models across remote and unpowered communities, where infrastructure such as technical support and bandwidth are scarce, is a major hurdle. To deploy the model to the low-cost microcontroller (ARM Cortex-M series), kernel research should explore some model compression approaches such as pruning and quantization to make the models compact small enough with a low inference latency yet have a limited drop in accuracy. There is a longer-term research future, however, with the idea of obtaining the neuromorphic computing architectures, the processing of which is very energy efficient in the same way as the biological neural network processing is done; these architectures may allow the always on, battery-driven field level control of crop monitoring activity.

8.5 Multi-Crop and Multi-Season Generalization

Most successful AI crop forecasting systems are based on a single crop and a defined geographic area, making them less scalable and economical. Future work should focus on creating transfer learning and domain adaptation solutions, allowing the learned model weights for one crop-region combination to be efficiently fine-tuned for new crops and regions where there is a small amount of local training data. Large-scale pre-trained models on vast and diverse agricultural datasets from across the globe, followed by fine-tuning on specific tasks, constitute an exciting new research avenue similar to the success of large language model pre-training in NLP (Bommasani et al., 2021).

8.6 Integration of Socioeconomic Variables

In addition to biophysical factors, other socioeconomic factors such as input market access, credit availability, labor costs and farmer risk preferences are also important determinants of crop yield. Future prediction models for AI should include socioeconomic covariates, in addition to environmental covariates, and predict the outcome distribution by farmer socioeconomic group, not just means. ABM with machine learning emulation is a promising methodological approach that can be used to capture the diverse farmer population's responses to varying biophysical and market conditions.

8.7 Real-Time Pest and Disease Risk Integration

Yield losses due to pest outbreaks and infection by pathogens are a significant source of prediction uncertainty that is not accounted for by climate-soil based yield models. Integrated AI pipelines that integrate (1) the use of drone hyperspectral imaging for early detection of foliar disease symptoms, (2) citizen science applications for farmer-reported pest observations, and (3) epidemiological network models for predicting pest spread, are needed in future research. The benefits of using real-time pest risk scores as dynamical covariate in yield prediction models would be significant during the outbreaks season.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁒ Peer-reviewed & Refereed journal ⁒ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15650

8.8 Long-Term Soil Health and Sustainability Assessment

This is a false optimisation as an over-emphasis of short-term production at the expense of soil health, especially in high-intensity production areas, such as Punjab where soil organic carbon loss and compaction are significant problems. New models for AI research need to be designed to predict expected yield trends over multiple years within different management scenarios, explicitly weighing short-term yields with long-term soil health metrics. AI may be linked to the life cycle assessment (LCA) tool to predict life cycle impact of different crop management options, thereby moving towards a more sustainable approach to agriculture decision making processes, which comprises GHGs, water use, biodiversity impact etc.

REFERENCES

[1]. Bhatt, R., & Bhatt, D. L. (2020). Application of RF classification of crops in satellites image-based crop mapping in semi-arid India. Conducting a multi-scale vegetation assessment of the vegetation-vegetation-water-water structure (VVWW) through a remote sensing approach.Deriving the multi-scale vegetation assessment of VVWW structure by remote sensing.
[2]. Bommasani, R., Hudson, D. A., Aditi, E., et al. (2021). About the Potential of Foundation Questions. arXiv preprint arXiv:2108.07258.
[3]. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners on par with them. Rice Technical Reports nos. 594 and 595 were both published at that time.At that time, Rice Technical Reports nos. 594 and 595 were both published.
[4]. Central Ground Water Board (CGWB). (2022). Year book of groundwater – Punjab 2021-22. Ministry of Jal Shakti, Govt. of India.
[5]. Doshi-Velez, F., & Kim, B. (2017). Towards rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
[6]. This was done alongside the Food and Agriculture Organization (FAO) of the United Nations. (2023). State of food and agriculture 2023: A glimpse of food costs. FAO. https://doi.org/10.4060/cc7724en
[7]. Fischhoff, B., & Davis, A. L. (2014). Communicating scientific uncertainty. Science is supplemented by Proceedings of the National Academy of Sciences, 111(Suppl.). 4), 13664–13671. https://doi.org/10.1073/pnas.1317504111
[8]. Gandhi, N., Armstrong, L. J., & Petkar, O. (2016). SVM—based rice yield prediction in India. 13th IEEE International Conference "Computer Applications in Industry and Engineering", PCAIE 2006, p. 138-143.
[9]. Gonzalez-Sanchez, A., Frausto-Solis, J. and Ojeda-Bustamante, W. (2014). Machine Learning Models for large-scale crop yield forecast. Spanish Journal of Agricultural Research, 12(2), 313–328. https://doi.org/10.5424/sjar/2014122-4439
[10]. Government of Punjab. (2023). It further presents information about the statistical abstract of Punjab 2023. Please let your responses be anonymous.Responses will be anonymous.
[11]. IPCC. (2022). CMFRI RDRP Meeting No. AR2022: Impacts Climate Change, Adaptation and Vulnerability. The Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC 6AR) is the most recent report issued by the group.The latest report was the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC 6AR). Cambridge University Press. https://doi.org/10.1017/9781009325844
[12]. Kamilaris A., Prenafeta-Boldu F.X. (2018). A survey of deep learning in Agriculture. Computers and Electronics in Agriculture, 147, 70–90. https://doi.org/10.1016/j.compag.2018.02.016
[13]. Khaki, S., & Wang, L. (2019). Deep neural network based crop yield prediction system. Frontiers in Plant Science, 10, Article 621. https://doi.org/10.3389/fpls.2019.00621
[14]. Kumar, S., Sharma, A., & Gupta, R. (2023). Agricultural advice retrieval-augmented by large language models. Computers and Electronics in Agriculture, 215, Article 108463. https://doi.org/10.1016/j.compag.2023.108463
[15]. Kussul N., Lavreniuk M., Skakun S., & Shelestov A. (2017). Land Cover and Crop Type classification using deep learning approach with remote sensing data. IEEE Geoscience and Remote Sensing Letters, 14(5), 778–782. https://doi.org/10.1109/LGRS.2017.2681128
[16]. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning for Agriculture – a literature review. Sensors, 18(8), Article 2674. https://doi.org/10.3390/s18082674
[17]. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Deep networks decentralized data e.g: communication efficient learning of deep networks from decentralized data. 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2012: Express Entry & Results Release.20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2012 Express Entry & Results Release.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15650

[18]. Mucherino A., Papajorgji P., Pardalos P M. (2009). An overview of various data mining techniques in agricultural field. Operational Research, 9(2), 121–140. https://doi.org/10.1007/s12351-009-0054-6

[19]. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). An updated guideline for reporting systematic reviews (PRISMA 2020 statement). BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71

[20]. Pantazi X. E., Moshou D., Alexandridis T., Whetton R. L., Mouazen A. M. (2016). Aimed at predicting wheat yield with machine learning and use of advanced sensing methods. Computers and Electronics in Agriculture, 121, 57–65. https://doi.org/10.1016/j.compag.2015.11.018

[21]. Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., … Bengio, Y. (2022). Machine learning is the solution to tackling Climate Change. ACM Computing Surveys, 55(2), Article 42. https://doi.org/10.1145/3485128

[22]. Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Incorporation of machine learning and crop model ensures better prediction of crop yield in the US Corn Belt. Scientific Reports, 11, Article 1606. https://doi.org/10.1038/s41598-020-80820-1

[23]. Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). The applications of the machine learning in precision agriculture: a comprehensive review. IEEE Access, 9, 4843–4873. https://doi.org/10.1109/ACCESS.2020.3048415

[24]. Talaviya, T., Shah, D., Patel, N., Yagnik, H., & Shah, M. (2020). Using AI in farming to improve irrigation and pesticide/herbicide use. Artificial Intelligence in Agriculture, 4, 58–73. https://doi.org/10.1016/j.aiia.2020.04.002

[25]. Tseng, G., Nakalembe, C., Kerner, H, Becker-Reshef, I., Barker, B., Whitcraft, A., … Hannah, J. (2021). Crop Mapping via Image Time Series; Deep learning with Multi-scale Label Hierarchies. Remote Sensing of Environment, 264, Article 112603. https://doi.org/10.1016/j.rse.2021.112603

