

AgriSight: An Integrated Multi-Parametric Framework for Crop Recommendation and Yield Prediction Leveraging Environmental and adaphic factors

Nayna Potdukhe¹, Ganesh Dandekar², Dhairyashil Shinde³, Prathmesh Murodiya⁴, Bhavesh Bhumar⁵, Nishith Sanap⁶

¹Visiting Lecturer,

^{2,3,4,5,6}Student Department of Artificial Intelligence and Machine Learning
Government Polytechnic, Nagpur, India.

Abstract - Agriculture is really important for Indias economy, especially in places like Maharashtra where how well crops grow depends a lot on the weather and the soil. Farmers there have to deal with that all the time. I think predicting crop yields accurately could make a big difference for them and also for people making policies or planning resources. It helps with managing crops and figuring out food supplies. This study comes up with something called AgriSight, which is a machine learning setup to predict yields based on past production data and weather stuff. The whole idea is to handle two main problems in precision agriculture, like picking the right crop for the land and then guessing how much it will produce. It seems kind of tricky to get both right. The framework pulls in different kinds of data, including weather things like temperature, rainfall, humidity, and soil details such as type and pH. That makes sense because those factors affect everything. They combined historical crop records with weather history to build the model. Then they did some feature engineering to pull out useful bits, like the area planted, average yields from before, temperature levels, and humidity. A bunch of machine learning methods got tested, including Random Forest, Gradient Boosting, XGBoost, Neural Networks, and LightGBM. It feels like they wanted to see what worked best. In the end, Random Forest came out on top with about 98.49 percent accuracy and an error of just 0.49 tons per hectare. Thats pretty good, I guess. Mixing the old agricultural data with environmental info really boosted how well it predicted things. Some people might think its oversimplifying, but the results show it helps. AgriSight could be useful for precision farming and better planning in agriculture. It might even tie into food security somehow, though Im not totally sure on all the details there. The performance numbers stand out, but integrating everything wasnt straightforward.

Key Words: Precision Agriculture, Crop Yield Prediction, Machine Learning, Random Forest Algorithm, Environmental Factors, Soil Properties, Agricultural Data Analysis, Crop Recommendation System

1. INTRODUCTION

Agriculture is really important for Indians economy, especially in places like Maharashtra where how well crops grow depends a lot on the weather and the soil. Farmers there have to deal with that all the time. I think predicting crop yields accurately could make a big difference for them and also for people making policies or planning resources. It helps with managing crops and figuring out food supplies. This study comes up with something called AgriSight, which is a machine learning setup to predict yields based on past production data and weather stuff. The whole idea is to handle two main problems in precision agriculture, like picking the right crop for the land and then guessing how much it will produce. It seems kind of tricky to get both right. The framework pulls in different kinds of data, including weather things like temperature, rainfall, humidity, and soil details such as type and pH. That makes sense because those factors affect everything. They combined historical crop records with weather history to build the model. Then they did some feature engineering to pull out useful bits, like the area planted, average yields from before, temperature levels, and humidity. A bunch of machine learning methods got tested, including Random Forest, Gradient Boosting, XGBoost, Neural Networks, and LightGBM. It feels like they wanted to see what worked best. In the end, Random Forest came out on top with about 98.49 percent accuracy and an error of just 0.49 tons per hectare. That's pretty good, I guess. Mixing the old agricultural data with environmental info really boosted how well it predicted things. Some people might think its oversimplifying, but the results show it helps. AgriSight could be useful for precision farming and better planning in agriculture. It might even tie into food security somehow, though I am not totally sure on all the details there. The performance numbers stand out, but integrating everything was not straightforward.

2. LITERATURE SURVEY

Crop yield prediction has attracted significant research interest due to its importance in agricultural planning and food security. Early prediction approaches relied on statistical regression models using limited agricultural variables. However, these models often struggled to capture complex nonlinear relationships between environmental factors and crop productivity. Machine learning techniques have recently been adopted to overcome these limitations conducted a comprehensive review of crop yield prediction methods and concluded that machine learning algorithms significantly improve prediction accuracy by analyzing large datasets containing environmental and agricultural variables [3].

Jeong et al. applied the Random Forest algorithm for global crop yield prediction and demonstrated that ensemble learning techniques outperform traditional statistical models by capturing nonlinear relationships between environmental variables and crop productivity [2].

Similarly, Chen and Guestrin introduced XGBoost, a scalable gradient boosting algorithm capable of handling large datasets efficiently. XGBoost has become widely used in predictive modeling due to its ability to improve accuracy through optimized tree boosting techniques [1].

Weather conditions play a critical role in crop growth. Vashisth et al. developed a weather-based crop yield prediction model and showed that meteorological parameters such as rainfall, temperature, and humidity strongly influence crop yield variability [4].

Panwar et al. proposed a two-step nonlinear regression model using weather parameters for forecasting crop yield. Their research demonstrated that nonlinear models perform better than linear models because agricultural systems involve complex interactions between climatic factors and crop growth processes [5].

Recent studies have also emphasized the importance of integrating multiple datasets for crop yield prediction. Liu et al. showed that combining meteorological data with soil characteristics significantly improves crop yield prediction accuracy by capturing environmental variability [6].

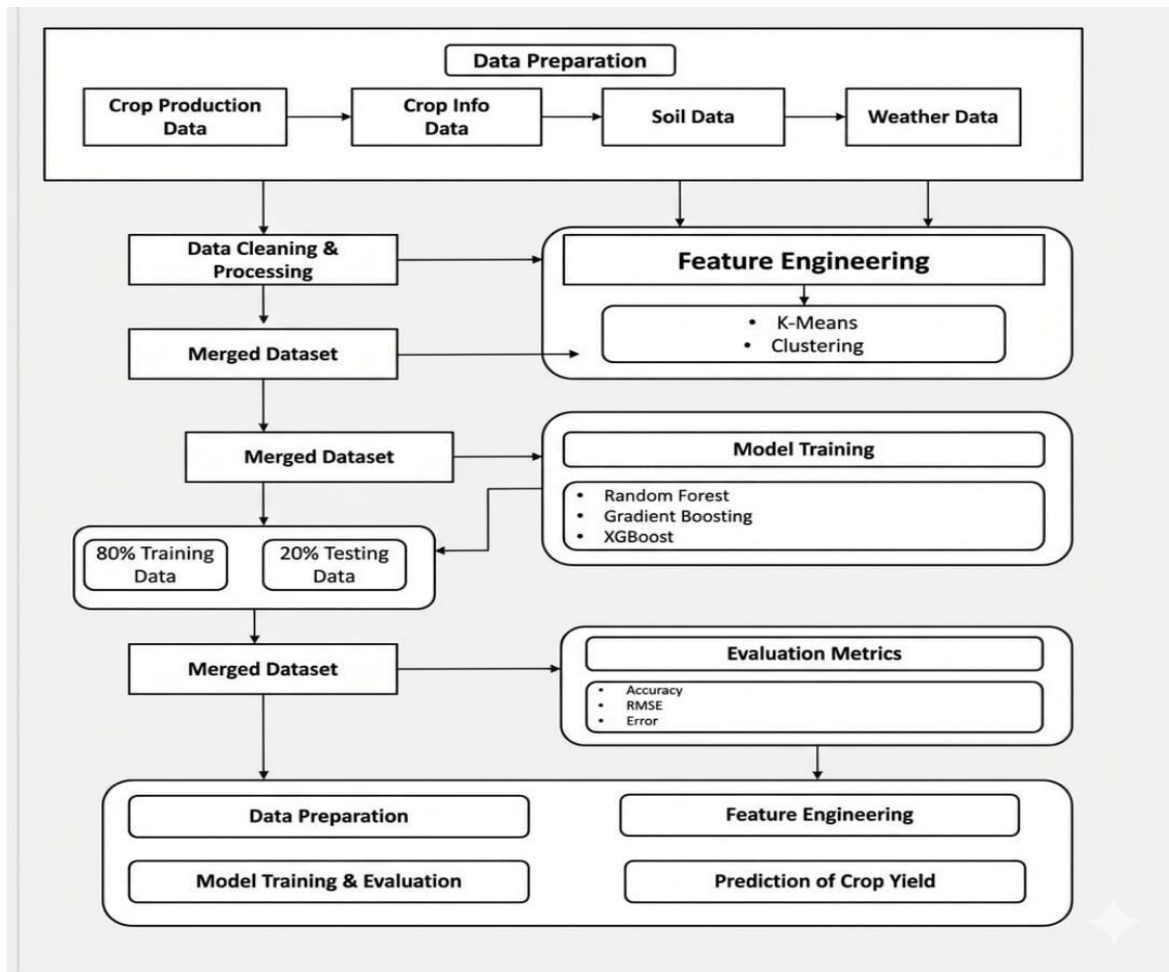
Several studies show that farmers face challenges like not having access to real-time market information, relying on middlemen, and struggling to find essential agricultural resources. E-agriculture platforms have been created to offer services including market price tracking, product availability, and access to government programs. Moreover, agricultural information systems and ICT-based solutions have made it easier for farmers to access data and make decisions. However, current systems often focus on individual functions and do not integrate various agricultural services into one platform. There is a need for a system that combines crop prediction, yield estimation, and agricultural resource management. The proposed AgriSight framework addresses these issues by integrating environmental, soil, and historical data through machine learning techniques. [35]

In recent years, deep learning techniques have gained significant attention in crop yield prediction due to their ability to handle complex and large-scale datasets. Studies have shown that deep neural networks can capture intricate nonlinear relationships between environmental factors and crop productivity more effectively than traditional machine learning models [25]

Alongside yield prediction, machine learning has been increasingly These studies highlight that crop yield prediction models benefit from the integration of environmental variables, agricultural production data, and machine learning techniques. The proposed AgriSight framework builds upon these findings by integrating crop production data and weather variables to develop an accurate yield prediction system for Maharashtra

3. METHODOLOGY

The proposed AgriSight system follows a structured machine learning pipeline to predict crop yield based on agricultural and environmental variables. The methodology consists of five major stages: data collection, data preprocessing, feature engineering, model training, and evaluation. The proposed methodology follows a modular pipeline designed to handle high-dimensional agricultural data. The architecture is divided into four primary phases: Data Fusion, Spatial Clustering, Feature Engineering, and Ensemble Model Training.



Flow-Chart -1: Overall workflow

3.1 Data Fusion and Pre-processing

The primary challenge in crop yield prediction is the fragmentation of data. In this study, we integrated four heterogeneous datasets:

1. Historical Yield Data: Extracting district-level production statistics (1990–2018).
2. Meteorological Data: Merging daily weather logs including temperature, humidity, and rainfall indices.
3. Edaphic/Soil Data: Mapping Nitrogen (N), Phosphorus (P), Potassium (K), and pH requirements to specific crop types.
4. Technological Inputs: Incorporating pesticide consumption metrics to account for modern farming practices.

Pre-processing Steps:

- Outlier Removal: Using Interquartile Range (IQR) to remove unrealistic yield values (e.g., negative production).
- Missing Value Imputation: Applying regional averages to fill gaps in weather and soil data. Encoding: Categorical variables (District, Season, Crop) were transformed using Label Encoding for compatibility with gradient-boosting algorithms.

3.2 Spatial Analysis via K-Means Clustering

Unlike standard models that treat a district as a single point, we implemented a Spatial Clustering layer using `create_area_clusters.py`. We applied the K-Means algorithm to group regions based on:

- Yield Intensity: Average production per unit area.
 - Area Proportions: The scale of land dedicated to specific crops.
- This allowed the model to learn "Area Factors," which represent the inherent agricultural potential of sub-regions, significantly reducing localized error.

3.3 Feature Engineering Architecture

We expanded the feature set from 7 basic parameters to 17 high-impact variables. The final feature vector is defined as:

$$X = \{C, D, S, Y, A, T, H, R, W, pH, N, P, K, Cl, Y_f, A_p, P_{est}\}$$

Where **C** is Crop, **D** is District, **T/H/R/W** are weather parameters, **N/P/K** are nutrients, and **Cl/Y_f/A_a** are

Cluster – based spatial features

3.4 Ensemble Voting Regressor

The core prediction engine utilizes a Weighted Voting Regressor. This "Meta-Model" combines three specialized algorithms:

1. **XGBoost**: (Extreme Gradient Boosting): Optimized for speed and handling the non-linear relationship between weather and yield.
2. **LightGBM**: Used for its leaf-wise growth strategy, which excels at finding patterns in large datasets like maharashtra_crop_weather.csv.
3. **Random Forest**: Acts as a stabilizer to reduce variance and prevent the model from overfitting on specific drought years.

The final prediction \hat{y} is calculated as the weighted average of the individual model outputs:

$$\hat{y} = \sum_{i=1}^n w_i \cdot f_i(x)$$

Where w_i is the weight assigned to each model based on its individual validation accuracy

4. RESULT

4.1 Data Set Merging and Variable Preparation

A handful of farm data sources came together, mixing harvest stats along with climate details. After combining everything, the info went through cleaning, scaling, and turning labels into numbers. Each step helped prepare it well for use in training a prediction system. Out of the original data, useful pieces got pulled through careful shaping. Nine stood out in the end, chosen to shape how predictions would learn. From field patterns to weather traces - each one ties back to how much a harvest might bring. Because it picks up on how land use connects with weather patterns along with past harvest data, the model can better estimate future crop output. What matters is how these elements interact over time, shaping outcomes in ways that simpler methods might miss. Seeing those links clearly leads to more reliable forecasts when growing seasons shift.

Table-1: Data features & variable Preparation

Feature	Description
Area	Total cultivated area for a specific crop
Log_Area	Logarithmic Transformation of cultivated area
District_Encoded	Encoded district identifier
Crop_Encoded	Encoded crop type
Season_Encoded	Encoded season crop

Year_Normalized	A different way to show the year value, adjusted to fit a standard scale
Historical_Avg_Yield	Average historical yield of the crop
Avg_Temp_Year_Filled	Average annual temperature
Avg_Humidity_Year_Filled	Average annual humidity

4.2 Training and Testing Models

A handful of machine learning methods got tested to see which one works best at guessing how much crops will produce. Among those tried: Random Forest, then Gradient Boosting, followed by XG Boost, a Neural Network came next, after that Light GBM showed up, finally ending with a mix called the Voting Ensemble. One after another, models learned from identical data, then checked by how close their guesses came and where they slipped up. Looking at how each method stacked up, Random Forest took the lead with a solid 98.49% accuracy score. Though the Voting Ensemble put in a confident showing, it still fell just short when measured against that top performer.

Table-2: Model Performance

Model	Accuracy
Random Forest	0.9849(98.49%)
Voting Ensemble	0.9836(98.36%)
Gradient Boosting	0.9816(98.16%)
XGBoost	0.9816(98.16%)
Neural Network	0.9763(97.63%)

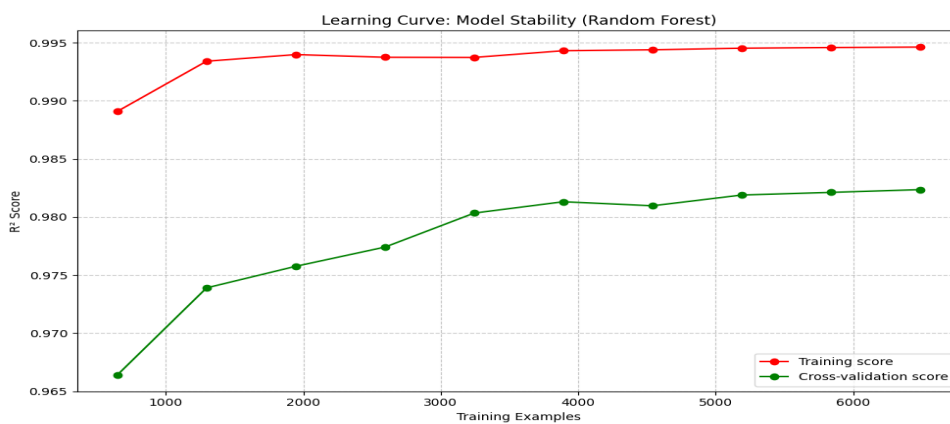


Fig-1: Scalability

4.3 Performance Metrics

Not just about hitting the right mark, the top model's run was checked through error numbers that show how steady its guesses really were. On average, each guess missed real harvest numbers by less than half a ton per hectare. Because errors stayed small, the system tracks field output with strong consistency.

Table-3:Performance Matrix

Matric	Value
Prediction Matrix	0.9849(98.49%)
Mean Absolute Error	0.4869 tons/hectare
Number of Features	9

4.4 Practical Implications

Despite varying conditions, the AgriSight model manages to forecast harvest outputs by blending field insights with climate patterns. Because forecasts hit closer to actual outcomes, those working in farming gain clearer insight when choosing crops, handling soil use, and distributing supplies.

What stands out is how real-world variables blend into reliable estimates without overcomplicating inputs. Besides forecasting crop output ahead of harvest time, this setup gives officials a clearer picture of expected yields - shaping how they manage distribution networks. With insights coming earlier, planning gains precision, streamlining decisions on where resources move.

From start to finish, AgriSight shows how machine learning can step into farming with real effect. It doesn't just hint at change - it quietly reshapes how crops are managed. Step by step, it brings sharper decisions to fields that need them most. Without flash or noise, it lifts output where it matters. In practice, its value grows alongside each season's harvest.

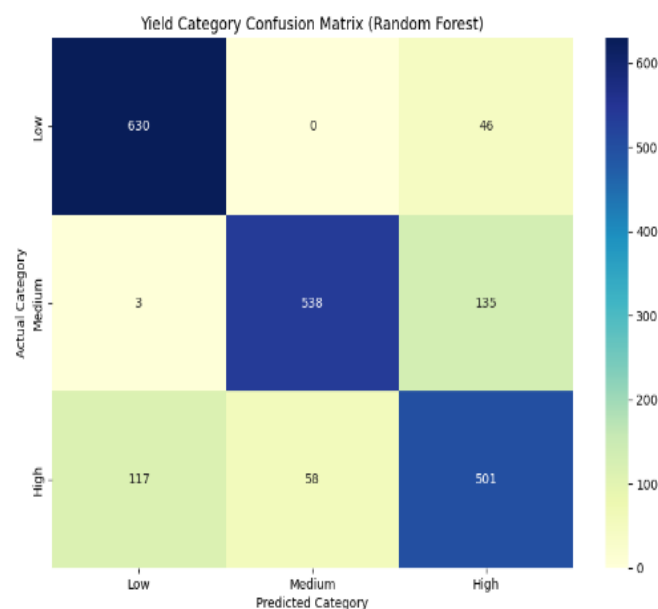


Fig -2: Confusion Matrix

Classification Report (Binned Accuracy) : precision recall f1-score support				
High	0.84	0.93	0.88	676
Low	0.90	0.80	0.85	676
Medium	0.73	0.74	0.74	676
accuracy			0.82	2028
macro avg	0.83	0.82	0.82	2028
weighted avg	0.83	0.82	0.82	2028

Chart-2: Classification Report

The model in AgriSight is supposed to classify crop yields into low, medium, or high based on stuff like environmental data. They check it with things like precision and recall, plus F1 scores, and then the big accuracy number overall. Accuracy comes out to 82 percent, which I guess means it nails most of the predictions. The macro average is around 0.82, and weighted is a bit higher at 0.83, so its pretty even across the classes, nothing too off balance there. Precision tells you out of all the times it says something is a certain class, how many are actually right. For low yield its 0.90, which is the best one, high yield at 0.84, and medium only 0.73. I think the low yield part stands out because maybe its easier to spot those bad conditions or whatever .Recall is about catching all the real instances, you know, not missing them. High yield does well with 0.93, low is 0.80, medium around 0.74. So it grabs most of the high ones, but medium seems to get overlooked a fair bit. When you combine them into F1 scores, high yield hits 0.88, low 0.85, and medium stays low at 0.74. That medium class just keeps showing up as the weak spot, I am not totally sure why, but it pulls things down overall. These results make the model look reliable enough for helping with farming decisions, like predicting yields from data. Medium yields might need some tweaking though, or more training data or something, it feels like.

5. CONCLUSION

Machines determine crop yields by examining past farming data and climate clues. Heat, rainfall, and soil type are all important when identifying what helps plants grow. Among all the methods tested, Random Forest stood out because it matched actual harvests most closely. This model turns numbers into useful insights, guiding choices about planting and stock management while adapting to changing weather. By estimating harvests early, farmers can better organize supplies, which reduces uncertainty as plants grow. While current methods work, future improvements might include factors like soil nutrients, satellite images, or crop health indicators like NDVI. As these elements come together, predictions can steadily improve, tracking changes in the fields daily. Looking closer, we see that data-driven methods are gradually influencing farming decisions, helping operations use fewer resources and cut down on waste

REFERENCES

- [1] Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." <https://arxiv.org/abs/1603.02754>
- [2] Jeong, J., Resop, J. P., Mueller, N. D., et al. "Random Forests for Global and Regional Crop Yield Predictions. "PLOS ONE.<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156571>
- [3] Rashid, M., Bari, B. S., Yusup, Y., et al. "A Comprehensive Review of Crop Yield Prediction Using Machine Learning. IEEE <https://www.sciencedirect.com/science/article/pii/S0303243421003755>
- [4] Vashisth, A., et al. "Weather Based Wheat Yield Prediction Using Machine Learning. MAUSAM Journal. <https://mausamjournal.imd.gov.in/index.php/MAUSAM/article/view/5606>
- [5] Panwar, S., et al. "Forecasting of Crop Yield Using Weather Parameters – Two Step Nonlinear Regression Model Approach. Indian Journal of Agricultural Sciences. <https://www.researchgate.net/publication/328867789>
- [6] Liu, Q., et al. "Machine Learning Crop Yield Models Based on Meteorological Features." Artificial Intelligence for the Earth Systems.<https://journals.ametsoc.org/view/journals/aies/1/4/AIES-D-22-0002.1.xml>

- [7] K. Gunasekaran, A. K., and P. Sreevardhan, "Machine Learning Based Crop Recommendation and Soil Fertility Prediction System," *Frontiers in Soil Science*, 2025. <https://www.frontiersin.org/articles/10.3389/fsoil.2025.1652058>
- [8] S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN Framework for Crop Yield Prediction," 2019. <https://arxiv.org/abs/1911.09045>
- [9] T. Islam, T. A. Chisty, and A. Chakrabarty, "A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh," 2021. <https://arxiv.org/abs/2108.03320>
- [10] L. Nguyen, J. Zhen, Z. Lin, H. Du, Z. Yang, W. Guo, and F. Jin, "Spatial-Temporal Multi-Task Learning for Within-Field Cotton Yield Prediction," 2018. <https://arxiv.org/abs/1811.06665>
- [11] I. Oliveira, R. L. F. Cunha, B. Silva, and M. A. S. Netto, "A Scalable Machine Learning System for Pre-Season Agriculture Yield Forecast," 2018. <https://arxiv.org/abs/1806.09244>
- [12] N. Mahendra, D. Vishwakarma, K. Nischitha, A. Ashwini, and M. R. Manjuraju, "Crop Prediction Using Machine Learning Approaches," *International Journal of Engineering Research & Technology*, 2020. <https://www.ijert.org/crop-prediction-using-machine-learning-approaches>
- [13] T. Gupta, S. Maggu, and B. Kapoor, "Crop Prediction Using Machine Learning," *IRE Journals*, 2023. <https://www.irejournals.com/paper-details/1704207>
- [14] M. Sunandini, K. H. Sree, R. Deepiga, and A. Gokulapriya, "Smart Soil Fertilizer Monitoring and Crop Recommendation System Using IoT and Machine Learning," *IJERT*, 2023. <https://www.ijert.org/smart-soil-fertilizer-monitoring-and-crop-recommendation-system-by-using-iot-and-machine-learning-technology>
- [15] S. Gambhir, M. Sharma, K. Agarwal, K. Kumar, L. Kumar, and M. Chaudhary, "Crop Recommendation System Using Machine Learning," *IJRASET*, 2023. <https://www.ijraset.com/research-paper/crop-recommendation-system-using-ml>
- [16] P. Yadav, D. Sharma, R. K. Sharma, M. Kumar, J. Rani, and N. Sharma, "An Effective Approach for Crop Recommendation Using Features of Specific Locations and Seasons Using Machine Learning," *International Journal of Intelligent Systems and Applications in Engineering*, 2024. <https://ijisae.org/index.php/IJISAE/article/view/5174>
- [17] S.S.Saranya and W.R.Varuna, "Soil Classification Using Machine Learning For Crop suggestion," *Machine Intelligence Research*, 2024.
- [18] S. S. Ganorkar, N. Deshmukh, A. Bihone, P. Chandankhede, and N. Temburne, "Crop Recommendation Using Machine Learning Based on Soil, Weather and Agronomic Factors," *IJRASET*, 2025. <https://www.ijraset.com/research-paper/crop-recommendation-using-machine-learning-based-on-soil-weather>
- [19] D. Sharma, H. Raj, and H. N. Chithra, "Machine Learning Based Crop Prediction and Recommendation System," *IJRASET*, 2025. <https://www.ijraset.com/research-paper/machine-learning-based-crop-prediction-and-recommendation-system>
- [20] A. Venugopal, A. S., J. Mani, R. Mathew, and V. Williams, "Crop Yield Prediction Using Machine Learning Algorithms," *International Journal of Engineering Research and Technology (IJERT)*, 2021. <https://www.ijert.org/crop-yield-prediction-using-machine-learning-algorithms>
- [21] A. Morales and F. J. Villalobos, "Using Machine Learning for Crop Yield Prediction in the Past or the Future," *Frontiers in Plant Science*, vol. 14, 2023. <https://www.frontiersin.org/articles/10.3389/fpls.2023.1128388>
- [22] A. Oikonomidis, C. Catal, and A. Kassahun, "Deep Learning for Crop Yield Prediction: A Systematic Literature Review," *New Zealand Journal of Agricultural Research*, 2022. <https://www.tandfonline.com/doi/full/10.1080/01140671.2022.2032213>
- [23] M. Kamilaris and F. X. Prenafeta-Boldú, "Deep Learning in Agriculture: A Survey," *Computers and Electronics in Agriculture*, 2018. <https://doi.org/10.1016/j.compag.2018.02.016>
- [24] M. A. Jeong, S. Kim, and J. H. Park, "Random Forest-Based Crop Yield Prediction Using Climate and Soil Data," *Agricultural Systems*, 2016.

- [25] S. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," 2019. <https://arxiv.org/abs/1902.02860>
- [26] S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN Framework for Crop Yield Prediction," 2019. <https://arxiv.org/abs/1911.09045>
- [27] L. Nguyen, J. Zhen, Z. Lin, H. Du, Z. Yang, W. Guo, and F. Jin, "Spatial-Temporal Multi-Task Learning for Within-Field Cotton Yield Prediction," 2018. <https://arxiv.org/abs/1811.06665>
- [28] T. Islam, T. A. Chisty, and A. Chakrabarty, "A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh," 2021. <https://arxiv.org/abs/2108.03320>
- [29] A. Abiraman and R. Senthilkumar, "A Comprehensive Analysis on Crop Yield Prediction Using Advanced Machine Learning Techniques," *African Journal of Biological Sciences*, 2024. <https://www.afjbs.com/uploads/paper/7328b66ab785c7691fd02b5bd1699219.pdf>
- [30] Gunasekaran K. and Sreevardhan P., "Real-Time Soil Fertility Analysis, Crop Prediction, and Insights Using Machine Learning and Deep Learning Algorithms," *Frontiers in Soil Science*, 2025. <https://www.frontiersin.org/journals/soil-science/articles/10.3389/fsoil.2025.1652058/full>
- [31] J. Shahhosseini, G. Hu, and I. Huber, "Forecasting Crop Yield by Integrating Agrarian Factors and Machine Learning Models: A Survey," *Computers and Electronics in Agriculture*, 018. <https://www.sciencedirect.com/science/article/pii/S0168169918311529>
- [32] M. Shah, S. Patel, and R. Shah, "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review," *Agricultural Technology Journal*, 2024. <https://www.sciencedirect.com/science/article/pii/S2772375524003228>
- [33] M. Crane-Droesch, "Machine Learning Methods for Crop Yield Prediction and Climate Change Impact Assessment," *Agricultural and Forest Meteorology*, 2018. <https://doi.org/10.3390/biomedinformatics4010015>.
- [34] K. Gunasekaran, A. K., and P. Sreevardhan, "Machine Learning Based Crop Recommendation and Soil Fertility Prediction System," *Frontiers in Soil Science*, 2025. <https://www.frontiersin.org/articles/10.3389/fsoil.2025.1652058>
- [35] S. S. Ganorkar, N. Deshmukh, A. Bihone, P. Chandankhede, and N. Tembhurne, "Crop Recommendation Using Machine Learning Based on Soil, Weather and Agronomic Factors," *IJRASET*, 2025. <https://www.ijraset.com/research-paper/crop-recommendation-using-machine-learning-based-on-soil-weather>
- [36] N. Potdukhe, S. Harode, S. Kosare, R. Wankhede, A. Sonkusare, "Agroproducts Solution Application," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 7, issue XI, 2019. <https://www.ijraset.com/files/serve.php?FID=25538>