# Climate-Driven Crop Yield Forecasting using Regression Model

*Kapil Keshao Wankhade[1], Ganesh Khekare[2]*
[1] Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia; [2] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, 632014, Tamil Nadu, India
Email ID [1]**pdf.kapil@lincoln.edu.my; [1]kaps.wankhade@gmail.com; [2]khekare.123@gmail.com**

**Abstract:** The importance of crop yield prediction lies in accurate yield prediction in climate-resilient agriculture, precision farming, and sustainable food production systems. This paper introduces the multi-source environmental information-based and supervised regression modeling, applying to the current study a well-structured data-driven model to predict crop yields. The suggested methodology initiates with the multi-source data collection that combines the usage of IoT-driven soil sensors, meteorological measurements, satellite-generated vegetation indices, and past records of crop yields. The correlation analysis and ensemble-based importance ranking are the feature selection methods used to determine important predictors. Predictive performance can be improved by feature engineering strategies to give agronomic indicators such as growing degree days, cumulative rainfall indices, and the interaction variables of climatic factors. Random Forest Regressor, Support Vector Regressor, and Gradient Boosting Regressor are supervised regression models that are trained to predict values of continuous crop yield. Root Mean Square error, Mean Absolute error, and coefficient of determination are used as measures of model performance. It has been shown that ensemble-based models are useful in capturing the nonlinear interactions of the environment and enhancing prediction.

**Keywords**: Crop Yield Prediction; Feature Engineering; Machine Learning; Regression Models; Precision Agriculture

## Introduction

Agriculture is one of the most climate-sensitive sectors and impacts food security, rural livelihoods, and economic security globally. The crop growth and yield are affected by variation in the rainfall, temperature, soils, and moisture. Due to the unpredictable nature of the climate and the frequent occurrence of extreme weather patterns, proper crop yield prediction has become an important factor in sustainable agriculture planning and risk reduction policies. Recent global assessments indicate the need for better climate predictive systems based on environmental cues and more complex analytical systems [1] based on yield variability in key crop production areas of the world, with the effects of climate change.

The crop has been used to estimate the yield response to climatic variables using traditional statistical approaches. The relationships between productivity trends, rainfall, and temperature were quantifiable with regression-based models. However, they are often unrealistic, and linearity assumptions are made and use only linear combinations of environmental factors that do not take into account the interactions between environmental variables in a nonlinear manner [2]. The complex and dynamic nature of climate systems requires further developments in the adaptive and data-intensive approaches to make reliable forecasts.

The crop yield prediction has been greatly improved by advancements in the machine learning field. On the other side, environmental nonlinear dependencies are effectively modeled by statistical models such as Random Forests and Support Vector Machines (SVM) [3]. These models can handle a huge volume of data as well as identify complex relationships between the features that is particularly useful in climate-based agriculture forecasting. Deep learning (DL) techniques can further be extended to offer capabilities to learn hierarchical representations in high-dimensional data sources, such as remote sensing imagery and temporal climate sequences. CNNs and Long Short-Term Memory (LSTM) networks have demonstrated good performance in estimating the spatio-temporal yield patterns in various agroclimatic regions [4]. Combining the satellite-retrieved vegetation indices (e.g., NDVI) with climatic variables increases the spatial generalization and the strength of prediction [5].

Besides the satellite information, Internet of Things (IoT) based environmental sensors will allow monitoring soil moisture, pH, temperature, and electrical conductivity in real time. This ground data is of high-resolution and enhances the predictive model sensitivity to short-term climatic variations [6]. Nevertheless, there are also obstacles in terms of data heterogeneity, complexity of preprocessing, model interpretability, and transferability across regions.

Recent systematic reviews note that there is a necessity to have integrated structures that combine multi-source environmental information with explainable and scalable machine learning systems [7]. These issues are important in terms of coming up with efficient working crop yield predictive systems that can be used to support precision agriculture and climate-informed decision-making.

The organization of the paper is as follows: Section 2 discusses related work, and the methodology is included in Section 3. The conclusion is concluded in Section 4.

## Related work

### A. Parameters for crop yield prediction

This subsection highlights the basic importance of planning the appropriate environmental and agronomic inputs. There are many factors determining the crop yield, some of which are variable, like climatic variations, soil properties, and health indicators of vegetation. Temperature and rainfall have a direct impact on the plant physiological processes, including photosynthesis, evapotranspiration, and flowering. Climatic stress, specifically extreme heat events and irregular precipitation, may greatly increase yield gaps and lower productivity, as highlighted in [8] and [9].

### B. Methods of feature selection and feature engineering

The feature selection is necessary to eliminate the redundant or unnecessary variables that are likely to cause noise in the model. Natural agricultural data of high dimension is likely to have attributes of climatic phenomena that are correlated. The methods based on the mutual information and correlation analysis are used to find statistically significant predictors [10]. Random Forest-based ensemble-based importance ranking offers this benefit because it quantifies the contribution of every feature to predictive accuracy [11].

### C. Regression measures of yield prediction

The use of machine learning regression models has become prominent because it is able to model nonlinear relationships. Random Forest which was brought about in [11] minimizes overfitting by the virtue of ensemble averaging and boot strapped sampling, therefore, it is applicable in heterogeneous agro-climatic data. The support vector regression [12] is good at working with large dimension feature spaces by maximizing margin boundaries and applying the kernel transformations.

**Methodology**

Fig. 1. is the overall methodological pipeline of prediction of crop yield by first acquiring multi-source data and then moving to preprocessing, feature selection, feature engineering, supervised regression modeling, evaluation, and eventual yield prediction. The format guarantees the systematic conversion of uncooked agro-environmental information into an enhanced predictive model in a systematic sequence of calculations.

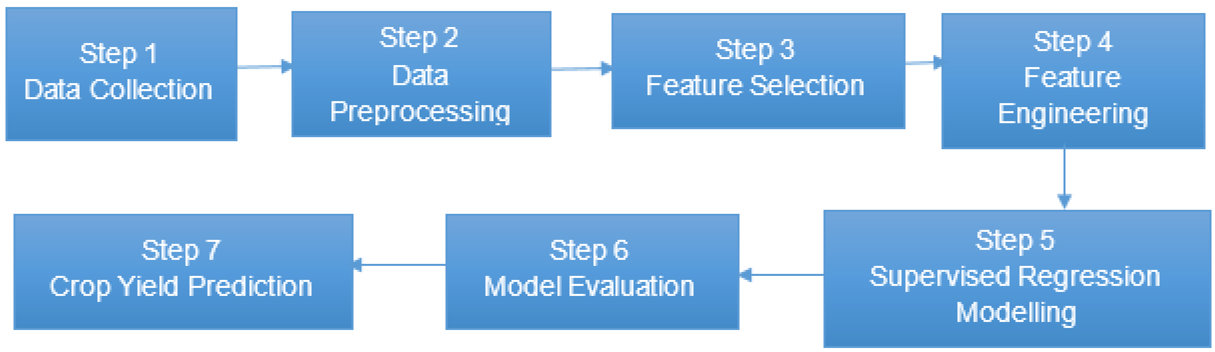


*Fig. 1. Proposed Framework for Climate-Driven Crop Yield Prediction Using IoT and Machine Learning Regression Models.*

A. Data collection

The proposed crop yield prediction framework is based on the data collection stage. Different heterogeneous agricultural datasets are collected in this stage to represent the environmental and agronomic situation that influences crop productivity. These data sources will comprise soil sensors equipped with Internet of Things technology, meteorological data, satellite-based vegetation indices, and historical crop yield data. Such a heterogeneous relationship between climate, soil properties, and plant development patterns is useful in yield modeling.

B. Data preprocessing

In the proposed system, data preprocessing steps play a vital role and ensure that the collected agricultural data are in pristine, uniform, and friendly machine learning format. Due to sensor failures, changes in environmental conditions, recording errors, etc., raw data collected from sensors, meteorological stations, and satellites in IoT environments often contains missing values, noise, and extreme environmental values. The preprocessing methods, such as handling of missing data, feature scaling, normalization, outlier detection, and data subsetting, also help to improve the performance and robustness of the predictive model. After these preprocessing steps, the normalized dataset can be used for feature selection and machine learning regression modeling. The ability of the model to be useful can seriously compromise the reliability of the predictions if the data acquired from the environment is not reliable.

C. Feature selection

The feature selection process is a significant part of the suggested crop yield prediction framework because it assists in identifying specific environmental variables that have a significant impact on crop productivity. Farm records tend to have numerous climatic parameters and soil parameters, most of which can be superfluous or have low predictive power on crop output. The addition of such irrelevant variables may result in more complex models and lower prediction accuracy. Hence, the method of feature selection is employed to remove the most uninformative variables that do not help in the prediction of yields.

The statistical and machine learning-based feature selection methods are employed in this study to determine the significance of each parameter of input. These tools examine the correlation between the environmental factors and crop production to determine which characteristics are to be included in modelling.

D. Feature engineering

The feature engineering process is to engineer features by converting the features and modifying them into new variables in order to better represent the underlying relationships between the environmental conditions and crop yield. Raw climatic and soil data might be inadequate for crop growth dynamics in agricultural prediction problems. Thus, agronomic predictors that are specific to the domain are developed to boost the prediction power of machine learning models.

Various derived characteristics are created in this work based on environmental characteristics like temperature, rainfall, and soil characteristics. These are constructed properties that are used to capture the time variation, cumulative environmental impacts, and non-linear interactions of the climatic factors. These artificial characteristics can make the regression model more representative of the complicated interactions between the environments and enhance the overall precision of crop yield forecasting.

E. Supervised regression modelling

In the suggested structure, regression models are taught through supervision to understand the connection between the environmental characteristics and crop production. The regression model is appropriate in this issue since crop yield is continuous numerical data that is subject to various climatic and soil conditions. After preprocessing, the regression models are trained on the optimized feature sets.

F. Model Evaluation

In testing the performance of the trained regression models, there are a number of statistical measures employed. These indicators reflect the difference between the expected crop yield information and the actual yield. Appropriate assessment helps to measure the effectiveness and reliability of the prediction model.

a. One of the commonly used evaluation measures is Root Mean Square Error (RMSE) which measures the average magnitude of the errors in the predictions.

b. The MAE (Mean Absolute Error): it is the average absolute error between the actual and predicted values.

c. The Coefficient of Determination (R2) measure the amount of variance in crop yield explained by the model.

The smaller the RMSE and MAE and the higher the R2, the more successful the model and its predictive capacity.

**Conclusions**

In this paper, the authors have suggested an organized machine learning model which is capable of predicting yield of crop with the help of environmental and agronomic characteristics with accuracy. The proposed methodology consists of different stages such as multiple-source data collection, data pre-processing, feature selection, feature engineering, supervised regression model and testing the model performance. It covers the interaction between environment and crop productivity, which is a combination of the data collected by the IoT sensors, meteorological data, soil data and vegetation indices. Machine learning models can be trained to learn nonlinear behavior patterns of the environment by using the regression assessment technique like RMSE, MAE and R2 and can give credible yield forecasts.

**References**


1. E. Najafi, N. Devineni, R. M. Khanbilvardi, and F. Kogan, “Understanding the Changes in Global Crop Yields through Changes in Climate and Technology,” Earth’s Future, vol. 6, no. 3, pp. 410–427, 2018.
2. D. B. Lobell and M. B. Burke, “On the Use of Statistical Models to Predict Crop Yield Responses to Climate Change,” Agricultural and Forest Meteorology, vol. 150, no. 11, pp. 1443–1452, 2010.
3. J. H. Jeong, A. M. Resop, N. D. Mueller, D. B. Lobell, “Random Forests for Global and Regional Crop Yield Predictions,” PLOS ONE, vol. 11, no. 6, e0156571, 2016.
4. S. Khaki, L. Wang, and S. V. Archontoulis, “A CNN–RNN Framework for Crop Yield Prediction,” Frontiers in Plant Science, vol. 10, p. 1750, 2020.
5. H. Kamangir, “Large-Scale Spatio-Temporal Yield Estimation via Deep Learning Using Satellite and Management Data Fusion,” Computers and Electronics in Agriculture, vol. 216, p. 108439, 2024.
6. E. Rajitha, “Time Series Forecasting of Crop Yield under Climate Variability Using IoT Data and LSTM Networks,” International Journal of Environmental Sciences, 2025.
7. P. Muruganantham, “A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing,” Remote Sensing, vol. 14, no. 9, p. 1990, 2022.
8. D. B. Lobell, K. G. Cassman, and C. B. Field, “Crop Yield Gaps: Their Importance, Magnitudes, and Causes,” Annual Review of Environment and Resources, vol. 34, pp. 179–204, 2009.
9. C. B. Field, “Climate Change Impacts on Agriculture,” Science, vol. 341, no. 6145, pp. 508–513, 2013.
10. I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” Journal of Machine Learning Research, vol. 3, pp. 1157 1182, 2003.
11. L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
12. V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.