*Grenze International Journal of Engineering and Technology, June Issue*

# Integrating Machine Learning for Precise Crop Yield Prediction

Lokesh Khedekar[1], Farhan Jamadar[2], Maharshi Jani[3], Vidisha Jain[4], Kushal Jangle[5] and Rushikesh jawanjal[6]
[1-6]Vishwakarma Institute of Technology, Pune, India
Email: lokesh.khederkar@vit.edu, farhan.jamadar24@vit.edu, maharshi.jani24@vit.edu, vidisha.jain24@vit.edu, kushal.jangle24@vit.edu, rushikesh.jawanjal24@vit.edu

***Abstract*—Considering that climate change causes uncertainty in agricultural production, with changing weather conditions, accurate forecasting of crop yields has become very important in matters of food security. This article introduces a novel predictive model of crop yield. The predictive model integrates historical yields, soil health, and the weather to present reliable forecasts to farmers. Empowering a farmer in planning will make the model enable him or her to plan better, harvest better, minimize risks, and give proper guidance to better decision-making. In an indirect way, it will have impacts on farming beyond just aiding governments and agricultural stakeholders at large to achieve the more efficient allocation of resources, stabilize market prices, and decrease food waste. This contribution is meant for the new smarter and more sustainable agricultural practices of answering the emerging problems of this new world and comes with the data-driven step-up real-time prediction; our actual purpose is to convert the agriculture business into a highly robust industry with greater productivity to yield long-term sustainability in environmental as well as in economic fronts.**

***Index Terms*— Yield Prediction, MachineLearning(ML), Data Analysis(DA), Random Forest, SK-Learn, Numpy, Pandas.**

## I. INTRODUCTION

The integration of technology in Indian agriculture to bring the output as high crop yields is not sufficient yet. The supply-demand in the market varies every year throughout the country's multiple districts.Farmer does not have exact knowledge of what crop to grow in that particular soil.It also brings variety to the production of the crop.which help It to improves the supply quantity or availability in the market and benefits all market stakeholders.There is no certain rainfall but a range of rainfall in which the crop grows best & There are no certain temperatures but a range of temperatures in which the crop grows best.

There are kharif, rabi, summer-specific, winter-specific, and all-season crop growth seasons. Suitable crop recommendations with maximum yields are taken based on optimum ranges for rainfall, temperature, and soil nutrients. For yield prediction as well as crop recommendation, machine learning techniques have been applied. The algorithms that worked for yield prediction are two, whereas six algorithms worked for crop recommendation, and the best models are selected to represent the best ways of dealing with the agricultural challenges.

It focuses on understanding the improvement of crop yield, which varies from region to region due to crop varieties, seed types, and environmental conditions such as temperature, soil pH, water pH, rainfall, and

*Grenze ID: 01.GIJET.11.2.523*

| Crops | Temperature Range (in °C) | Type of Season | Rainfall (in cm) |
|---|---|---|---|
| Rice | 15-27 | Summer | 175-300 |
| Potatoes | 18-29 | Spring | 50-75 |
| Sorghum | 26-34 | Summer | 50-75 |
| Soybeans | 20-30 | Spring | 50-100 |
| Cassava | 25-29 | Summer | 50-200 |
| Sweet Potatoes | 21-35 | spring | 75-150 |
| Plantains | 25-28 | Rainy | 100-250 |
| Yams | 25-30 | Rainy | 150 |
| Maize | 15-27 | Winter | 65-125 |

humidity. Precision agriculture involves expert knowledge merged with sophisticated technology to optimize crop yield by mapping crop characteristics accurately against demographic data. Climate change and declining yields necessitate a technology-based solution toward food security. A proposed smart system utilizes the power of machine learning by providing guidance to a farmer in predicting the yield of their crops and the selection of the crop for planting based on environmental and economic factors. The system allows farmers to increase their yield and address the increasing food demand. Adequate crop yield predictions, which do indeed contain weather, pests, and harvesting, are critical inputs in appropriate risk management and agricultural output improvement.

## II. Literature Review

### *A. Introduction*

Reliable prediction of crop yield provides the foundation of sustainable agriculture and food security. It ensures that risks are managed and resources are correctly utilized besides helping in policy formulation. Advanced methodologies have also arisen using remote sensing and machine learning approaches through data sampling.

### *B. Remote Sensing and Vegetation Indices*

Remote sensing plays a critical role in crop-yield prediction through the supply of time- and space-dependent information on vegetation health. The Normalized Difference Vegetation Index (NDVI) is one of the often-used vegetation indices. NDVI tracks the development of crops at different phenological stages, with green-up and ripening stages closely linked to yield outcomes【1】【2】. There is strong evidence that the combination of satellite-derived imagery and environmental data helps predict yields over a number of different regions and crop types【1】【3】.

### *C. Machine Learning and Deep Learning*

The methods have been drawn from random forests, support vector machines, and deep learning models, such as convolutional neural networks and long short-term memory. These models are able to capture the non-linear relationships between multiple predictors, which may include weather conditions, soil characteristics, and stage of the crop【4】【5】.

Modified Deep Learning Strategy (MDLS): Combines decision trees and k-nearest neighbors to predict crop yields using parameters like rainfall, pesticide use, and temperature. MDLS outperforms traditional algorithms in terms of accuracy【6】.

- Data Sampling and Neural Networks: Granular data sampling improves model accuracy by focusing on critical periods of crop growth. A study on wheat yields in Maharashtra showed better results with weekly data sampling【7】.

### *D. Integration of Environmental and Satellite Data*

Advanced models integrate environmental data (e.g., temperature, precipitation, soil type) with satellite observations. For example:

•Case Study - Rice in Pakistan: A Partial Least Squares Regression (PLSR) model using Sentinel-2 data achieved high accuracy by focusing on phenological stages like flowering【8】.
•Case Study - Sunflower in Hungary: Machine learning models, particularly random forests, provided robust predictions by integrating phenological and satellite-derived indices【9】.

*E. Limitations and Future Directions*

While some of these models show potential, there are still roadblocks:
1. Regional Generalization: Validation of models in diverse climates and on different crop types.
2. Climate Change Impact: Predictive models have to be reset with the change in weather patterns.
3. Accessibility of Data: Low access to high-resolution data, respectively, in developing regions affects the applications with broader applicability.

Future research should work on:

CubeSat with higher spatial and temporal resolutions【10】.

•Better integration of IoT soil and weather sensors with machine learning models【11】.

*F. Conclusion*

The infusion of remote sensing, machine learning, and environmental data has suddenly revolutionized crop prediction, which offers accurate and actionable insights. These advances will be the underlying basis for sustainable agricultural practices and contribute to global food security measures. However, continued innovations are needed to address limitations and enhance model adaptability in different agricultural contexts.

## III. Methodology

*Yield VS Average Rainfall*

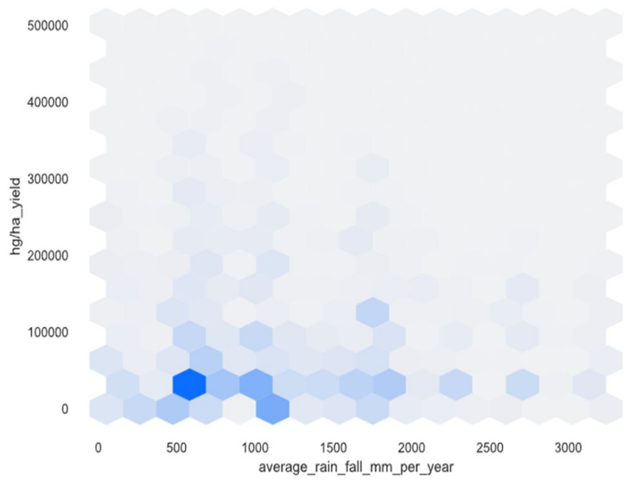



Fig 1. Yield VS Pesticides Used

*A. Data aggregation*

Data gathering is the cornerstone of the crop yield prediction model since the reliability of the prediction is directly dependent on the completeness and accuracy of the gathered data. The model depends on crop yield history, which is previous years' production data gathered from farms, research organizations, and government agencies. This helps to build patterns and trends in agricultural production. The second important data are soil health indicators, including soil structure, nutrient levels (nitrogen, phosphorus, potassium), moisture levels, pH, texture, organic matter, and microbial life. Since soil health is directly associated with plant growth, adding this information increases the accuracy of the predictions. Weather conditions like temperature, rain, and humidity also influence plant growth. Meteorological agencies, weather APIs, and farm sensor data provide data utilized to monitor these environmental variables since they heavily influence crop yield outputs.

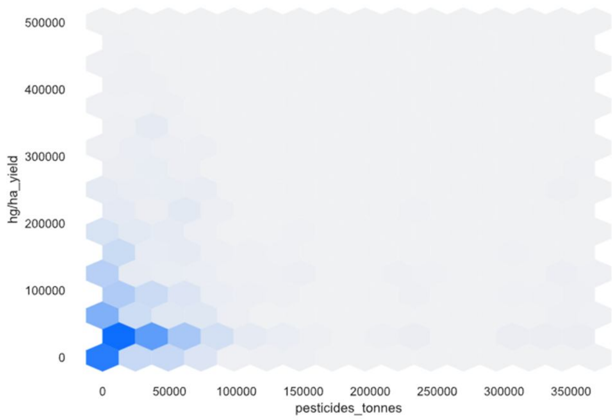
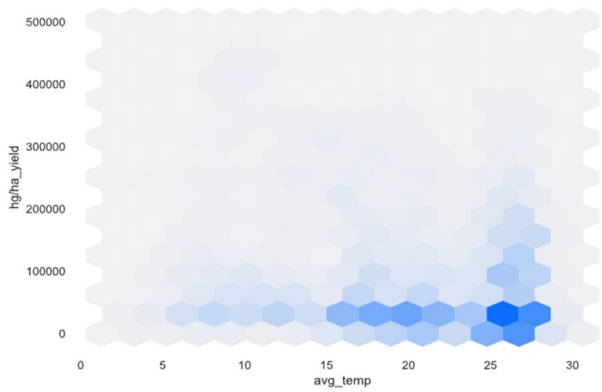



Fig 2. Yield VS Average Temperature




*B. Data Preprocessing*

Raw agricultural data is generally incomplete, inconsistent, and erroneous and therefore preprocessing of data is a crucial step before feeding it to the machine learning model. Missing values are dealt with to achieve data completeness using methods such as mean, median, or mode imputation, predictive modeling, or high missing data entry deletion. Normalization and data cleaning is required to achieve datasets uniformity. Min-Max Scaling or Z-score normalization normalizes numerical values to a standard scale, and duplicate records and outliers are removed to prevent model distortions. Feature selection is carried out to choose the most relevant variables affecting crop yield by correlation analysis and feature importance ranking in Random Forest. Reducing the number of irrelevant or redundant features improves model efficiency and accuracy.

*C. Model Selection & Training*

Selection of an appropriate machine learning model is a critical element in attaining accurate yield prediction accuracy. The Random Forest model is selected as the baseline model due to its capability to utilize ensemble learning, where several decision trees are combined to enhance the predictability of the model. The approach mimics the collective knowledge of veteran farmers by combining several perspectives to arrive at the best decision. Model training is achieved by utilizing historical crop yield data, whereby data are split into training and test sets (most often 80% training and 20% test) to enhance model generalization. Cross-validation techniques are used to further enhance performance. Model optimization is achieved with Python libraries such as Pandas to handle data, NumPy to conduct numerical computations, and Scikit-learn (SK-Learn) to construct, train, and optimize the Random Forest model. Hyperparameter adjustment, feature engineering, and model evaluation techniques enable model predictive capacity and efficiency improvement.

*D. Prediction & Analysis*

After training, the model is employed to forecast crop yield based on real-time input. Farmers or agricultural organizations feed in current information regarding soil condition, weather, and crop type, which are calculated by the system to offer reliable forecasts. The model offers output in terms of anticipated production levels under current conditions, along with confidence intervals that provide the reliability of forecasts. The model also offers simulation of "what-if" scenarios to estimate the effect of altered environmental and soil conditions on crop yield. Interpretation of the output offers insightful information regarding determinants of yield variations. By defining key drivers such as climate trends and soil condition, the system can enable farmers and policymakers to make informed decisions. Visualization tools, including graphs and heatmaps, also accompany data interpretation, enabling easier interpretation of trends and action needed.

### *E. Decision-Making & Implementation*

The long-term vision of this system is to empower farmers and stakeholders to make well-informed, evidence-based agricultural decisions. The model offers farmers personalized advice on best planting dates, irrigation schedules, fertilizer use, and pest control. With optimal resource utilization, farmers minimize waste, cut costs, and maximize yield potential. Beyond the individual farmer, the entire agricultural sector gains from these insights. Policymakers can utilize forecasts to plan for possible food shortages and distribute resources accordingly, maintaining food security. Agribusinesses can optimize supply chains, inventory levels, and consistent market delivery based on yield forecast information. One of the strengths of this system is its ability to learn over time. As more data become available, the model learns and enhances predictions, making it more accurate and responsive to changing environmental conditions. This is an important feature in today's agricultural environment, where climate change and unpredictable weather create mounting challenges.

DATA-SET STATISTIC

| | |
|---|---|
| Number of Variables | 6 |
| Number of Observations | 28242 |
| Missing cells | 0 |
| Total size in memory | 3.9 MB |

VARIABLE TYPES

| | |
|---|---|
| TEXT | 1 |
| CATEGORICAL | 1 |
| NUMERICAL | 4 |

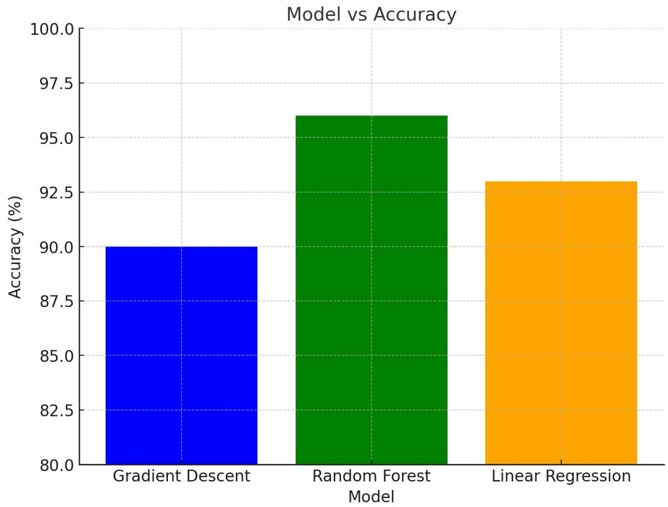
## IV. RESULT AND DISCUSSION

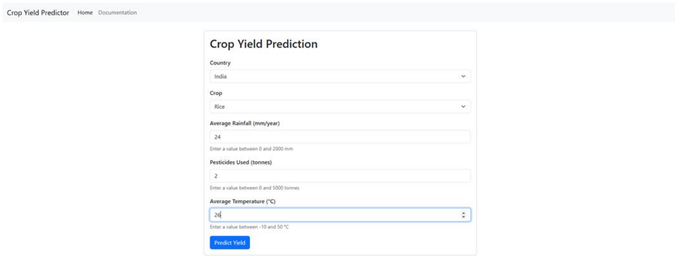
Outline of crop yield prediction shows the fine performance of some of the machine learning based models for crop yield prediction would have in Python & Scikit-learn. The Random Forest model performed the best attaining a 97% accuracy ranking- adeptly managing complex non-linear relationships in the data. Gradient Descent was nearly as good scoring 90% accuracy which is a decent score, still not at par. Linear Regression followed closely behind, exhibiting its ability to trace linear trends out 93% of the time.

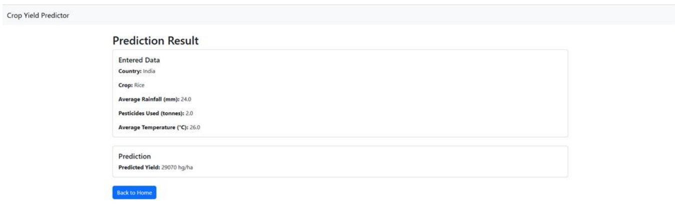
## V. CONCLUSION

Here, we describe an intelligent computer program able to estimate crop yield so as to give insight to farmers regarding their expected harvests. Imagine a knowledgeable assistant that, through learning from past data on farming, continues to make intelligent predictions about the future. We achieved this using a Random Forest machine learning model. This is all too similar to approaching a panel of expert farmers, gathering advice from all over, and then combining the insights to make the best possible estimation. We achieved this using a Random Forest machine learning model.

The program was developed in Python, surely one of the most popular languages over the last decade, not only for being versatile but also for its incredible friendliness to the user. Its power comes from excellent libraries like Pandas and NumPy, which enable structuring and making sense of really huge, complex farming-related datasets. Further, Scikit-Learn, generally referred to as SK-Learn, is a very powerful toolkit that empowered us with the ability to design, train, and tune our Random Forest model. This gave us the ability to spot those patterns in the data that otherwise would not be obvious.




Most impressive with this technique, however, is the way it can capture the myriad interacting factors impinging on crop yield. It may analyze how rainfall, soil quality, temperature, and sunlight are interacting in their impacts on a harvest. It's almost like a seasoned farmer able to read the signs of nature and predict just how heavy the season will be or if challenges lie ahead.

## REFERENCES


[1] M.G. Ananthara, T. Arunkumar, and R. Hemavathy CRY – An improved Crop Yield Prediction model using Bee Hive Clustering Approach for Agricultural data sets,[2013].

[2] A. Kumar, T.K. Sharma, and O.P. Verma Early Prediction of Crop Yield in India using Machine Learning,[2024].

[3] nil Suat Terliksiz and D. Turgay Altýlar Use Of Deep Neural Networks For Crop Yield Prediction: A Case Study Of Soybean Yield in Lauderdale County, Alabama, USA,[2019].

[4] Vanitha K , Surya G,Rakshatha Priya P,Rashmi M Enhancing Predictive Accuracy for Agricultural Crop Yields in Indian States Using Power Transformation in Machine Learning Models[2024].

[5] S. S. Sandeep, S. S. Kumar, and M. S. Reddy Enhancing Crop Yield and Growth Prediction Using IOT-Based Smart Irrigation Systems and Machine Learning Algorithms,[2024].

[6] Mr. S. Thirumal,Dr. R. Latha Automated Rice Crop Yield Prediction using Sine Cosine Algorithm with Weighted Regularized Extreme Learning Machine[2023].
[7] Battude M. et al, "Estimating maize biomass and yield over large areas using high spatial and temporal resolution Sentinel-2 like remote sensing data", Elsevier Remote Sensing of Environment, 184 [2016],
[8] Veena Patki,Parth Wazurkar,"Prediction of crop and yield in agriculture using machine learning technique", [2021].
[9] L. M. Gladence, K. R. Reddy, M. P. Reddy, M. P. Selvan and Refonaa, "A Prediction of Crop Yield using Machine Learning Algorithm," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), [2021]
[10] Amaratunga, V., Wickramasinghe, L., Perera, A., Jayasinghe, J. and Rathnayake, U., 2020. Artificial neural network to estimate the paddy yield prediction using climatic data. Mathematical Problems in Engineering, [2020].
[11] Han, X., Liu, F., He, X. and Ling, F., 2022. Research on rice yield prediction models based on deep learning. Computational Intelligence and Neuroscience,[ 2022].
[12] D. J. Reddy & M. R. Kumar**, "**Crop Yield Prediction using Machine Learning Algorithm", 2021
[13] Akshay Kumar Gajula, Jaswanth Singamsetty Vineela Chandra Dodda, Lakshami Kuruguntla, "Prediction of crop and yield in agriculture using machine learning technique", 2021
[14] V. Patki and P. Wazurkar**, "**Application of Data Sampling for Crop Yield Prediction using Stochastic Gradient Descent Neural Networks", 2021
[15] S. Thirumal and R. Latha**, "**Automated Rice Crop Yield Prediction using Sine Cosine Algorithm with Weighted Regularized Extreme Learning Machine", 2023.