Ann. Agric. Res. New Series Vol. 47 (1) : 101-108 (2026)

Leveraging predictive models for accurate crop yield forecasting-smart harvest approach

Gunjan Chauhan[1]*, Sanjeev Tomar[1], Anshika Panwar[2], Anushree Misra[3], Ranjit Kumar Paul[4], Md Yeasin[4], Anil Kumar[5] and Sanjeev Panwar[6]

Amity University, Noida Uttar Pradesh
**Corresponding Author's Email: 12309gunjan@gmail.com*

Received: December 2025; Revised Accepted: March 2026

Abstract

This research introduces *Smart Harvest*, a data-driven forecasting system developed to predict the yield of key crops such as sugarcane, wheat, rice, and maize. By combining historical yield records with rainfall statistics, the model captures the intricate relationship between crop production and climatic variation. The study evaluates several predictive techniques, including ARIMA, ARIMAX, Support Vector Regression (SVR), Random Forest, and Artificial Neural Networks (ANN). Through a comparative performance analysis using metrics like RMSE, MAE, and MAPE, the results highlight the effectiveness of rainfall-integrated models in improving forecasting accuracy. The findings provide a robust foundation for enabling timely, informed agricultural planning and risk management strategies.

Keyword: Crop yield forecasting, predictive analytics, machine learning, artificial neural networks (ANN), time series prediction, smart agriculture

Introduction

Agriculture continues to be the lifeblood of India's economy, employing over 50% of the nation's labour force and contributing significantly to GDP, food security, and rural development. As a sector deeply influenced by seasonal variability, agriculture in India is both vital and vulnerable. Among the key metrics used to assess agricultural performance, **crop yield** stands as a fundamental indicator of productivity, efficiency, and sustainability. Yield not only impacts a farmer's income but also plays a crucial role in determining national food stock levels, commodity pricing, and trade policies. Consequently, ensuring accurate and timely crop yield forecasts phrase is essential for effective decision-making at both micro and macro levels.

However, predicting crop yield is an inherently complex task. Yield is influenced by a broad spectrum of variables—ranging from biological and genetic factors to soil health, pest incidence, farming practices, and environmental conditions. Among these, rainfall remains one of the most dominant and unpredictable elements, particularly in a country like India, where nearly 60% of agriculture is rain-fed. The timing, amount, frequency, and geographical distribution of rainfall during the growing season have a direct impact on crop health, germination rates, flowering, and harvest potential. Different crops exhibit varied responses to rainfall patterns. For instance, **rice** requires consistent water availability and is highly sensitive to monsoon variability.

In contrast, **wheat** is generally grown in cooler, drier conditions and is more sensitive to temperature shifts than rainfall quantity.

1-Student, Associate Professor AMITY, 2- Student BVCE Delhi, 3-ASV USA, 4-Senior Scientist IASRI, 5-ADG Coordination, ICAR, 6-Principal Scientist ICAR

Crops like maize and sugarcane are moderately resilient but still depend heavily on rainfall during critical growth phases. Understanding the specific relationships between rainfall trends and crop-specific yield outcomes is crucial to developing effective forecasting systems that can adapt to India's agricultural diversity.

In the past, crop forecasting has largely relied on traditional statistical techniques or farmer intuition. Methods like historical average or simple linear regression, while straightforward, often fail to capture the nonlinear, lagged, and multidimensional nature of climate-yield interactions. Moreover, these models typically lack adaptability, cannot account for real-time fluctuations, and are not easily generalizable across multiple crop types or regions. With the increasing impact of climate change, such limitations have become even more pronounced, necessitating a shift toward more intelligent, flexible, and scalable prediction tools.

To address these challenges, there has been a growing interest in applying advanced analytical methods—such as time series modeling, machine learning (ML), and artificial intelligence (AI)—to agricultural forecasting. These models can process large volumes of historical and real-time data, recognize hidden patterns, and learn from complex, nonlinear relationships between variables. Tools like ARIMA (AutoRegressive Integrated Moving Average) can model temporal trends in yield, while ARIMAX extends this capability by including exogenous variables such as rainfall. Support Vector Regression (SVR) and Random Forest Regression (RFR), on the other hand, offer the flexibility to handle high-dimensional inputs and capture intricate dependencies among features. Additionally, Artificial Neural Networks (ANN) simulate human-like learning and are capable of modeling complex functions when properly trained.

In this research, we present Smart Harvest, a unified forecasting system that evaluates and compares the performance of these models for four major crops: sugarcane, wheat, rice, and maize. Using historical yield and rainfall data, the study builds individual predictive models for each crop and compares them using key metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The results aim to identify the most accurate and generalizable model, offering a foundation for smarter agricultural planning and risk mitigation.

Furthermore, the Smart Harvest framework serves as a scalable and adaptable platform that can be extended to additional crops, locations, and climatic variables. It demonstrates how data-driven insights can be translated into actionable strategies, not only for farmers but also for policymakers, researchers, and agribusiness stakeholders. By aligning agricultural practices with predictive intelligence, this system seeks to reduce uncertainty, enhance productivity, and contribute to sustainable and climate-resilient agriculture.

Materials and Methods

This section outlines the materials and step-by-step methodology adopted to forecast crop yields using machine learning and time series models. The methodology ensures reproducibility, clarity, and transparency in the predictive modeling process, spanning data acquisition to model evaluation.

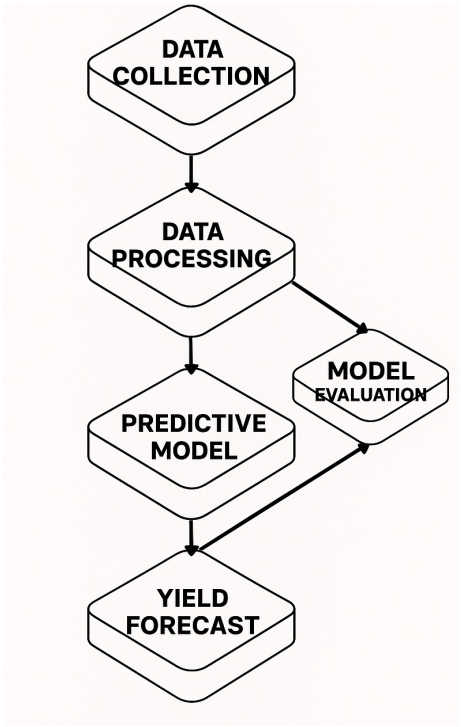

Fig. 1. Data Flow Diagram for the Predictive Analytics

- **Data Collection**: Historical crop yield data for **sugarcane, wheat, rice, and maize** was collected from government agricultural reports and institutional databases such as Agricultural statistics at a glance and IITM. Corresponding rainfall data, which serves as the exogenous variable, was compiled for the same years and regions. The selected data includes 20+ years of annual observations to ensure time-based learning.

To gather and preprocess comprehensive datasets including crop yields and rainfall over multiple years for four major crops *(SUGARCANE,WHEAT, MAIZE, RICE).*

1.(Agricultural Statistics AtAGlance) 1950-2016/2020

https://agriwelfare.gov.in/en/Agricultural_Statistics_at_a_Glance

2.(Indian Meteorological Department) 1950-2016/2020

https://mausam.imd.gov.in/

Datasets were pre-processed to make them consistent in time index (e.g., transforming "1950-51" into a single representative year such as "1950") and merged with respect to the year. Handling missing values was done via interpolation techniques, and data was normalized where necessary for machine learning models.This Study was done in the year 2025-26.

- **Data Integration and Preprocessing**: After ensuring chronological alignment, the yield and rainfall datasets were merged using the common "Year" column. Rows with missing values were removed or imputed using linear interpolation. The dataset was further standardized to reduce scale differences across variables. Lag features were engineered to allow models to learn from historical patterns.
- **Feature Engineering**: Lagged variables of yields (up to 4 years) were constructed as input features to capture temporal dependencies. For ARIMAX, rainfall was incorporated as an exogenous input variable. These engineered features allowed models to forecast not just based on recent performance but also based on trends and climate variability.
- **Model Training**: All models—ARIMA, ARIMAX, SVR, Random Forest, and ANN—were trained individually. ARIMA and ARIMAX models were optimized using AIC/BIC criteria to determine the best order parameters. SVR used a linear kernel with fixed C and epsilon values. Random Forest was trained using 100 trees, while ANN was trained for 100 epochs with early stopping.
- **Evaluation Metrics**: Model performance was evaluated on both training and test datasets using RMSE, MAE, and MAPE. These metrics helped to compare prediction accuracy and consistency across models. Train-test splits were done chronologically to avoid data leakage and mimic real-world forecasting.

Predictive Models and Equations

This section outlines the mathematical foundations and practical use cases of the six predictive model used in the study. These models were selected for their effectiveness in time series forecasting, regression capabilities, and ability to capture both linear and nonlinear patterns within agricultural yield data. Each model was trained using preprocessed historical yield and rainfall data, and tested using standard evaluation metrics.

Linear Regression

Linear Regression is a basic yet powerful statistical technique used to model the relationship between input features and a target variable. It assumes a straight-line relationship and is ideal for baseline comparison. In this project, it is used to forecast crop yields based on lagged historical values.

Equation: $Y = \beta_0 + \beta_1 X + \varepsilon$

- Y is the dependent variable (target output),
- X is the independent variable (input feature),
- β_0 is the intercept,
- β_1 is the slope coefficient,
- ε is the error term

ARIMA

ARIMA is a classical time series model that combines autoregression, differencing, and moving average components. It is effective for datasets with strong temporal patterns. In this project, ARIMA is used to model crop yields without external variables. Equation:

$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta\_q \varepsilon_{t-}\_q + \varepsilon_t$

where:

Y_t = Actual value at time t

φ_i = Coefficients of the autoregressive series

θ_j= Coefficients of the moving average series

ε_t= Random error (or white noise)

ARIMAX

ARIMAX extends the ARIMA model by incorporating external features, such as rainfall, which can influence crop output. By integrating these exogenous variables, ARIMAX is able to model cause-effect relationships beyond the autoregressive nature of time series data. This model was found to be particularly effective for crops like rice and sugarcane, which are highly climate-sensitive.

Equation:

$Y_t = c + \varphi_1 Y_{t-1} + ... + \varphi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + ... + \theta\_q \varepsilon_{t-}\_q + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + \varepsilon_t$

- X_{kt} are exogenous input variables
- All other terms are similar to ARIMA

SVR

Support Vector Regression (SVR) uses kernel functions to model complex, non-linear relationships between input features and target values. It's particularly useful for small- to medium-sized datasets with strong feature dependencies.

Equation:

$f(X) = w^t X + b$

- ε is the margin of tolerance (epsilon-insensitive tube)
- SVR tries to fit as flat a function as possible within ε deviation.

Random Forest

Random Forest is an ensemble technique based on decision trees. It builds multiple trees using bootstrapped datasets and aggregates their predictions. This model performs well with non-linear data and is robust against overfitting. Random Forest consistently delivered the best results across all crops due to its capacity to handle high-dimensional and noisy data.

$y = (1/T) \Sigma h_t(x)$

where:

y = Final predicted value

T = Decision tree count

$h^t(X)$ = Prediction from the t-th tree

Each tree is trained on a random sample (bootstrap) from the data

Artificial Neural Network (ANN)

ANNs simulate the learning mechanism of the human brain by passing input through interconnected layers of artificial neurons. With sufficient training, ANNs can model highly complex and nonlinear relationships. In this study, the ANN model achieved strong performance, particularly for maize and wheat, where multiple input patterns interact nonlinearly.

$y = f(\Sigma w_j \cdot \sigma(\Sigma v_{ij} x_i + b_j) + b)$

where: x_i = Input variables (rainfall, previous yield, etc.)

v_{ij}, = Weights between input and hidden layer

w_j, = Weights between hidden and output layer

σ = Activation function (e.g., sigmoid or ReLU)

f = Output function (e.g., linear or SoftMax)

b_j, b = Biases

Model Comparison

To select the most appropriate model for crop yield prediction, all models that were put into practice—ARIMA, ARIMAX, SVR, ANN, and Random Forest—were tested and compared on the basis of both quantitative measures and qualitative factors. All the models were trained on 80% of data and tested on the other 20%, while assessment was done using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The accuracy measures are defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

All the analyses were conducted in Python (v3.x) with the help of libraries such as statsmodels, scikit-learn, TensorFlow, and Keras

for machine learning and time series modelling, and pandas and matplotlib for data handling and plotting.

Results and Discussion

To evaluate model effectiveness, each algorithm was applied to the same cleaned dataset of historical crop yields and rainfall. The key metrics used were Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), which provide a comprehensive view of both bias and variance in model predictions.

The table below presents the performance of each model on both training and test datasets. From the results, it is evident that Random Forest achieved the lowest RMSE and MAE across most crops, indicating its robustness and predictive power. ARIMAX also performed strongly, particularly in modeling rainfall-dependent crops like rice and sugarcane. ANN produced competitive results with strong generalization ability,

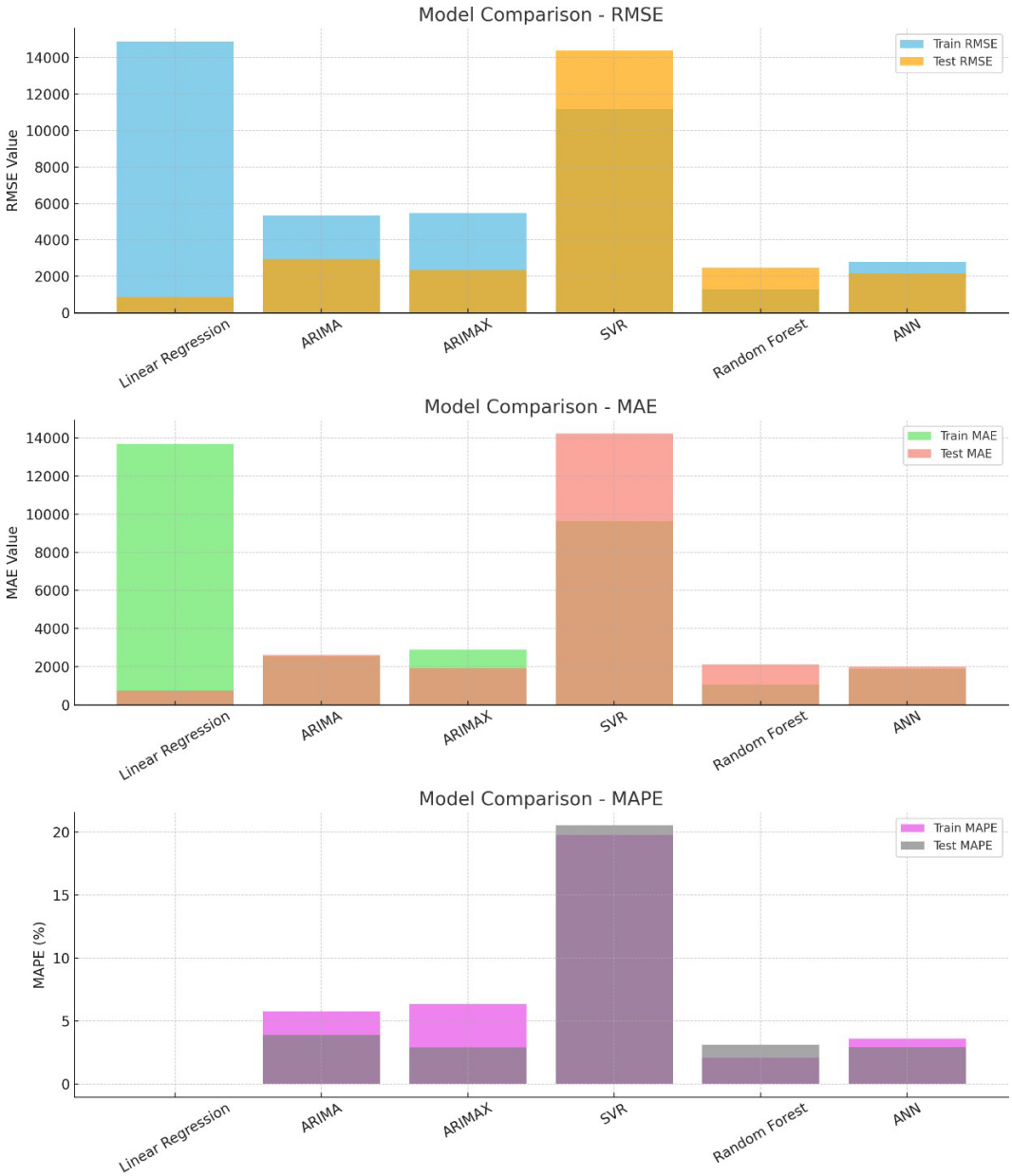

Fig. 2. Model comparison of RMSE, MAE & MAPE

Table 1. Performance Summary of All Models

Model	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MAPE	Test MAPE
Linear Regression	14873.41	861.86	13704.39	752.09	-0.01	0.04
ARIMA	5337.19	2934.0	2546.41	2622.65	5.75	3.88
ARIMAX	5480.1	2355.32	2883.93	1912.21	6.34	2.88
SVR	11182.1	14386.34	9639.84	14215.68	19.76	20.54
Random Forest	1300.39	2470.54	1045.93	2119.9	2.1	3.08
ANN	2782.8	2173.7	1866.15	1999.35	3.59	2.92

while SVR and Linear Regression had comparatively higher error margins.

These results emphasize the importance of model selection based on the nature of the data and complexity of relationships. Machine learning models like Random Forest and ANN, which can capture nonlinearities and interactions, are more suited for agricultural yield forecasting than linear or univariate time-series models.

Based on the comparative analysis, Random Forest and ARIMAX models emerged as the most effective models for crop yield forecasting.

- **Random Forest** achieved the lowest prediction errors in terms of RMSE (Train: 1300.39, Test: 2470.54) and MAPE (Train: 2.10%, Test: 3.08%). This indicates exceptional accuracy and consistency in both training and test datasets. Its ability to handle high-dimensional data and capture complex interactions contributed to its superior performance.
- **ARIMAX** demonstrated strong results in generalizing predictions to unseen data. With a Test MAPE of 2.88%, it closely followed Random Forest. By incorporating rainfall as an exogenous factor, ARIMAX effectively modeled the climatic dependencies crucial to yield outcomes.
- **ANN** also performed well and showed potential in modeling nonlinear patterns, especially in crops like maize and wheat. While not outperforming Random Forest, it proved to be a strong competitor.
- **SVR and Linear Regression**, while useful for benchmarking and comparison, showed higher errors and lacked the complexity handling required for multi-variable agricultural datasets.

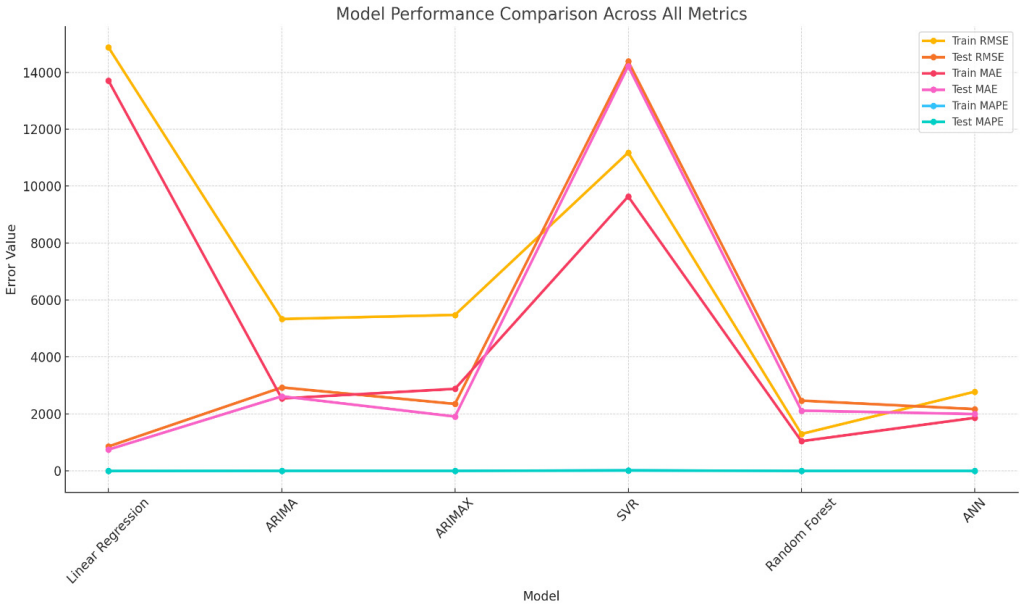

Fig. 3. Model performance comparison across all metrics

Fig. 4. ANN model: Actual vs. predicted yields in maize, sugarcane wheat and rice.

Conclusion

This research successfully developed and evaluated a predictive framework for crop yield forecasting using historical yield and rainfall data. Multiple models, i.e. (ARIMA, ARIMAX, SVR, ANN, and Random Forest) were tested to assess their forecasting power for key crops like sugarcane, wheat, rice and maize. The study applied structured preprocessing, feature engineering and a consistent evaluation framework using metrics such as RMSE, MAE, and MAPE to ensure fair comparison.

Among all the models evaluated, Random Forest emerged as the most robust and accurate model across both training and test datasets. It consistently achieved the lowest RMSE and MAPE values, showing strong generalization and resistance to overfitting. The model's ensemble-based nature and ability to capture nonlinear interactions make it highly suitable for complex agricultural datasets influenced by numerous variables

like climate, rainfall, and seasonal lags.

While ARIMAX also demonstrated competitive performance, particularly by integrating exogenous variables like rainfall, it did not outperform Random Forest in most scenarios. SVR and Linear Regression had higher error rates, and ANN showed promise but was outmatched by Random Forest in consistency. In conclusion, *Random Forest stands out as the best-fit model* for practical deployment in smart agriculture platforms, offering high accuracy, flexibility, and ease of integration into decision-support systems.

References

Alam, N. M., Mazumdar, S.P., Mitra, S., Saha, R., Pandey, S.K., Panwar, S. and Gouranga Kar 2021. Predictive model for fibre yield estimation of tossa jute (*Corchorus olitorius*) in India. *Indian Journal of Agricultural Sciences*, **91**(6): 837-841

Singh, K.N., Singh, K.K., Kumar, S., Panwar, S. and Gurung, B. 2019. Forecasting crop yield through weather indices through LASSO. *Indian Journal of Agricultural Sciences*, **89**(3): 540-544.

Sanjeev Panwar, Anil Kumar, Ranjit Paul, N M Alam, Sonia Tomar, Nitin Kumar and Abhishek Rathore, 2019. An Alternative Method for Yield Forecasting using Weather Indicies Approach and Non-Linear Statistical Modeling. *Indian Journal of Extension Education*, **55**(2): 111-115

Panwar, S., Kumar, A., Singh, K.N., Ranjit Kumar Paul, Bishal Gurung, Rajeev Ranjan, N M Alam and Abhishek Rathore 2018. Forecasting of crop yield using weather parameters– two step nonlinear regression model approach. *Indian Journal of Agricultural Sciences,* **88** (10): 1597-1599

Sanjeev Panwar*, K.N. Singh, Anil Kumar, Bishal Gurung, Susheel Kumar Sarkar, Sivaramane N and Abhishek Rathore, 2017. Pre harvest forecasting of crop yield using non-linear regression modelling: A concept. *Indian Journal of Agricultural Sciences*, **87**(5): 685-689

Paul, R.K. and Garai, S. 2021. Performance comparison of wavelets-based machine learning technique for forecasting agricultural commodity prices, *Soft Computing*, **25**(20): 12857-12873.

Paul, R.K., Vennila, S., Yeasin, M., Yadav, S.K., Nisar, S., Paul, A.K., Gupta, A., Malathi, S., Jyosthna, M.K., Kavitha, Z., Mathukumalli, S.R., and Prabhakar, M. 2022. Wavelet Decomposition and Machine Learning Technique for Predicting Occurrence of Spiders in Pigeon Pea. *Agronomy*, **12**: 1429

Garai, S., Paul, R.K., Rakshit, D., Yeasin, M., Emam, W., Tashkandy, Y. and Chesneau, C. 2023. Wavelets in Combination with Stochastic and Machine Learning Models to Predict Agricultural Prices. *Mathematics*, **11**: 2896.

Singh, A.K., Paul, R.K., Sarkar, A., Yeasin, Md., Sinha, K., Pal, S. and Paul, A.K. 2024. A Novel Ensemble Machine Learning Approach for Forecasting Oilseeds Prices in India. *Economic Affairs*, **79** (4): 978-992.