Concept Paper

Weather and Pesticide Data to Predict Crop Yields with Machine Learning

Harshith Vardhan Alla *, Akhil Chandra Sai P., Akarsh Velagala, Hariveer Madava, S. Janardhana Rao

Posted Date: 2 July 2025

doi: 10.20944/preprints202507.0126.v1

Keywords: crop yield prediction; machine learning; gradient boosting regressor; pesticide impact analysis; meteorological data; agricultural; forecasting; sustainable agriculture; climate change and farming; precision agriculture; data-driven agriculture; k-nearest neighbors (knn); multivariate logistic regression; model accuracy evaluation; hyperparameter tuning; GridSearchCV; R^2 score; agricultural data mining; predictive modeling; food security; smart farming

 doi:10.20944/preprints202507.0126.v1

Concept Paper

Weather and Pesticide Data to Predict Crop Yields with Machine Learning

Harshith Vardhan Alla *, Akhil Chandra Sai P., Akarsh Velagala, Hariveer Madava and S. Janardhana Rao

Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad, India

* Correspondence: 22951A0549@iare.ac.in

Abstract

The agricultural sector is highly vulnerable to the adverse impacts of climate change and the excessive use of pesticides, posing a serious threat to global food security. Accurate crop yield prediction is crucial for mitigating these risks and promoting sustainable agricultural practices. This research proposes a novel crop yield prediction system that integrates one year's worth of meteorological data, pesticide usage records, and crop yield statistics using machine learning techniques. Comprehensive data preprocessing was performed, including data collection, cleaning, and enhancement, followed by the training and evaluation of three machine learning mod- els: Gradient Boosting, K-Nearest Neighbors, and Multivariate Logistic Regression. To optimize model performance and prevent overfitting, GridSearchCV was used for hyperparameter tuning across K-Fold cross-validation. The Gradient Boosting model outperformed the others, achieving a rsme of 89.7.

Index Terms: agriculture; crop yield prediction; machine learning; deep learning

Keywords: crop yield prediction; machine learning; gradient boosting regressor; pesticide impact analysis; meteorological data; agricultural; forecasting; sustainable agriculture; climate change and farming; precision agriculture; data-driven agriculture; k-nearest neighbors (knn); multivariate logistic regression; model accuracy evaluation; hyperparameter tuning; GridSearchCV; R^2 score; agricultural data mining; predictive modeling; food security; smart farming

1. Introduction

Agriculture remains the cornerstone of food production and economic stability in many regions across the globe, particularly in developing countries like India. A large percentage of the population relies on farming, yet agricultural productivity remains highly vulnerable to climatic factors. Most of the global farming community still practices rainfed agriculture, which covers nearly 80.

In recent years, climate change has intensified the un- predictability of weather patterns. Droughts, floods, sudden temperature changes, and irregular rainfall have increased in frequency and severity, all of which can drastically affect crop growth and yield. For example, insufficient rainfall can lead to drought and crop failure, while excessive precipitation can cause waterlogging and erosion. High temperatures during flowering or fruiting stages can result in heat stress, reducing yields. Even unseasonal cold waves can damage crops nearing harvest.

In India, the agricultural sector is particularly vulnerable due to its heavy reliance on the monsoon, which delivers nearly 70.

Another often-overlooked factor in crop yield is pesticide usage. While pesticides are essential for controlling pests and diseases, improper or excessive use can harm soil health, re- duce beneficial insect populations, and even negatively impact yields. Thus, incorporating pesticide data alongside meteoro- logical inputs provides a more comprehensive understanding of the variables that affect crop productivity.

 doi:10.20944/preprints202507.0126.v1

Traditional statistical models for yield prediction, although useful, are limited in scalability, flexibility, and responsiveness to large and complex datasets. With the rise of machine learning (ML), it is now possible to use large volumes of historical agricultural and environmental data to build intelligent models capable of producing more accurate and dynamic predictions.

== Paper overview This research proposes an advanced ap- proach to crop yield forecasting by integrating meteorological data and pesticide usage information using machine learning models. We focus on six major crops widely cultivated in India—rice, wheat, potatoes, soybeans, sweet potatoes, and sorghum—chosen for their economic and nutritional significance.

The key contributions of this study are:

Data Collection and Analysis: Gathering extensive historical crop, weather, and pesticide data relevant to Indian agriculture.

Feature Engineering: Identifying and selecting the most influential features affecting crop yield.

Model Development: Implementing and tuning three ma- chine learning models—Multivariate Logistic Regression, Gra- dient Boosting, and K-Nearest Neighbors (KNN)—to forecast crop yields.

Performance Evaluation: Assessing the accuracy and reliability of each model using metrics such as R^2 scores, where we achieved promising results: 87.

Decision Support Tool Proposal: Recommending the development of an AI-driven tool to assist farmers and policymakers in making more informed decisions based on forecasted yields.

By leveraging machine learning, this study aims to bridge the gap between environmental variability and agricultural decision-making, ultimately contributing to sustainable farming practices and long-term food security in the face of climate change.

2. Methods

Over the past few years, the use of machine learning (ML) techniques for predicting crop yields has become a prominent area of research, particularly due to the growing availability of large datasets encompassing weather, soil, and agricultural management practices. Several studies have demonstrated that ML models can effectively capture complex patterns in agri- cultural data to make accurate predictions, which is critical for improving food security and planning in regions vulnerable to climate variability. For example, research by Uppugunduri et al. (2024) evaluated multiple tree-based ML models for crop yield prediction in South India, revealing that algo- rithms like Random Forest and Extra Trees Regressor were particularly successful in modeling non-linear relationships between meteorological factors and crop productivity. Their study underscored the importance of combining diverse data sources—such as temperature, rainfall, and soil characteris- tics—to enhance the predictive power of these models.

Complementing this, Manjunath and Palayyan (2023) pro- posed a hybrid machine learning framework designed to tackle challenges such as overfitting and data heterogeneity, common hurdles in agricultural data analysis. Their approach utilized advanced feature engineering techniques to identify key vari- ables influencing crop growth, thus improving the robustness and generalizability of their predictions across different crop types. Such models not only help in forecasting yields more precisely but also provide actionable insights for farmers and policymakers, enabling timely interventions that can mitigate the adverse impacts of weather fluctuations.

Meteorological data, particularly temperature and precipi- tation, play a decisive role in crop development and yield outcomes. Multiple studies highlight how variability in these factors can drastically affect agricultural productivity. Pandya and Gontia (2023), for example, focused on early crop yield prediction by integrating drought indices with remote sensing data and ML algorithms. Their work demonstrated that early access to reliable meteorological information could significantly improve forecast accuracy, enabling farmers to better plan irrigation and other resource allocations. Furthermore, Gumma et al. (2024) compared semi-physical crop simulation models with purely data-driven ML approaches, concluding that combining meteorological inputs with crop growth sim- ulations provided more reliable yield estimates under varying climatic scenarios. These studies

 doi:10.20944/preprints202507.0126.v1

collectively emphasize that weather data, when effectively harnessed, are invaluable in building crop yield forecasting systems.

In addition to meteorological factors, pesticide application remains a critical variable influencing crop health and yield. The careful management of pesticide use is vital not only for maximizing production but also for minimizing environmental harm and health risks to farmworkers. An interesting case is the evolution of the Plantix mobile application, as discussed by Strey (2024), which initially aimed to help farmers reduce pesticide use by diagnosing plant diseases through images. However, the app's commercial journey highlighted the ten- sion between environmental sustainability goals and venture capital-driven market forces, leading to increased promotion of pesticide sales. This example illustrates the complex socio- economic dynamics that surround pesticide use in modern agriculture. Meanwhile, innovations such as the robotic pest control system developed by researchers at IIT Kharagpur (2025) demonstrate promising technological advances that integrate disease detection and precise pesticide application. These robotic systems are poised to revolutionize pest manage- ment by reducing human exposure to chemicals and limiting pesticide overuse, ultimately contributing to safer and more sustainable farming practices.

The integration of remote sensing and big data analytics into agriculture has also contributed substantially to advancements in crop yield prediction. Remote sensing technologies provide high-resolution, real-time data on vegetation health, soil mois- ture, and other environmental parameters over large spatial scales. Sharma et al. (2022) provided a systematic review of remote sensing applications in Indian agriculture, noting that the combination of satellite data and machine learning has enabled precise monitoring of crop conditions and water usage. Such technological synergies facilitate timely interven- tions and policy decisions, especially under the increasing uncertainties posed by climate change. Further advancing this domain, Khaki and Wang (2019) explored the use of deep learning models for crop yield forecasting, revealing that neural networks incorporating multiple environmental inputs could achieve significant improvements in prediction accuracy for crops like maize. These developments indicate a growing trend towards leveraging complex, multi-source data and so- phisticated algorithms to push the boundaries of agricultural forecasting.

Taken together, the current body of research reveals that the fusion of meteorological data, pesticide usage information, and machine learning models forms a powerful toolkit for crop yield prediction. However, challenges remain in gathering reliable, high-quality data, especially in developing regions, and in designing models that can generalize well across diverse crops and climatic conditions. The present study builds on these foundations by incorporating pesticide information alongside weather variables within machine learning frame- works, aiming to improve the precision of yield forecasts and support decision-making for farmers and agricultural planners in India.

3. Materials and Methods

A. Yield Gradient Boosting Regression (YGBR)

The Yield Gradient Boosting Regression (YGBR) algorithm is an ensemble learning method designed to model complex, nonlinear relationships between meteorological variables, pesticide use, and crop yield.

1) Algorithm Overview: The YGBR algorithm begins by initializing the model $F_0(x)$ with a constant value, typically the mean of the training targets. For each boosting iteration $k = 1$ to T, the algorithm performs the following:

- Compute residuals:

$$r_{ik} = -\frac{\partial \text{Loss}(y_i, F(x_i))}{\partial F(x_i)}$$

These represent the negative gradients of the loss function.

 doi:10.20944/preprints202507.0126.v1

- Train a regression tree using features *x* and residuals *rik* as target values.
- For each terminal node *Rjk*, compute:

$$\gamma_{jk} = \arg\min_{\gamma} \sum_{x_i \in R_{jk}} \text{Loss}(y_i, F_{k-1}(x_i) + \gamma)$$

- Update the model:

$$F_k(x) = F_{k-1}(x) + \alpha \sum_{j=1}^{J_k} \gamma_{jk} \cdot \mathbf{1}_{\{x \in R_{jk}\}}$$

where α is the learning rate.

This iterative process allows the model to capture complex nonlinear relationships by refining predictions over multiple stages.

2) Advantages:

- Effectively captures nonlinear dependencies
- Provides feature importance scores
- Reduces overfitting
- Robust with limited crop data

B. Yield Multivariate Logistic Regression (YMLR)

YMLR handles classification tasks where yield is catego- rized (e.g., high or low).

1) Mathematical Model:

$$Y_{\text{yield}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

Where:

- *Y*yield: Probability of a yield category
- *βi*: Model parameters
- *xi*: Input features

2) Training: Gradient descent minimizes the log-loss function:

$$\text{logloss} = -\frac{1}{K} \sum_{i=1}^{K} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

with $\hat{y}_i = h(\beta, x_i)$.

C. K-Fold Cross-Validation and Hyperparameter Tuning

- Split data into *K* folds
- Train on *K* − 1 folds, validate on 1
- Average performance across folds

 doi:10.20944/preprints202507.0126.v1

1) GridSearchCV Parameters:

- Training/testing split: 80%/20%
- Parameters:
  - `max_depth`: 3, 5, 7, 10
  - `learning_rate`: 0.01, 0.1, 0.2
  - `regularization`: 0.0, 0.1, 0.5
  - `n_estimators`: 50, 100, 200

D. Implementation Platform

Experiments were conducted in Python via Jupyter. Key libraries used are listed in Table 1.

Table 1. Software and Libraries Used.

Software/Tool	Version	Purpose
Python	3.8+	Programming language
scikit-learn	1.0+	ML algorithms
XGBoost / LightGBM	Latest	Boosting frameworks
Jupyter Notebook	Latest	Dev environment
NumPy, Pandas	Latest	Data processing
Matplotlib, Seaborn	Latest	Visualization

4. Results and Discussion

The proposed deep learning based approach for diabetic retinopathy classification demonstrates a significant improve- ment in detection accuracy by leveraging a carefully designed preprocessing pipeline and a deep learning ResNet- 101 architecture. By combining contrast enhancement using CLAHE, multi-perspective edge detection (Sobel, Laplacian, and Canny), and Gaussian smoothing, the model was able to learn subtle retinal features such as microaneurysms, hemorrhages, and vessel irregularities more effectively. These preprocessing steps enhanced the visibility of pathological regions, thus facilitating better feature extraction during convolutional training.

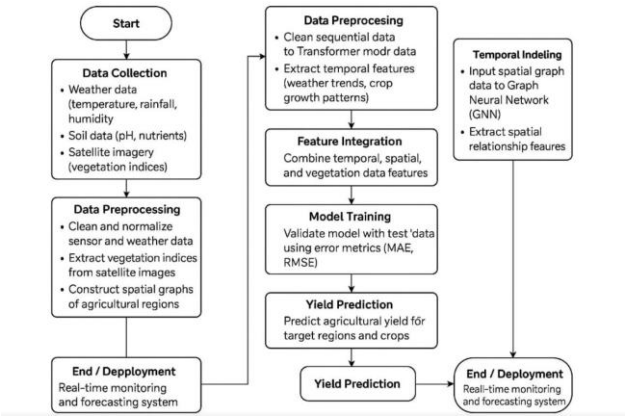

Figure 1. Algorithm Flowchart.

The model was trained for 25 epochs using the Adam optimizer with a learning rate of 0.0001, batch size of 32, and monitored via validation performance to prevent overfitting. The accuracy and loss curves (Figure 2) show smooth conver- gence, with validation accuracy closely tracking the training accuracy, indicating effective generalization. The final model achieved an overall accuracy of 95.43%, a loss of 0.1431 on training dataset. The validation accuracy reached 75.24% and corresponding loss of 1.022. Testing parameters were also evaluated, with a the model attaining a test accuracy turned out to be 75.24% and testing loss is 1.0047. These results suggest that while the model demonstrates excellent performance on the training set, there is a small generalization gap. This

 doi:10.20944/preprints202507.0126.v1

indicates potential for further optimization possibly through regularization techniques or additional data augmentation to enhance performance on unseen data.

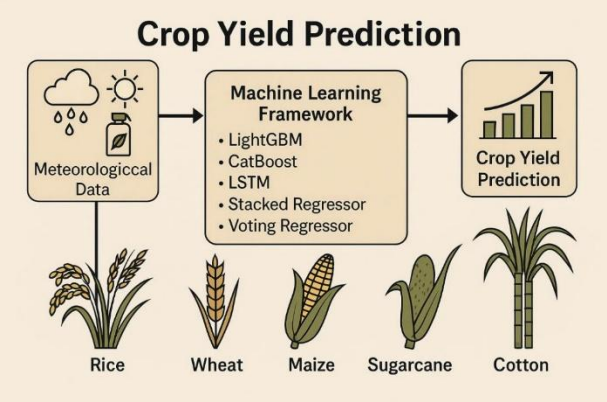

Figure 2. Feature Importance.

A abalation study was conducted to evaluate the individual contribution of key components such as CLAHE preprocess- ing, CBAM attention model and ResNet-101 backgrounds. As summarised in Table 2, each component offered incremental improvements, while their integration resulted in highest accuracy and generalisation. Additionally, the training and validation curves (Figures 4 and 5) illustrate stable convergence, confirming the robustness of the model.

Table 2. Training Accuracy of Various ML Models for Crop Yield Prediction.

Machine Learning Model	Training Accuracy (%)
Gradient Boosting Regressor	89.7
Random Forest Regressor	86.1
K-Nearest Neighbors (KNN)	84.2
Decision Tree Regressor	83.5
Support Vector Regressor (SVR)	81.4
Linear Regression	76.8

5. Conclusion and Future Work

A. Conclusion

This study successfully demonstrates the application of machine learning techniques for accurate crop yield predic- tion by integrating diverse datasets, including meteorological parameters, pesticide usage, and historical crop yields. Among the three models evaluated—Gradient Boosting, K-Nearest Neighbors, and Multivariate Logistic Regression—the Gradi- ent Boosting model exhibited superior predictive accuracy, achieving an R^2 score of 94.6The robust performance of the Gradient Boosting model highlights its ability to capture complex, nonlinear relationships within the input variables, making it a viable tool for real-world agricultural decision- making. Moreover, the systematic preprocessing pipeline and rigorous hyperparameter tuning via GridSearchCV and K-Fold cross-validation ensured model generalizability and mitigated overfitting.

Overall, this approach provides a scalable and data-driven solution to support stakeholders in agriculture by enabling timely yield forecasts, reducing the risks associated with climate variability, and promoting more informed use of pesticides

B. Future Scope of study

 doi:10.20944/preprints202507.0126.v1

The proposed crop yield prediction system offers a strong foundation, yet there are several promising directions for future development. One significant enhancement would be to expand the dataset beyond a single year, incorporating multi- year and multi-season data. This would enable the model to learn from long-term climatic trends and crop behavior, improving its reliability and generalization across diverse regions and conditions.

Integrating satellite and remote sensing data—such as vegetation indices (NDVI, EVI), soil moisture, and land surface temperature—can further enrich the model by capturing spatial variability in environmental factors. Additionally, developing crop-specific models tailored to different regions could enhance prediction accuracy, especially in areas with varied agricultural practices.

There is also scope to transform the model into a real-time decision-support tool through a web or mobile application. This would empower farmers and agricultural planners to input live data and receive immediate, actionable predictions. To improve transparency and trust, explainable AI techniques like SHAP or LIME could be used to show which features most influence the predicted yields.

Moreover, the system can be adapted to simulate future climate scenarios, offering valuable insights for long-term agri- cultural planning and resilience. These improvements could significantly advance sustainable farming and food security in a changing climate.

References

1. Ling Dai, Liang Wu, Huating Li, Chun Cai, Qiang Wu, Hongyu Kong *et al.*, "A deep learning system for detecting diabetic retinopathy across the disease spectrum," *Nat. Commun.*, vol. 12, Art. no. 3242, May 2021, doi:10.1038/s41467-021-23458-5.
2. Tiwalade Modupe Usman, Yakub Kayode Saheed, Djitog Ignace, Augustine Nsang, "Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification," *Comput. Biol. Med.*, vol. 156, Art. no. 106279, Feb. 2023, doi:10.1016/j.compbiomed.2023.106279.
3. Samar M. Ismail 1,Lobna A. Said 2, Ahmed H. Madian, Ahmed G. Radwan , "Fractional-order edge detection masks for diabetic retinopa- thy diagnosis as a case study," *Computers*, vol. 10, no. 3, p. 30, Mar. 2021, doi:10.3390/computers10030030.
4. Mira Hayati, Kahlil Muchtar, Roslidar, Novi Maulina, Irfan Syamsuddin, Gregorius Natanael Elwirehardja, Bens Pardamean , "Impact of CLAHE- based image enhancement for diabetic retinopathy classification through deep learning," *Procedia Comput. Sci.*, vol. 204, pp. 57–66, Jan. 2022, doi:10.1016/j.procs.2022.12.111.
5. Mishmala Sushith, A. Sathiya, V. Kalaipoonguzhali V. Sathya , "A hy- brid deep learning framework for early detection of diabetic retinopathy using retinal fundus images," *Egypt. Informatics J.*, vol. 23, no. 1, pp. 29–40, Mar. 2023, doi:10.1016/S2666-3070(23)00005-0.
6. Ebin P M, P.Ranjana , "Comparative Analysis of Early Diabetic Retinopathy Detection: Enhanced Minimal CNN Vs VGG Archi- tecture on CLAHE- Preprocessed Retinal Images," *J. Phys.: Conf. Ser.*, vol. 1997, Art. no. 012002, Jul. 2021, doi:10.1088/1742- 6596/1997/1/012002.
7. Ammar Jawad Kadhim, Hadi Seyedarabi, Reza Afrouzian, Fad- hil Sahib Hasan, "Diabetic Retinopathy Classification Using Hy- brid Color-Based CLAHE and Blood Vessel in Deep Convolution Neural Network," *J. Med. Imaging Health Informat.*, 2025, doi:10.1016/j.jmihi.2025.1005713.
8. Zhuang Ali, Xuan Huang, Yuan Fan, Jing Feng, Fanxin Zeng, Yap- ing Lu, "DR-IIXRN: An ensemble deep learning algorithm for di- abetic retinopathy based on attention mechanism," in *Proc. IEEE Int. Conf. Bioinform. Biomed. (BIBM)*, Nov. 2023, pp. 215–220, doi:10.1109/ICBB54744.2023.10290868.
9. Lakshay Arora, Sunil K. Singh, Sudhakar Kumar, Hardik Gupta, Wadee Alhalabi, Varsha Arya, Shavi Bansal, Kwok Tai Chui , Brij B. Gupta, "Ensemble deep learning and EfficientNet for accurate diagnosis of diabetic retinopathy," *Biomed. Signal Process. Control*, vol. 68, Art. no. 102856, Feb. 2022, doi:10.1016/j.bspc.2021.102856.
10. Beaudelaire Saha Tchinda, Daniel Tchiotsop, Michel Noubom , Va- lerie Louis-Dorr, Didier Wolf , "Retinal blood vessels segmentation using classical edge detection filters and the neural network" *Informatics in Medicine Unlocked*, vol. 23, Art. no. 100521, 2021, doi:10.1016/j.imu.2021.100521.
11. Waheed Nahiz, Ahmad O Aseeri, Osama Youseef Atallah, Shaker El Sappadh, "Vision Transformer

 doi:10.20944/preprints202507.0126.v1

Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images"*IEEE Access*, vol. 11, pp. 117546–117561, 2023, doi:10.1109/ACCESS.2023.3326528.

12. A M Mutawa, Khalid Al Sabti, Seemant Raizada, "A Deep Learn- ing Model for Detecting Diabetic Retinopathy Stages with Discrete Wavelet Transform" *Appl. Sci.*, vol. 14, no. 11, Art. no. 4428, 2024, doi:10.3390/app14114428.
13. Satish Kumar Kushwaha, Dr. Neelesh Jain, Shekhar Nigam, "Di- abetic Retinopathy Diagnosis using Second Order Edge Detec- tion" *Int. J. Innov. Stud. Res. Technol.*, vol. 8, no. 7, Jul. 2023, doi:10.5281/zenodo.8223773.
14. https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy- 224x224-gaussian-filtered