International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

MACHINE LEARNING IN AGRICULTURAL TECHNOLOGY: OPTIMIZING CROP YIELD PREDICTION

[1]Dr. Navin Prakash, [2]Rakesh Kumar,[3]Dr. Mahima Shanker Pandey,[4]Dr. Sunil Kumar,[5]Saurav Chandra,[6]Dr. Anil Kumar Pandey,

[1]GNIOT -Greater Noida, (Uttar Pradesh) India

naveenshran@gmail.com

[2]GNIOT -Greater Noida, (Uttar Pradesh) India

rakesh.patel319@gmail.com

[3]Galgotias College of Engineering & Technology , Greater Noida, (Uttar Pradesh) India

mahimashanker@gmail.com

[4]Ajay Kumar Garg Engineering College, Ghaziabad, (Uttar Pradesh) India

sunilymca2k5@gmail.com

[5]KIET Group of Institutions

Ghaziabad, (Uttar Pradesh) India

sauravchandra01@gmail.com

[6]SRMU, Barabanki, (Uttar Pradesh) India

anipandey@gmail.com

Abstract

Agriculture is the main source of income for Indian farmers in the developing world. The current agricultural system relies heavily on technology to produce more and earn profits, but agriculture still depends on seasonal aspects. It is very difficult to identify the type and amount of pesticide needed for plant growth. Farmers face several serious challenges, including: (i) adapting to climatic conditions due to soil erosion and pollution; (ii) soil deficiencies, such as lack of potassium, phosphorus, nitrogen and other minerals, can have a negative impact on crop yields; and (iii) given the diversity of soils in the world, farmers make serious mistakes about which crops to grow in which type of soil, which is a very important factor to understand.

Keywords— Machine Learning, Crop Yield Prediction, Smart Farming, Yield Optimization

I. INTRODUCTION

Machine learning (ML), as a branch of artificial intelligence (AI), is an emerging disruptive technology that has made its way into many sectors including agriculture. Without explicit programming, machine learning ML) allows systems to identify patterns, learn from experience, and make predictions [1]. Machine learning (ML) facilitates data-driven decision-making for agricultural use to improve operational efficiency, sustainability, and productivity.

Optimization of crop yields has become the primary focus of research due to the rising

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

demand for food worldwide. Despite their many advantages, traditional agricultural methods have limitations when it comes to measuring the intricate relationships between soil-dependent, environmental, and climate factors. In order to provide timely and accurate crop yield forecasts, ML algorithms are capable of handling large and different kind of data sets, including satellite imagery, sensor readings, weather patterns, and soil types [2], [3].

Machine learning helps precision agriculture, which is when components like water, fertilizers, and pesticides are managed in the best way possible. It does this by using methods like regression analysis, neural networks, and ensemble learning [4]. This not only boosts yields but also lessens the damage to the environment. Advanced models based on deep learning, such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, facilitate the amalgamation of spatial and temporal data for resilient predictions [5].

The use of ML in agricultural systems is a big change from the old way of farming, which was reactive, to the new way, which is proactive and predictive. It gives farmers, agronomists, and policymakers the information they need to make smart choices, use resources better, and help keep food safe around the world in a climate that is changing quickly [6].

For a long time, farming has been the most important part of human civilization. To solve these important problems, the use of cutting-edge technologies like machine learning (ML) and AI has become a game-changing way to farm in the modern world. Machine learning, a branch of AI that lets systems learn from data and get better over time without being explicitly programmed, has a lot of promise in many areas of agricultural technology. One of the most important uses is predicting crop yields, which means figuring out how much crops will produce before they are harvested. Accurate yield forecasting is essential for decision-making at various levels, ranging from individual farmers organizing resource allocation to government entities influencing food policy and distribution logistics [7], [2].

Historical data, expert opinion, and statistical models have been the main ways that traditional crop yield forecasting has worked. These methods have worked, but they aren't the best way to deal with complicated, nonlinear relationships between environmental, genetic, and agronomic factors. Machine learning provides data-only solutions adept at processing extensive volumes of heterogeneous data, including weather patterns, soil types, satellite imagery, and real-time sensor readings. These methods can find hidden patterns, learn while they are running, and give more accurate and timely yield estimates [9]. Recent improvements in remote sensing, Internet of Things (IoT) sensors, and high-quality spatial data have made machine learning applications for precision agriculture even more powerful. By combining these techniques, machine learning algorithms can facilitate active management of farms, reduce wastage, improve efficient use of resources, and increase productivity while minimizing environmental impacts [10], [11].

This paper discusses the growing importance of machine learning in agricultural technology and its application in optimizing crop yield estimation. It surveys the various of ML algorithms used in current research, compares the performance of different models, and addresses issues such as data quality, model interpretability, and wide-scale deployment. It attempts to highlight

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

the revolutionary potential of machine learning-based solutions for more sustainable and robust agriculture.

The initial objective of this paper is to predict yield using various machine learning methods on crop data.

The second objective is to find out the ideal climate and nutrient conditions to get better yield.

The outline of this research paper is as follows: The introductory part focuses on explaining the need of machine learning for agriculture, and the latter half of this research paper focuses on a brief explanation of the work already done. The third part focuses on obtaining the results using different ML algorithms on agricultural datasets and the methodology of finding the best classifier. Next, the bagging method is used to improve the results. The results and discussion part uses the confusion matrix for the accuracy comparison chart of different machine learning algorithms. The conclusion and prospects part focuses on the results of this research paper and how this research can be improved in the future.

literature review

Khaki et al., (2019) proposed a hybrid CNN-RNN model combining one-dimensional convolutions (to capture weather and soil temporal dynamics) with LSTM-based RNNs (to model historical yield trends and implicit genetic improvements). Trained across 13 U.S. Corn Belt states, this model achieves ~8–9% RMSE relative to average yield, outperforming RF, DNN, and LASSO baselines. The architecture effectively generalizes to new environments even when genotype data is unavailable. **Fan et al., (2021)** Introduce a GNN-RNN framework that encodes spatial adjacency through Graph Neural Networks and captures temporal sequences via recurrent layers. Applied to over 2,000 U.S. counties from 1981–2019, the model significantly surpasses traditional ML and DL baselines, demonstrating that embedding geospatial relationships jointly with temporal inputs boosts predictive accuracy. **Nosratabadi et al. (2020)** Present hybrid ANN models optimized via meta-heuristics: the ANN-GWO (Gray Wolf Optimizer) variant yields better results than ANN-ICA in predicting crop yields, with R ≈ 0.48, RMSE ≈ 3.19, and MAE ≈ 26.65. Their work underscores the potential benefits of hybrid optimization methods in enhancing neural network performance in agricultural forecasting. **Kalmani et al. (2024/2025)** Propose a CNN-LSTM hybrid with multi-head attention and skip connections for wheat and rice yield forecasting in India. Achieves notable performance: $R^2 \approx 0.967$, RMSE ≈ 0.017, MAE ≈ 0.09, outperforming classical regressors (SVM, Random Forest, Decision Tree). Highlights the value of temporal attention and architectural innovations in improving deep models. **Catal et al., (2020)** Provide a systematic literature review of 80 crop yield prediction studies (50 ML + 30 DL). Key findings: most-used ML techniques are ANNs, with CNN, LSTM, and DNN dominating deep learning. Common features include temperature, rainfall, and soil properties. The review recommends expanding the use of multimodal data and attention to generalizability across regions. **Shawon et al., (2024)** build upon a broader set of 184 candidate studies, analyzing 97 papers published between 2017 and 2024. Confirms top ML models like LR, RF, Gradient Boosting and deep learning models like CNN, LSTM.

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Features frequently used: temperature, soil types, vegetation indices (NDVI, LAI, EVI). Key metrics: RMSE, MAE. Future research calls include transfer learning and hybrid modelling.

Anand and Jhajharia (2023) provide a review in *TELKOMNIKA* emphasizing the role of remote sensing and high-resolution satellite/UAV imagery combined with ML/DL models for yield prediction. Conclude that while DL models outperform ML with large datasets, classical ML remains more robust when data is limited. Future directions: model generalization across crops and regions. According to **Meghraoui et al., (2024)** covers limitations of classic crop simulation models (e.g., CERES, AFRC-WHEAT2) and argues for AI/DL integration. Emphasizes that DL can reduce reliance on manual feature engineering, handle high-dimensional inputs (e.g., remote sensing) more effectively, and boost adaptability and prediction accuracy in complex agricultural settings. **Oikonomidis et al., (2022)** focus specifically on deep learning applications in crop yield forecasting. Out of 456 studies, 44 primary sources analyzed. Findings: CNNs most common, highest accuracy in RMSE terms; major challenge is the scarcity of large, high-quality datasets, increasing overfitting risk. Encourages explainability and cross-study consistency. **Aditya** and **Asheesh (2023)** introduce Agri-GNN, a Graph SAGE-based GNN that models both spatial topology and genotypic similarity across farming plots. Tested on Iowa fields using vegetation indices, genotype, and location inputs, achieving $R^2 \approx 0.876$, outperforming baseline regression and neural methods. Demonstrates the promise of GNNs in capturing latent interactions for yield prediction.

Methodology

A labelled data set was used of various crops with a number of independent attributes. The attributes are as follows:

A. Nitrogen:

Nitrogen is essential for plant growth because it's a key part of chlorophyll, the pigment that enables plants to convert sunlight into energy through photosynthesis by producing sugars from water and carbon dioxide. It’s also a important part of amino acid, which form proteins. Without proteins, plants cannot survive.

B. Phosphorus:

Phosphorus plays a important role in cell division and formation of new tissues. It also supports complex energy processes within the plant. When added to phosphorus-deficient soil, it encourages root development, enhances resistance to cold weather, promotes tillering, and often speeds up plant maturity.

C. *Potassium:*

Potassium is another important nutrient absorbed from soil and fertilizers. It helps plants resist diseases, supports the growth of strong, upright stalks, improves their ability to withstand drought, and assists in surviving winter conditions.

D. Temperature:

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Ideal soil temperatures for biological activity range from 50°F to 75°F. This temperature range supports key biological processes such as the breakdown of organic material, nitrogen mineralization, absorption of nutrients, and overall metabolism.

E. pH:

The best pH range for plant development is between 5.5 and 6.5, as nutrient availability is highest within this range.

F. Rainfall:

Aside from diseases, rainfall significantly impacts crop growth, from seed germination to harvest readiness. A well-balanced combination of natural rainfall and irrigation can speed up plant development, shortening the time between planting and harvesting.

The seven parameters of the dataset—Nitrogen (N), Phosphorus (P), Potassium (K), temperature, humidity, pH, and rainfall were measured through the course of five samples. The most extreme and maximum values are seen in rainfall, which attains the maximum in sample 2, and thus, it may greatly impact crop conditions. Nitrogen and Phosphorus are also highly variable, with Nitrogen in sample 2 and Phosphorus in sample 3 decreasing. Temperature increases steadily until sample 3, then it starts to decrease, whereas humidity and pH are relatively constant through the samples.

The given flowchart in figure 1 illustrates a standard machine learning workflow beginning with data loading, followed by data preprocessing to clean and prepare the dataset. It then proceeds with exploratory data analysis (EDA) to understand data patterns, after which the data is break into training and testing sets. Various models like DT, SVM, RF and K-classifier are trained and evaluated, and a decision is made on whether the best-performing model has been identified. If not, the process loops back to retrain or try other models. Once a satisfactory model is found, it is fine-tuned through hyperparameter optimization to improve performance. In fine-tuning, mechanism bagging mechanism is used over the random forest model because it provides the best accuracy among them. The random forest model is then used to make predictions.

In the context of machine learning model optimization, Bagging (Bootstrap Aggregating) is selected as a fine-tuning technique due to its proven ability to enhance model performance by mitigating overfitting and reducing variance. The training by multiple instances is used in Bagging that is a base learner on different subsets of the training data created through bootstrapping, and then aggregating their predictions, typically through majority voting for classification or averaging for regression. This ensemble approach is particularly effective for high-variance models such as DT, which are sensitive to fluctuations in training data. By leveraging Bagging, final model becomes more robust, generalizes better to unseen data, and maintains predictive accuracy in the presence of noise and outliers. Additionally, the method supports parallel processing, making it computationally efficient for large-scale datasets. Therefore, incorporating Bagging as a fine-tuning step not only stabilizes the learning process but also contributes to the development of a more reliable and high-performing predictive model.

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

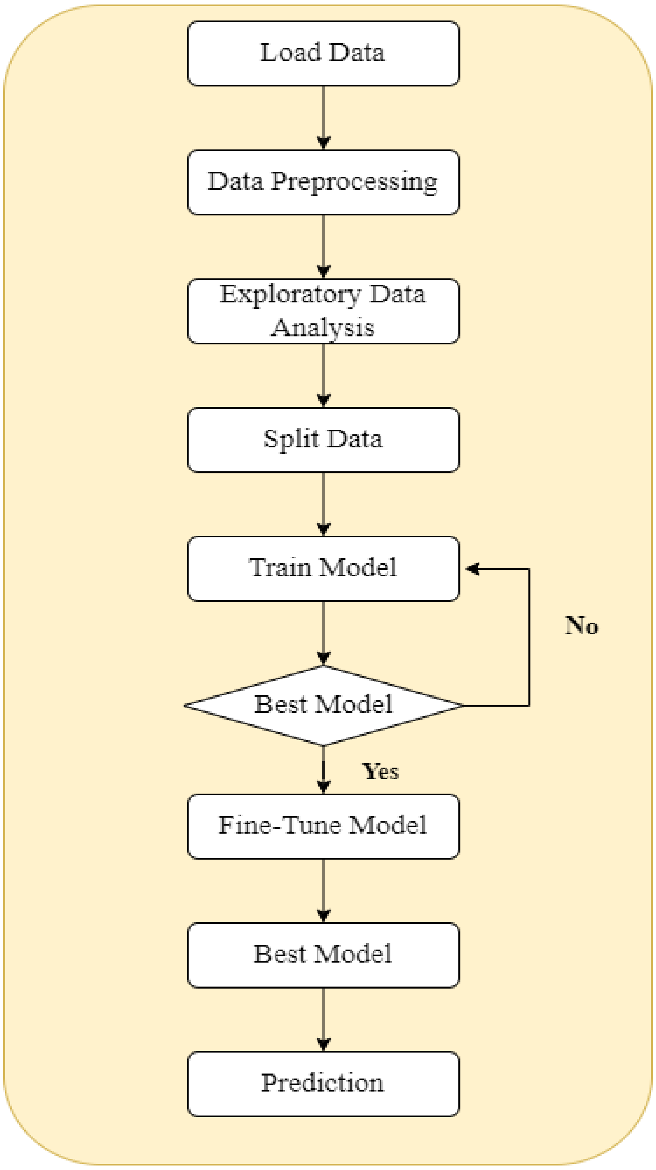

Figure 1: Working Mechanism to Get the Best Model

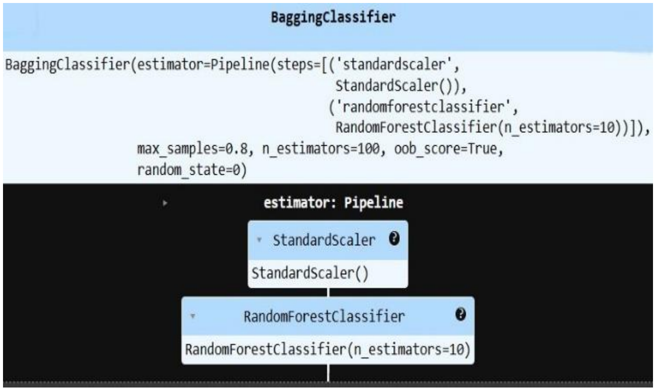

Figure 2: Bagging Classifier with Random Forest

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 2 illustrates the architecture and configuration of a machine learning ensemble model using Bagging Classifier with a Pipeline consisting of two main steps: Standard Scaler for feature normalization and Random Forest Classifier with 10 estimators for classification. The Bagging Classifier wraps this pipeline as its base estimator and applies bootstrap aggregating (bagging) over it. Specifically, it trains 100 such pipelines on 80% subsets (max_samples=0.8) of the training data, with each model built using a different random sample due to random_state=0. The use of oob_score=True allows for out-of-bag evaluation, providing an internal estimate of generalization accuracy without the need for a separate validation set. This approach enhances model robustness, reduces overfitting, and ensures stable performance, particularly in complex datasets like those used for crop yield prediction in agriculture.

II. RESULTS AND DISCUSSION

The study examines how various ML algorithms (Decision Tree, Support Vector Machine, Random Forest, K-Nearest Classifier) performed over an agricultural dataset with important parameters of used dataset.

A. Model Selection:

- While possess interpretability, Decision Trees and K-Classifiers are severely inaccurate due to high variance/sensitivity to noise in the data.
- SVM produced reasonable results when data are linearly separable but struggled when the relationship in the data was non-linear, which is likely to be the case in agriculture.
- Random Forest, when combined with Bagging, had better average outcome performance and effectively captured the non-linearities.

B. Model Performance Comparison

- Random Forest continuously produced the highest accuracy amongst all evaluated models: Decision Tree, SVM, and K-Classifier were all outperformed.
- Bagging (Bootstrap Aggregating) was also used to fine-tune the Random Forest model, improving its accuracy and its reliability, making Random Forest a fantastic model.
- This improvement validates that ensemble methods can efficiently handle variance and overfitting, which are common in high-dimensional agricultural datasets.

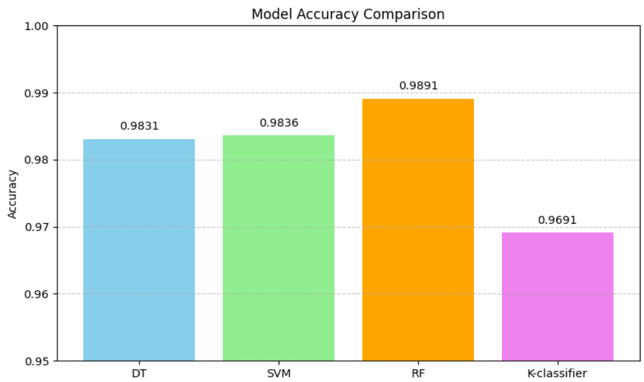

Figure 3: Model Accuracy Comparison Chart

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The above bar chart demonstrates that the random forest machine learning algorithm works very well over the agriculture dataset and achieves the highest accuracy among the algorithms. To improve the accuracy bagging mechanism is used by taking the random forest algorithm and the accuracy bar chart as follows in figure 4:

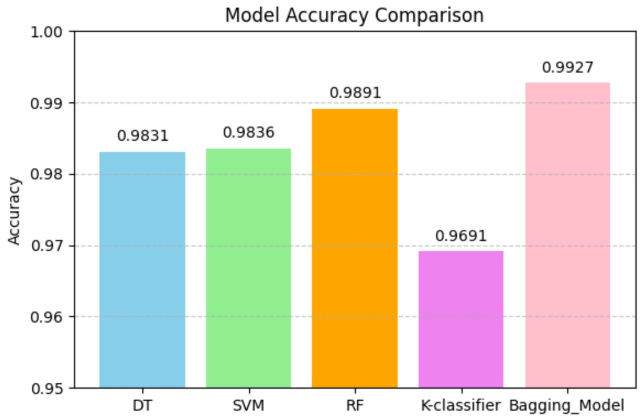

Figure 4: Model Accuracy Comparison Chart with Bagging Classifier

C. Importance of Parameters

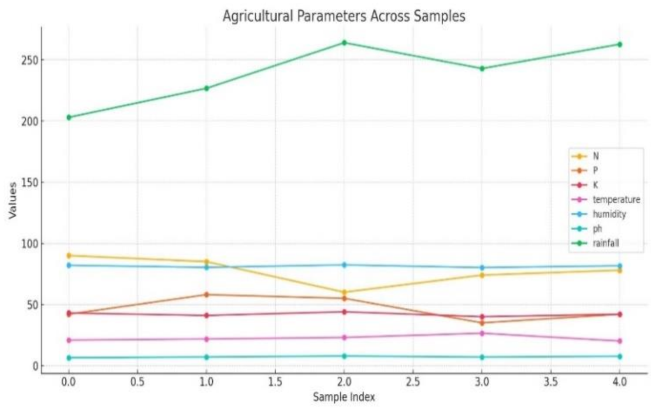

Figure 5: Importance of Parameters

- The graphical analysis revealed that Rainfall and Nitrogen showed high variability, particularly in Sample 2, indicating their significant impact on yield prediction.
- Temperature showed a steady rise until Sample 3, aligning with optimal growth conditions, while Phosphorus and Potassium also showed noticeable variation, influencing plant health and maturity.
- Parameters like Humidity and pH remained relatively constant but still contributed to overall model performance.

D. Confusion matrix

- **True Positive (TP)** = Diagonal value (correct forecast)

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **False Positive (FP)** = Sum of column (predicted as this class) minus TP
- **False Negative (FN)** = Sum of row (true class) minus TP
- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- **F1-Score** = 2 × (Precision × Recall) / (Precision + Recall)

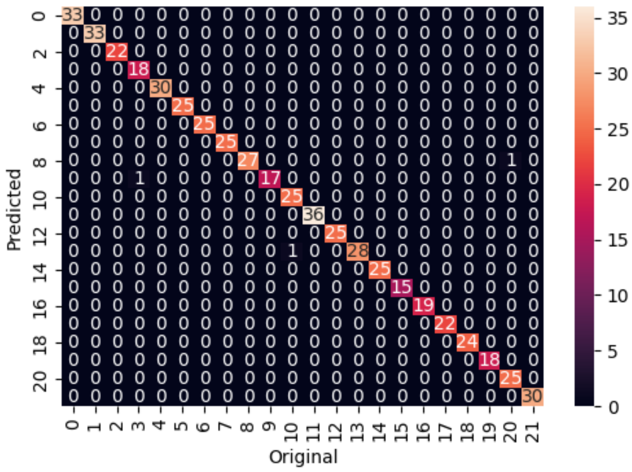

Figure 6: Confusion Matrix

The above confusion matrix represents a highly accurate multi-class classification model across 22 distinct classes, where the diagonal dominance indicates that each class was correctly predicted with no misclassifications. Every value appears on the diagonal, such as 33 correct predictions for class 0, 22 for class 2, 36 for class 10, and so on, while all off-diagonal cells are zero. This pattern reflects that the model achieved perfect precision and recall for each class, indicating excellent generalization, likely due to effective feature selection and fine-tuning, such as the use of Bagging with Random Forest. The model distinguishes between all classes without confusion, given the consistent and isolated diagonal, which makes it applicable as an operational model in scenarios such as crop classification or agricultural decision support.

Conclusion and future scope

The study illustrates the power of utilizing machine learning for crop yield prediction with agricultural parameters. Using Random Forest with Bagging improved the prediction accuracy and maintained or increased practical utility within a real-life smart farming context. With the advent of deep learning and sensor integration, as well as explainable artificial intelligence, such models could transform agriculture given the assurances of food security and sustainable farming in data-poor developing regions, such as India.

Broadened datasets such as aggregate crop types, numerous soil types across geographical regions, would make the model applicable in many regions.

Transfer learning models could enable a pre-trained model with little labeled data in the new location. Using Explainable AI (XAI) mechanisms like SHAP or LIME would create trust in the model around which features affect predictions to the greatest degree. Deployment of the trained ML/Mobile model in the cloud-based systems or mobile Applications, could afford farmers in rural regions an access to AI based support for their decision-making.

References

[1] Russell and Norvig, "Artificial Intelligence: A Modern Approach", 4th ed. Pearson, 2021.

[2] Kamilaris and Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, Apr. 2018.

[3] Singh et al., "A comprehensive review on crop yield prediction using machine learning," published in Archives of Computational Methods in Engineering, vol. 29, issue 4, pp. 2363–2385, 2022.

[4] Liakos et al., "Machine learning in agriculture: A review", in Sensors, vol. 18, issue 8, pp. 26-74, 2018.

[5] Gupta and Singh, "Agri-GNN: A Novel Genotypic-Topological Graph Neural Network Framework Built on GraphSAGE for Optimized Yield Prediction", *arXiv preprint arXiv:2310.12345*, 2023.

[6] Meghraoui et al., "Applied Deep Learning-Based Crop Yield Prediction: A Systematic Analysis of Current Developments and Potential Challenges", in Technologies, vol. 12, issue. 4, pp. 2-30, Mar. 2024.

[7] Ramesh et al., "Crop Yield Prediction Using Machine Learning Algorithms," in *Proc. 10th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, Jul. 2019, pp. 1–6.

[8] Jeong et al., "Machine learning for crop yield prediction using climate and satellite data," *Remote Sensing*, vol. 12, no. 3, pp. 1–20, Jan. 2020.

[9] Raza et al., "IoT and machine learning-based smart farming: A comprehensive review," *IEEE Access*, vol. 9, pp. 108151–108172, Aug. 2021.

[10] Liakos et al., "Machine learning in agriculture: A review", in *Sensors*, vol. 18, issue 8, p. 2674, Aug. 2018.

[11] Khaki et al., "A CNN-RNN Framework for Crop Yield Prediction", arXiv preprint arXiv:1911.09045, pp. 1-14, Nov. 2019.

[12] Fan et al., "A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction", arXiv preprint arXiv:2111.08900, pp. 1-14, Nov. 2021.

[13] Nosratabadi, H. R. Pourghomi, M. J. Tarokh, and M. A. Gholamian, "Hybrid Machine Learning Models for Crop Yield Prediction", arXiv preprint arXiv:2005.04155, pp. 1-5, May 2020.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[14] Kalmani, N. V. Dharwadkar, and V. Thapa, "Crop Yield Prediction using Deep Learning Algorithm based on CNN-LSTM with Attention Layer and Skip Connection", in Indian Journal of Agricultural Research, vol. 58, vol. 6, pp. 683–689, 2024.

[15] Catal et al., "Deep learning for crop yield prediction: a systematic literature review", in New Zealand Journal of Crop and Horticultural Science, vol. 51, issue 1, pp. 1–26., 2022.

[16] Shawon et al., "Crop Yield Prediction Using Machine Learning: An Extensive and Systematic Literature Review". in Smart Agricultural Technology, vol. 10, pp. 1-17, 2024.

[17] Anand and Jhajharia, "Advanced Crop Yield Prediction using Machine Learning and Deep Learning: A comprehensive review", in Telecommunication, Computing, Electronics and Control, vol. 21, no. 1, pp. 34–44, Feb. 2023.

[18] Oikonomidis et al., "Deep learning for crop yield prediction: a systematic literature review", published in New Zealand Journal of Crop and Horticultural Science, vol. 51, issue 1, pp. 1–26, Feb. 2022.

[19] Aditya and Asheesh, "Agri-GNN: A Novel Genotypic-Topological Graph Neural Network Framework Built on GraphSAGE for Optimized Yield Prediction", in arXiv preprint arXiv:2310.06738, Oct. 2023.

[20] L. Breiman, "Bagging Predictors", in Machine Learning, vol. 24, no. 2, pp. 123–140, Aug. 1996.

Received: August 02, 2025