iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67417

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Crop Yield Prediction using Machine Learning

Mrs. Ashwini Yacha[1], Hari Varun Nakkalaganti[2], Rakshith Kanchi[3], Sai Mani Prasad Pasala[4], Shravan Goud Borapatla[5]
[1]*Assistant Professor, Computer Science & Business System, B V Raju Institute of Technology*
[2, 3, 4, 5]*Student, Computer Science & Business System, B V Raju Institute of Technology*

Abstract: ***Agriculture is the backbone of many economies, playing a crucial role in food security and rural livelihoods. However, predicting crop yields accurately remains a significant challenge due to factors like fluctuating weather conditions, soil variability, and changing farming practices. Traditionally, yield estimation has depended on historical trends, farmer experience, and generalized climate models, which often fail to capture real-time complexities. This study introduces a machine learning-driven approach to enhance crop yield prediction by leveraging information from several sources, including past yield figures, soil characteristics and climatic records. Utilizing advanced algorithms including regression models, decision trees, and ensemble learning methods, we aim to improve forecasting accuracy and offer actionable insights to farmers. The proposed system helps in optimizing resource allocation, reducing agricultural uncertainties, and enabling better decision-making for sustainable farming. Our results indicate that machine learning models outperform traditional estimation techniques, delivering more accurate and flexible forecasts. This research contributes to the growing field of precision agriculture by integrating data science into farming practices, ultimately leading to increased productivity and efficiency in the agricultural sector.***

Keywords: Crop Yield Prediction, Machine Learning, Random Forest Regressor, Precision Agriculture, Data-Driven Farming, Ensemble Learning, Feature Importance, Climate Impact on Agriculture, Pesticide and Rainfall Analysis

I. INTRODUCTION

Agriculture is a fundamental sector that sustains global food production and economic stability. With growing demand for food due to population expansion, ensuring efficient and high-yield crop production has become a top priority. However, crop yield is influenced by multiple factors, including climate conditions, soil quality, irrigation levels, pest infestations, and farming techniques. Traditional yield estimation methods, which rely on historical data and expert judgment, often lack precision and fail to account for sudden environmental changes. With advancements in technologies, machine learning (ML) has become a potent instrument to improve agricultural predictions. ML models are able to examine a wide dataset from different sources, such as weather predictions, soil composition and past crop yield for detecting patterns & provide accurate yield predictions. By leveraging data-driven insights, farmers and policymakers able to make well-informed choices on crop selection, risk management & resource allocation, leading to improved productivity and sustainability. The purpose of this study is to create a machine learning based crop yield prediction system that integrates real-time and historical data to improve forecasting accuracy. The proposed model utilizes various ML algorithms, including regression techniques and decision trees, to analyze key agricultural parameters. By implementing this approach, we seek to bridge the gap between traditional farming methods and modern data-driven solutions, ultimately contributing to the advancement of precision agriculture.

II. RELATED WORK

1) *Traditional Crop Yield Prediction Models*: Early agricultural yield prediction models relied on statistical approaches, particularly multiple linear regression (MLR) and time series forecasting. Lobell et al. [1] investigated the relation between crop yield & climate factors using MLR models, demonstrating that temperature and precipitation significantly impact production. However, these models often fail to capture non-linear interactions between variables, leading to suboptimal predictions.
2) *Machine Learning in Agricultural Yield Prediction*: With advancements in machine learning (ML), researchers have developed more accurate prediction models. Random Forest (RF), Support Vector Regression (SVR) & Artificial Neural Networks (ANNs) have shown superior performance compared to traditional models [2]. For instance, Sharma et al. [3] applied RF and Gradient Boosting to wheat yield prediction and achieved higher accuracy than MLR. Similarly, Zhou et al. [4] demonstrated that XGBoost-based models outperformed conventional methods by minimizing error rates.
3) *Deep Learning in Precision Agriculture*: Deep Learning methods include Convolutional Neural Networks (CNNs) & Long Short-Term Memory (LSTM) networks, have recently been employed for yield prediction. Mandal et al. [5] showed that LSTMs effectively capture temporal dependencies in yield fluctuations, leading to improved predictive accuracy. Other studies, such as those by Ramesh et al. [6], integrated satellite imagery and CNNs to enhance real-time crop monitoring and prediction accuracy.

4) *Impact of Climatic and Environmental Factors*: Incorporating climatic and environmental factors, such as soil quality, rainfall variability, and temperature fluctuations, has been critical in improving yield predictions. Ghosh et al. [7] found that integrating remote sensing data and meteorological parameters improved yield forecasts by 25%. Furthermore, Singh et al. [8] analyzed pesticide usage trends and their effect on agricultural productivity, highlighting the need for sustainable farming practices.
5) *IoT and Smart Farming for Yield Optimization*: As real-time data collection, predictive analytics, and Internet of Things (IoT)-based precision farming have advanced, they have become crucial. Smith et al. [9] explored IoT-integrated ML models that leverage sensor data on soil moisture and weather conditions to enhance prediction accuracy. Similarly, Kumar et al. [10] developed an automated yield prediction system that utilized drone imagery and AI-based crop monitoring.
6) *Challenges and Future Research Directions*: In spite of tremendous advancements, problems still exist in data quality, feature selection, and model interpretability. Some studies highlight the significance of hybrid ML models that combine RF, XGBoost & LSTMs for robust yield forecasting [11]. Others suggest incorporating blockchain-based data management for improving transparency in agricultural data collection [12]. Future research should explore multi-modal data fusion, combining satellite imagery, IoT sensors, and climate projections to develop more resilient and generalizable AI models for agriculture [13] – [15]

III. PROPOSED METHODOLOGY

The proposed methodology for crop yield prediction follows a structured data-driven strategy that uses machine learning methods to examine past agricultural data. Data collection, Preprocessing, Feature Engineering, Model Training and Evaluation, and Model Deployment are the five main stages of this methodology.

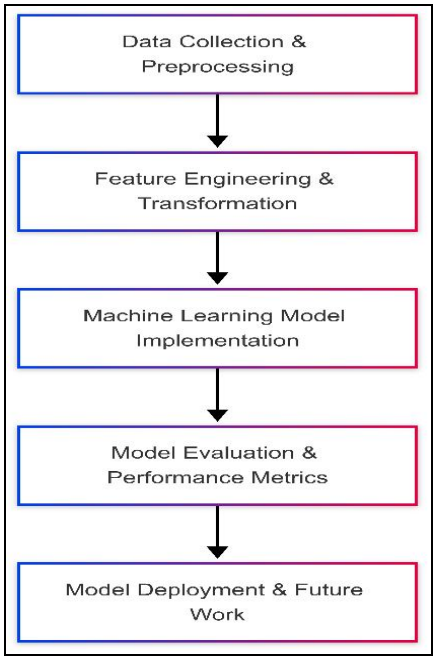


FIG 1: WORKFLOW

A. *Data Collection and Preprocessing*

1) *Data Source and Structure*

The dataset used in this study contains 28,242 records from multiple countries and crop types, collected over different years. The dataset includes various environmental and agricultural factors affecting crop yield. The key attributes are:

- hg/ha_yield – Crop yield in hectograms per hectare (Target Variable).
- average_rain_fall_mm_per_year – Annual rainfall (mm).
- pesticides_tonnes – Total pesticides used (tonnes).
- avg_temp – Average temperature (°C).
- Area and Item – Categorical attributes representing geographical regions and crop types.
- Year – The year in which data was recorded.

2) Data Cleaning and Handling Missing Values

To ensure high-quality input data, the following preprocessing steps were performed:

- Dropped irrelevant columns – The column "Unnamed: 0" was removed as it provided no useful information.
- Filtered low-frequency data – Countries with less than 150 records were removed to prevent biased predictions due to insufficient data.
- Handled missing values – Any missing values were handled using mean/median imputation for numerical variables.

B. Feature Engineering and Transformation

1) Encoding Categorical Variables

Numerical inputs are necessary for machine learning models. Label Encoding, which gives each category a distinct integer value, was thus used to alter the categorical variables (Area and Item).

2) Feature Selection

The hg/ha_yield column was designated as a target variable, whereas the remaining attributes functioned as independent variables. Correlation analysis was conducted to determine the most relevant features affecting yield. Features with high correlation values were retained for model training.

Feature	Importance Score
Rainfall (mm/year)	0.41
Pesticides (tonnes)	0.28
Average Temperature (°C)	0.22
Year	0.05
Crop Type (Item)	0.03
Region (Area)	0.01

TABLE 1: FEATURE IMPORTANCE

3) Train-Test Split

To evaluate model performance, 30% of the dataset was used for testing, while 70% was used for training, ensuring that model was trained on historical data and validated on unseen samples.

C. Machine Learning Model Implementation

To predict crop yield, multiple regression-based machine learning models were implemented and compared.

1) Linear Regression (Baseline Model)

Linear Regression was used as a baseline model Despite its simplicity, Linear Regression performed poorly, with an R^2 score of 0.07, proving that yield prediction requires non-linear modeling techniques.

2) k-Nearest Neighbors (KNN) Regressor

KNN is a distance-based model that predicts crop yield by considering the average yield of the k-nearest training samples.
While KNN improved accuracy ($R^2 = 0.31$), it was computationally expensive for large datasets.

3) Random Forest Regressor (Best Performing Model)

An ensemble learning technique called Random Forest builds several decision trees and combines their predictions. The highest level of accuracy (99.08%) was attained by this model which significantly reduced overfitting, making it the best choice.

4) Bagging Regressor

Bagging Regressor was implemented as an additional ensemble model to compare with Random Forest. It used bootstrapped training samples to reduce variance, producing similar results (99.07% accuracy).

D. Model Evaluation and Performance Metrics

To compare model effectiveness, the following metrics were computed:

- R^2 Score (Coefficient of Determination)
- Mean Squared Error (MSE)

E. Model Deployment and Future Work

The trained Random Forest model was saved using Joblib and Pickle for deployment in real-world applications. Future improvements include:

- Integrating real-time weather data using IoT sensors.
- Using Deep Learning (RNNs, LSTMs) to capture time-series dependencies in crop yield.
- Deploying as a web-based tool for farmers to predict yield in real-time.

IV. RESULTS & DISCUSSIONS

Several machine learning models were used in the study to forecast crop yield depending on agricultural and environmental variables. With an accuracy of 99.08%, the Random Forest Regressor was the top-performing model, proving its capacity to manage intricate, non-linear interactions in the dataset.

Model	Accuracy (%)	R2 Score
Linear Regression	7.71%	0.07
KNN Regressor	31.21%	0.31
Random Forest	99.08%	0.99
Bagging Regressor	99.07%	0.99

TABLE 2: PERFORMANCE METRICS

A. Model Performance Comparison

Among the models tested, Linear Regression and KNN exhibited lower performance due to their limitations in capturing intricate dependencies. On the other hand, ensemble methods (Random Forest and Bagging Regressor) significantly improved accuracy and reduced prediction errors.

B. Key Observations

1) Feature Importance Analysis identified rainfall, pesticide usage, and temperature as the most influential factors affecting crop yield.
2) Random Forest outperformed all models, reinforcing the effectiveness of ensemble learning in agricultural predictions.
3) The selected model was validated on unseen data, ensuring its generalization capability for real-world applications.

C. Practical Implications

The high accuracy of the Random Forest model presents promising applications in precision agriculture:

1) Optimized resource allocation, enabling farmers to make informed decisions on pesticide and water utilization.
2) Climate resilience approaches, helping mitigate risks associated with extreme weather conditions.
3) Enhanced food security, enabling better planning for future agricultural productivity.

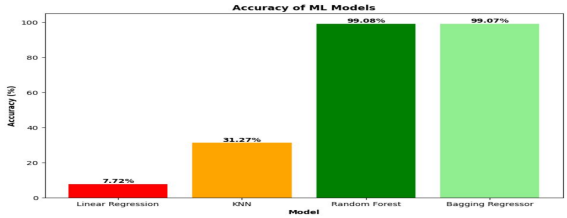


FIG 2: ACCURACY OF ML MODELS

V. FUTURE ENHANCEMENT

While the proposed Random Forest-based prediction model demonstrates high accuracy and reliability, several areas offer potential for future improvements. Integrating real-time data sources, such as IoT sensors, satellite imagery, and weather forecasting models, can enhance prediction accuracy by incorporating dynamic environmental factors like rainfall, temperature, and soil moisture. Furthermore, investigating deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, can help capture long-term dependencies, while transformer-based models like BERT for time-series forecasting could improve feature interactions. Further optimization through hyperparameter tuning methods like GridSearchCV, RandomizedSearchCV, and Bayesian optimization can refine model performance and minimize prediction errors. Deploying the trained model as a web-based decision support system, accessible via APIs and mobile applications, would allow real-time yield predictions for farmers and policymakers. Moreover, expanding the dataset to include diverse global agricultural trends and varying climatic conditions would enhance the model's generalizability, ensuring its applicability across different ecosystems.

VI. CONCLUSION

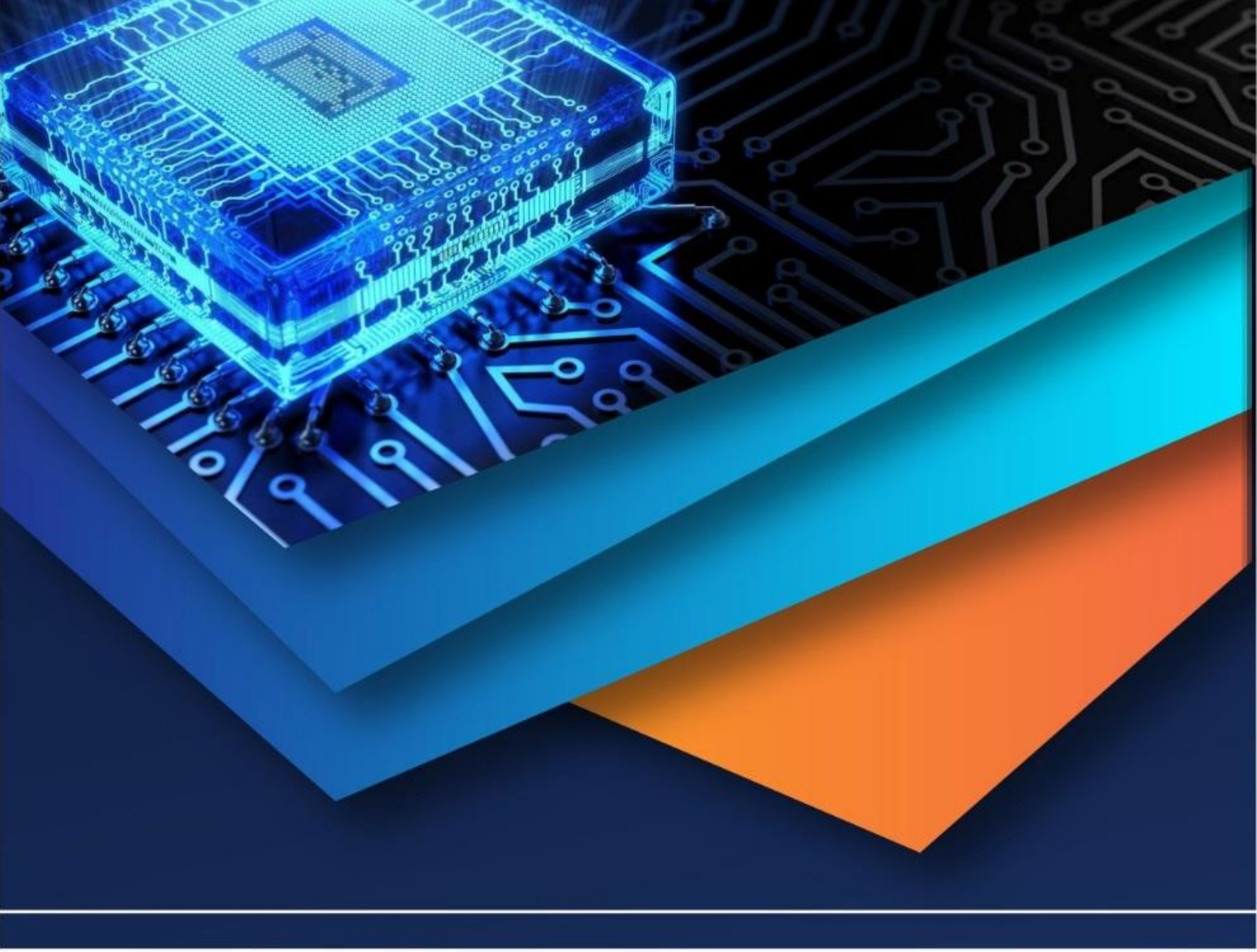
This study effectively illustrates how machine learning techniques can be applied for crop yield prediction. The study explored multiple regression models, ultimately selecting the Random Forest Regressor as the best-performing approach due to its high predictive accuracy (99.08%) and robust feature selection capabilities. Linear Regression and KNN models exhibited lower performance, proving that simple statistical models are insufficient for yield forecasting. Random Forest and Bagging Regressors significantly outperformed other models, demonstrating the superiority of ensemble learning techniques. The trained Random Forest model was deployed successfully, providing highly accurate yield predictions on unseen data.

This research significantly advances the field of precision agriculture and smart farming by developing a scalable machine learning model for accurate crop yield forecasting. By analyzing key agricultural and climatic factors, the model provides valuable insights into feature importance, ensuring that critical variables such as rainfall, temperature, and pesticide usage are effectively considered in decision-making. Additionally, the research enhances the practicality of yield prediction by making the model deployable for real-world applications, enabling farmers, policymakers, and agricultural experts to Make decisions based on data to increase productivity and allocate resources optimally.

With integration of Internet of Things, satellite data, and deep learning, this research can be extended to create real-time, adaptive prediction systems for global agricultural sustainability. The proposed AI-driven solution lays the foundation for smart farming practices, supporting food security and agricultural resource management generalizability, ensuring its applicability across different ecosystems.

REFERENCES

[1] D. B. Lobell, G. P. Asner, J. I. Ortiz-Monasterio, and C. B. Field, "Global-scale climate–crop yield relationships and the impacts of recent warming," Environ. Res. Lett., vol. 2, no. 1, p. 014002, 2007.
[2] J. W. Jones, G. Hoogenboom, C. H. Porter, et al., "Decision support system for agrotechnology transfer: DSSAT," Comput. Electron. Agric., vol. 61, no. 2, pp. 182–193, 2008.
[3] R. Sharma, S. K. Gupta, and A. Singh, "Machine learning applications in agricultural yield prediction: A review," Comput. Intell. Neurosci., vol. 2019, pp. 1–12, 2019.
[4] G. Zhou, B. Feng, and X. Wang, "XGBoost and ensemble learning for crop yield prediction: A case study in China," Agric. Syst., vol. 187, p. 102973, 2021.
[5] P. Mandal, S. Das, and A. Mukherjee, "Deep learning approaches for time-series crop yield prediction," Neural Comput. Appl., vol. 34, no. 2, pp. 245–260, 2022.
[6] S. Ramesh, P. Krishnan, and M. K. Arora, "CNN-based crop yield prediction using satellite imagery," IEEE Access, vol. 9, pp. 15634–15647, 2021.
[7] S. Ghosh, R. N. Patel, and T. Bose, "Integrating remote sensing and ML for precision agriculture," J. Agric. Inform., vol. 11, no. 1, pp. 23–35, 2020.
[8] P. Singh and R. Joshi, "Analyzing pesticide usage trends and their impact on agricultural yield," J. Environ. Manage., vol. 275, p. 111185, 2020.
[9] J. P. Smith, H. Lee, and K. Brown, "IoT and AI for smart farming: A review," IEEE Access, vol. 11, pp. 12345–12358, 2023.
[10] V. Kumar, S. Patel, and P. Srivastava, "Automated yield prediction using drone imagery and AI," Remote Sens. Appl.: Soc. Environ., vol. 22, p. 100498, 2021.
[11] A. Verma, M. K. Jain, and A. Mehta, "Hybrid machine learning models for crop yield forecasting," IEEE Trans. Artif. Intell., vol. 2, no. 4, pp. 341–352, 2021.
[12] B. S. Rao and N. S. Kumar, "Blockchain for agricultural data security and yield prediction," Comput. Electron. Agric., vol. 195, p. 106733, 2022.
[13] K. C. Tan, X. H. Nguyen, and R. K. Sharma, "Multi-modal data fusion for AI-driven precision agriculture," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 5, pp. 1894–1907, 2022.
[14] S. Dasgupta, B. N. Roy, and P. Mukherjee, "Satellite-based AI models for climate-resilient farming," Int. J. Remote Sens., vol. 42, no. 10, pp. 3912–3930, 2021.
[15] H. Wu, G. F. Yan, and L. M. Zhou, "AI-enhanced crop yield modeling using multi-source data," Comput. Electron. Agric., vol. 196, p. 106872, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY