International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(2): 577-590 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i02.03182

A Machine Learning Model for Crop Yield Prediction Using Remote Sensing Data

Kavita Jhajharia[1]*, Neha V Sharma[2], Pratistha Mathur[1]

[1]Information Technology, Manipal University Jaipur, Rajasthan, India, [2]Data Science Engineering, Manipal University Jaipur, Rajasthan, India. *Corresponding Author's Email: Kavita.jhajharia@jaipur.manipal.edu

Abstract

Precisely estimating crop yields is a critical aspect of agricultural planning, resource allocation, and food security. Satellite data integrated with machine learning algorithms have recently become a potential solution for predicting crop yield at local and global levels. The present study provides detailed investigation of satellite-based crop yield prediction using machine-learning algorithms. The proposed methodology integrates satellite imagery data with precipitation data. We use machine learning algorithms for predictive modelling, random forests, support vector machines, decision trees, linear regression, and k-nearest neighbour. Extensive investigations are conducted to examine the effectiveness of the proposed method. The study employs multi-year satellite imagery and corresponding crop yield data from various agricultural regions to develop predictive models. The models are trained and tested while considering temporal and spatial variations. Model accuracy and reliability are evaluated through performance metrics, including mean absolute error and root mean square error. The study's findings indicate that using machine learning algorithms for satellite-based crop yield prediction yields a significant level of accuracy compared with standard techniques. According to the research conducted, it has been found that among all the methods that were implemented, the support vector machine method has shown better performance. Integrating satellite-based techniques and machine learning algorithms presents a viable and scalable approach to predicting crop yields.

Keywords: Crop Yield Prediction, Machine Learning, Remote Sensing, Satellite Imagery, Support Vector Machine.

Introduction

In India, agriculture is biggest single segment contributor of the economy, 13.05% share of total Gross Domestic Product (GDP) and approximately 55% of the total household are dependent on agriculture (1). In India, agriculture land area is approximately 60.3% of total land (2). In this study for the prediction Rajasthan state has been considered. In Rajasthan state, the agriculture contributes approximately 25% to the state's GDP and employs 65% of the population (3). Agriculture land is 53% of total land where 75% land is rain fed and rest 25% is irrigated land. The population around the globe is projected to reach 9.8 billion in 2050, and population in India is expected to reach 1.7 billion (4). The area of cultivation is limited, but the population is increasing worldwide which leads to unbalance the food supply and demand chain (5). To correspond to the demand and supply of the food without affecting environmental elements prediction of population and crop production in one way to handle the situation. Crop yield estimation in advance is useful in the developing countries where agriculture has major impact on the economy (6). Precise crop yield forecasting is critical for agricultural management, agronomic challenges, international crop commerce, national food policy, and the administration of irrigation and fertilization practices (7). The crop yield prediction at the farm scale assists farmers in making timely decision for forthcoming problems, such as choose different crop or give up a crop at the initial stage of growth (8-10). The researchers working in the respective domain are interested in development of a mathematical model for better prediction with limited data. These mathematical models consider environmental, climatic, soil, area and other datasets to and follow some fundamental protocol. The traditional model with data from surveys and historical knowledge of prior years is valuable for a small field scale land, but it is difficult to estimate for the larger regions or countries. Recent advancement in the technology and data collection, processing and

(Received 13th November 2024; Accepted 13th April 2025; Published 30th April 2025)

storage has been more efficient than before. Machine Learning (ML) has demonstrated its efficacy in data analysis and agricultural research, particularly in the areas of crop classification and yield prediction (11). The utilization of machine learning (ML) when combined with data analysis provides possibilities for enhanced comprehension and exploration of the agricultural domain. Machine learning techniques are known to be particularly effective in handling noisy data and are capable of revealing non-linear relationships. This, in turn, can assist farmers in making informed decisions by providing them with predictive capabilities. The algorithms improve the performance with the increase in the input data, thus, ML can process a huge amount of data and produce useful information (12). Satellites cover large area and provide real time, multi temporal, and multi spectral data (13, 14). Satellite imagery can produce a tremendous amount of data for a crop model by verifying correlation between crop parameters and spectral reflectance. Remotely sensed data produced by satellite is most suitable choice for crop yield prediction because of its repetitiveness, and multi spectral information. The spectral features of plant canopies, allow different remote-sensing platforms to monitor vegetation dynamics and spatial and temporal variability. The identification of plant health and stress using satellite data is based on a significant relation between simple transformations of reflected red and near-infrared radiation. Several such transformations, defined as vegetation indices (VIs), are based on the distinct spectral signature of green vegetation in the red and near-infrared parts of the spectrum and serve as the foundation for quantitative assessment of vegetation state using satellite data.

NDVI (Normalized Difference Vegetation Index) is defined as the difference between the near-infrared (NIR) and red (RED) bands normalised by their sum calculated using equation 1:

$$NDVI = \frac{NIR-RED}{NIR+RED} \quad [1]$$

The primary drawback of NDVI is that it is affected by soil (brightness and colour), the environment (region covered by cloud and cloud shadow), and foliage canopy shadow. Another issue with NDVI is that it soon becomes saturated in dense vegetation (15). This is due to the non-linearity of the NDVI index. NDVI is useful for studying wide regions and getting an approximate indication of photosynthetic activity. Other indices with built-in feedback systems should be utilised for more qualitative examination. The Enhanced Vegetation Index (EVI) is the very commonly used alternative VI that eliminates few of the shortcomings of NDVI. The soil and atmosphere maintain a strong correlation by which reducing one will increase the other. To correct the impact of soil and atmosphere a feedback method was introduced (16). EVI compensates for some canopy background noise and Climatic variable, and it is more delicate in densely vegetated regions (17, 18). The equation 2 describes the EVI:

$$\frac{EVI=G*((NIR-RED)}{(NIR+C1*RED-C2*BLUE+L))} \quad [2]$$

C1 and C2 are coefficients that adjust for atmospheric resistance, G is gain factor and L corrects for soil background. The blue band on satellites rather contains noisy data and does not always have the highest data quality, hence EVI has limited utility.

Soil-adjusted vegetation index, (SAVI) intends to decrease the impact of soil background on the vegetation signal by introducing a soil adjustment factor in the NDVI equation's denominator, (19):

$$\frac{SAVI=((NIR-R)}{(NIR+R+L))*(1+L)} \quad [3]$$

Where: L is an empirically calculated constant that minimises the vegetation index sensitivity to soil background reflectance change. When L is 0, SAVI equals NDVI. L is generally about 0.5 for moderate vegetation cover levels, R is red band and NIR is near infrared. The factor (1 + L) ensures that the SAVI range is the same as the NDVI range, namely [-1, +1].

In addition to NDVI, there is an easy-to-use vegetation index for vegetation identification known as CVI (Chlorophyll Vegetation Index) (20). The CVI is calculated by combining the NIR/green channel SR (Simple Ratio) with the red/green channel SR. This is done to reduce the susceptibility of plants with varying canopies. The CVI is based on two assumptions: 1. the NIR/green channel SR is very sensitive to leaf chlorophyll concentrations in the canopy cover and is unaffected by the LAI (Leaf Area Index); and 2. The LAI has no effect on the NIR/green channel SR. 3. The red/green SR channel indicates the relative density of vegetation and soil and may be used to normalise LAI and reduce the sensitivity of

vegetation structural parameters. The equation 4 describes the CVI:

$$\frac{CVI=NIR}{GREEN*\frac{RED}{GREEN}} \quad [4]$$

In satellite imagery, the Normalized Difference Water Index (NDWI) is utilized to identify open water areas (21). The NDWI index is commonly compared to the Normalized Difference Moisture Index because it appropriately reflects moisture content. In fact, the two are computed and utilised in quite different ways. The NDWI is computed using a visible green and near-infrared combination, allowing it to measure small variations in the water content in leaves and water reservoirs. The equation 5 describes the NDWI:

$$NDWI = \frac{(GREEN-NIR)}{(GREEN+NIR)} \quad [5]$$

The visible green frequencies increase the water surface's usual reflectivity. The near-infrared wavelengths emphasise terrestrial vegetation and soil characteristics while reducing water features' low reflectivity. A higher NDWI value signifies adequate moisture, whereas a low value signifies water stress. The dimensionless NDWI product ranges from -1 to +1 based on the hardwood content as well as the kind of vegetation and cover. High NDWI values imply that water content in the leaf is high. The NDWI rate decreases during instances of water stress. The rapid advancement of Remote Sensing technology establishes a strong technological foundation for the comprehensive use of Indian agriculture data. Satellite remote sensing techniques can offer resource managers with an effective and cost-effective method for obtaining real time data for natural resource development and management. ML provides a variety of strategies for recognising patterns and trends in massive data and has been proven to be significantly predictive. The use of Machine learning advancements in the Indian agriculture industry is quite low, and the majority of accessible data is not even digitised. Much processing is required before it can be used in machine learning applications. The structure of the research paper is outlined as follows:

Section 1 introduces the overview of the research work with importance of Machine learning and remote sensing in agriculture, section 2 furnishes the literature survey to the study. section 3, methods and materials, focuses on the study area, data collection, pre-processing, vegetation index extraction from the remote sensing, and methodologies implemented on the data. Section 4 is focusing on the result and discussion, and section 5 concludes the research. The progress of machine learning algorithms has improved yield estimation (22). In a research study, multi-spectral and multi-temporal satellite images were used to predict crops using a variety of ML models. The comparison of random forest with multiple linear regressions conducted to estimate the crop yield at global scale (23). For model training and testing, authors used agricultural yield data from several sources and provinces: The first three are the gridded worldwide wheat grain yield, the second is the maize production from US, and the third is the potato and maize silage production in the north-eastern coast region. Random Forest was proven to be very competent of forecasting crop yields. The correlation between a number of factors, including yearly rainfall data, cultivation area, and food price, and associated impacts on rice crop production is conduced and is identified using regression analysis (24). The Results indicate that moderate variation in data features, each of which is clearly relevant to crop productivity. Implementing the Regression Analysis yields the affected value R2 = 0.7. This R2 result demonstrates conclusively that all the data variables have an average effect of 70% on crop yield. The study could be expanded by including the minimum support price, weather conditions, soil parameters, and others that affect crop yields, as well as using different data mining and statistical methods to examine yield variables. Sequential minimal optimization (SMO) classifier was used on the Maharashtra state dataset (25). The Indian Government's openly accessible data were the source of the dataset used to forecast the rice crop production. The study focuses on all the important factors for crop yield prediction for kharif season for four years. Various validation matrices were used in the research to validate the results. The study compares SMO approach with other techniques and the findings demonstrate that other implemented techniques performed better than SMO classifier. Rainfall, temperature, area and season were considered as features to predict the crop in India (26). Linear regression, random forest, artificial neural networks, logistic regression and XGBoost were implemented and among all random forest perform better than all the algorithms. A study primarily emphasis on

estimation of significant kharif crops in Visakhapatnam, Andhra Pradesh (27). In this study, researchers primarily forecast the volume of rainfall using Memory Augmented Neural Networks (MANNs), and afterwards, utilising the rainfall data and the region allocated to that specific crop, authors predict the volume of main kharif crops produced using support vector regression (SVR). Appropriate agricultural strategies may be developed by utilising the MANNs-SVR technique in order to boost crop production. According to the comparison, the suggested technique beats previous machine learning algorithms in forecasting the kharif crop production. Another research was conducted on the same objective and results indicates that random forest performs better on temperature, rainfall, humidity, pH., and crop name as features (28). Machine learning algorithms were used for soil classification, crop yield prediction and fertilizer recommendation using 5 years data of multiple crops (29). Random forest performed better for soil classification with 86% accuracy and support vector machine provided better results for crop yield prediction with 99% accuracy. In a study, authors used weather data, soil data, agro-management data and crop data from 2016 to 2019 of study region Shinkiari to predict the crop (30). Authors implemented machine learning algorithms and Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were utilized to evaluate the performance of the models. Among all the applied models XGBoost regressor performed better with RMSE as 0.15 t/ha. Principal component method was used to examine the impact of climate change on strawberry crop (31). The evaluation of connections between climatic characteristics and strawberry production can provide valuable information as well as timely demonstration of yield predictions that can be used for the benefit of strawberry growers. The findings of the experiments demonstrated a significant association between climate variables and strawberry yield, providing a foundation for yield prediction with a lead time of three to five months. In order to provide assistance to farmers introduced a crop yield prediction model that makes use of data mining methodologies known as Random Forest and regression tree (32). The approach that has been suggested is effective for forecasting sugarcane disease far in advance and is also helpful in providing suggestions to improve crop productivity. Authors discovered that the projected results from the RF algorithm are extremely near to the real values, and when compared to decision tree, RF gives better results in forecasting sugarcane crop disease. These observations were based on the findings of the experiments that were carried out. Remote sensing is an efficient tool for determining the identification, traits, and growth potential of essentially all crops (33, 34). Crop behaviour is determined by the crop's nature, interactions with solar radiation and other climatic conditions, and the presence of chemical nutrients and water inside the host medium (35). NDVI, EVI and land surface temperature (LST) were used as data features and implements machine learning and deep learning algorithms at municipality level for soybean yield prediction (36). The research identified the prediction accuracy of the model in middle of the season. The results show that long short-term memory outperforms all the other models with mean absolute error of 0.42 mg/ha. The study describes the integration of satellite and climate data improves the prediction accuracy. A study was conducted on three countries and five different crops data of approximately twenty years (37). The study presents a generic methodology highlighting reusability, accuracy and modularity and tests the workflow on thirteen different case studies. The authors suggested with more data addition and more predictive features can improve the overall prediction framework. Grassland biomass of two intensive agricultural grassland fields was used in Ireland for estimation using multiple models (38). In situ measurements, measured on weekly basis, were utilised for model development for the first analysis area (Moorepark) over a period of 12 years and the second analysis area (Grange) over a period of 6 years. All three machine learning models were provided five vegetation indicators. Model assessment showed that the adaptive neuro-fuzzy inference system (ANFIS) achieved better with RMSE Moorepark = 11.07; RMSE Grange = 15.35 biomass estimation than the Artificial Neural Network (ANN) and Multiple Linear Regression (MLR). Deep learning techniques and remote sensing data were implemented for the development and model execution (39). This

model has been used to forecast the soya bean crop in Argentina and Brazil. The findings of the research indicate that the proposed approach obtained a level of precision in forecast of the soya bean crop yield that is comparable to that reached by the existing methods. Exciting developments have been made in the domain of transfer learning that have the potential to enhance prediction performance in regions that lack complete data. These regions, in particular, stand to benefit from a reliable crop forecast tool that is both affordable and affordable. A study acquired data to forecast irrigation recommendations in addition to monitoring and regulating the crop (40). A dataset was specifically created by combining information gathered from soil-sensors distributed across four primary plots, by meteorological department, and from physical irrigation records. On this dataset, several regression and classification techniques were employed to train models that could predict the weekly irrigation pattern suggested by the agronomic. To identify the factors that consistently increased prediction accuracy, eight distinct subsets of parameters were used in the model development process. Gradient Boosted Regression Trees, which had an accuracy of 93%, and Boosted Tree Classifier, which had an accuracy of 95%, were found to be the top regression and classification models, respectively (on the test-set). Moreover, data that weren't improving the model's prediction rate were highlighted. The resultant model can make irrigation planning more easily for agronomists. A study explores ability to estimate agricultural productivity is crucial for global food security, and crop price forecasting can help farmers escape price crash (41). This research examines the application of remote sensing data and deep learning algorithms to forecast agricultural yields and farmer pricing. It is discovered that the introduced ensemble of Convolutional Neural Network- Long Short-Term Memory (CNN-LSTMs) is the good at predicting annual soybean yields. With a 31% improvement in average Root Mean Square Error, it exceeds methods described in the literature (RMSE). Crop yield prediction is influenced by various agronomic factors such as soil characterises, irrigation practices, and pest management. Machine learning models can improve predictive accuracy by integrating these factors. Environmental factors such as climate change and sustainable resources utilization are crucial to ensure long-term sustainability. The combination of data containing multiple satellite indices, environmental data, and temperature data to estimate the crop in China is also used in the studies (42). When developing models for predicting yield, one linear regression approach, two machine learning (ML) methods, and three ML methods were used. According to the findings, the individual machine learning approaches performed better than the linear regression methods. Furthermore, the ML ensemble model increased the performance of the single ML models. In addition, models with a greater number of inputs had superior performance. Additionally, the amalgamation of remote sensing data and environmental data demonstrated stronger yield prediction ability than separate inputs.

Methodology

Research has shown that machine learning techniques can be highly effective in dealing with noisy data and are capable of revealing non-linear relationships. Machine Learning (ML) has been found to be highly effective and widely adopted for predicting crop yield. It is hypothesized that there exists a significant correlation between climate data and satellite data, with potential overlapping effects. In this study, satellite and climate data features were integrated to predict crop yields in Rajasthan, India. The study utilized a variety of sources to gather data, such as annual crop plantation area, crop yield data, climatic data, and satellite data. To gain a comprehensive understanding of the data, it was necessary to integrate the different sources of information.

Figure 1: Area of the Study (43)

Study Area

The present study concentrated on predicting wheat yield in the Rajasthan region extending from 27° 23' 28.5972'' N and 73° 25' 57.4212'' E, (Figure 1) which contribute over 7.49% of the country's wheat production. Wheat is usually sown in the winter (November–December) and harvested in the spring (March–April) (31) (Figure 1). The region relies significantly on precipitation for agriculture, with 6.661 million hectares irrigated and 11.688 million hectares rainfed, with an average rainfall of 56 cms, fluctuating between 15 and 90 cms (44, 45).

Data Sources

Satellite captures the data beyond the range of the human eye, and we gain access to other information in the other ranges of electromagnetic spectrum like infrared. Satellite interprets or captures these ranges and opens the unseen secrets of earth for the better interpretation. The true state of crop becomes evident with the help of vegetation indices. Agriculture domain is one of the main consumers of earth remote sensing services which uses space monitoring techniques and unmanned devices.

Five distinct vegetation indices were obtained from the satellite data, namely the NDVI, EVI, SAVI, CVI, and NDWI. These indices were utilized to estimate the dynamics of above-ground vegetation in relation to biomass and photosynthesis. The five vegetation indices (VIs) were obtained from Landsat 7 and Landsat 8 collection 1. The spatial resolution of the data was 30m, and the temporal resolution was 16 days. The utilization of Landsat 8 satellite enhances the potential of acquiring cloud-free data across the globe. The characteristics of both the satellites have been described in table 1. The band comparison of both the satellites is represented in Figure 2.

Table 1: Satellite Landsat 7 and Landsat 8 Characteristics

Satellite	**Landsat 7 and Landsat 8**
Sensors	Operational Land Imager sensor, Thermal Infrared Sensor and Enhanced Thematic Mapper Plus
Cycle	16-day repeat cycle
Resolution	30-metre
Bands used	Near Infrared, Blue, Red, Green

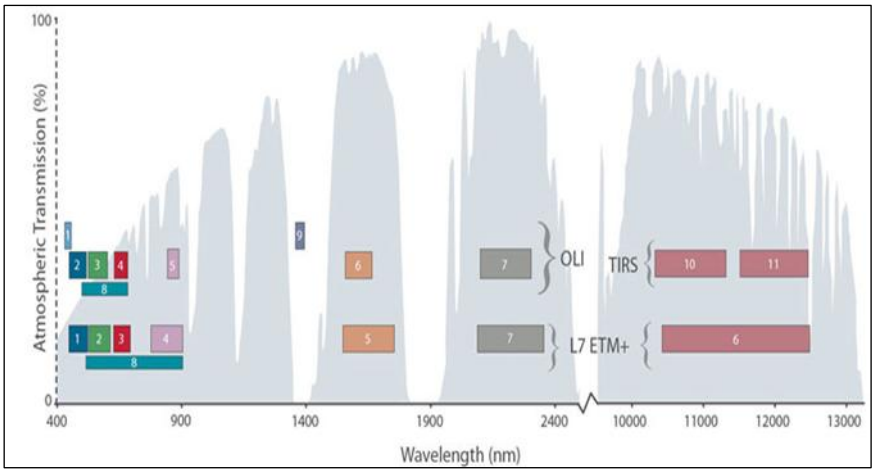

Figure 2: Band Comparison of the Landsat 7 and Landsat 8 (46)

The research focuses on the rabi season in Rajasthan, which spans from October to March. During this period, the cultivation of essential crops such as wheat, barley, mustard, and pulses takes place, which plays a crucial role in shaping the agricultural landscape of the region. Rainfall data is an essential factor for crop yield in Rajasthan due to arid to semi-humid climate (47). The adverse impacts of global climate change on agricultural output are manifest through extreme temperatures and unusually low rainfall (48). Annual rainfall throughout the state differs considerably. The majority of Rajasthan's monthly precipitation falls between July and August. We identified the rainfall for the prediction of crop yield. We acquired the data from https://www.indiawaterportal.org/met_data for each district from 2010 to 2020. Table 2 describes the sample features of data.

Table 2: The Final Integrated Data with Satellite and Statistical Data

District Name	Crop Year	Season	Rainfall (mm)	NDVI	EVI	SAVI	CVI	NDWI
AJMER	2015	Rabi	412.5	0.2091	0.51	0.4623	0.2134	0.4511
AJMER	2016	Rabi	534.8	0.2801	0.79	0.4817	0.2973	0.4798
AJMER	2017	Rabi	483.4	0.2736	0.65	0.4759	0.2941	0.4674
AJMER	2018	Rabi	426.66	0.2698	0.73	0.4861	0.2843	0.4743
AJMER	2019	Rabi	747.5	0.2842	0.85	0.4986	0.2987	0.4876
AJMER	2020	Rabi	436.5	0.2861	0.21	0.4609	0.2991	0.4532
JAIPUR	2015	Rabi	359.05	0.2791	0.1	0.4862	0.2857	0.4753
JAIPUR	2016	Rabi	555.81	0.2784	0.11	0.4985	0.2833	0.4878
JAIPUR	2017	Rabi	319.9	0.2753	0.29	0.4874	0.2798	0.4763
JAIPUR	2018	Rabi	520.71	0.2832	0.2	0.4991	0.2972	0.4888
JAIPUR	2019	Rabi	688.9	0.2856	0.65	0.5091	0.2987	0.4967
JAIPUR	2020	Rabi	545.7	0.289	0.57	0.5172	0.2992	0.4996
UDAIPUR	2015	Rabi	647.5	0.2843	0.1	0.4809	0.2978	0.4714
UDAIPUR	2016	Rabi	832.54	0.2891	0.12	0.4908	0.299	0.4882
UDAIPUR	2017	Rabi	829.7	0.2866	0.61	0.4841	0.2983	0.4737

The datasets were divided into two parts initially. The split was done for training as well as testing of the system. This also helps the algorithm to run on our less powerful machine with considerable efficiency. The data were split into 70 and 30 for training and testing respectively and the same process was followed to achieve higher accuracy. Therefore, this ratio was used in the final model.

The objective of the study is to predict the crop yield prediction in Rajasthan region using machine learning techniques. Therefore, we have developed a framework for current study region using machine learning algorithms. Cross-validation strategies are utilized for the evaluation of machine learning algorithms. Error evaluation is used to determine the optimal model, where a lower error value implies a better model fit. The machine learning model's performance was evaluated by calculating the Mean Absolute Error (MAE), and Root-mean-square Error (RMSE), between the observed and predicted agricultural yield. RMSE permits evaluating the standard deviation of the error for a typical individual observation as compared to any sort of "total error." In data science, RMSE is used as a heuristic for training models and to assess the accuracy of trained models (49). RMSE is computed using the equation 6, given as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (w_j - \hat{w}_j)^2} \quad [6]$$

Where w indicates predicted value, $\hat{w}$ represents measured value, and n is the predicted values.

MAE is best practice for calculating the average extent of errors in a prediction set excluding their direction. MAE is the mean of test sample's absolute difference between the estimated and actual values of the when all individual differences are given equal weight. MAE is computed using the equation 7, given as:

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |w_j - \hat{w}_j| \quad [7]$$

Where w denotes predicted value, $\hat{w}$ signifies measured value, and n is the predicted values.

Results and Discussion

The forecast of crops is a difficult but essential task for the effective management of available resources. The ability to accurately forecast agricultural yields has significant implications for crop insurance, harvest management, and strategic planning. Since remote sensing techniques provide real time and accurate information about the more extensive land covers, the data gathered through this method is perfect for crop forecasting. The agricultural industry analyses drought stress, categorises cropland cover, and forecasts production with the assistance of data obtained via remote sensing. The yield of the crop is dependent on a number of elements, including the temperature and amount of precipitation experienced throughout the growing season, as well as the management of the soil, the use of fertilisers, and other climate data. Numerous VIs, such as NDVI, EVI, LAI, SAVI, and others, which have been used to predict crop yields, are in addition to the many types of information which are significantly provided by remote sensing. Numerous investigations have shown that remote sensing data can effectively provide quantitative insights into agricultural yield, and enhanced models have been developed to evaluate crop yield with greater precision. The rate of crop growth is a key factor in determining the expected yield. In order to determine the progression of plant growth and development, process-based models take into account a wide variety of agricultural, environmental, and other management methods. Despite this, they do not account for all of the factors that have a substantial statistical impact on crop output and do not reflect all of them. Utilizing satellite images, remote sensing compiles Real time information on agricultural fields. The satellite images do not include any mistakes made by humans, and they are readily available without charge in accordance with open information strategies. However, satellite data only gives indirect views of agricultural yields. Because of this, relying on statistical models to get yield predictions from satellite observations is necessary. As predictors, statistical models make use of climatological parameters as well as the outcomes of the methodologies that came before them. Machine learning is extensively used to extract data from satellite photos for the purpose of making agricultural decisions. These judgments involve both the input variable and the independent variable. Machine learning is a technique that enables a system to acquire the ability to perform a specific task by learning from the input data provided. The system can learn to generate the desired output based on the input data it receives. In contrast to simulation crop models, the present study aims to investigate a different approach. Machine learning has many advantages over traditional approaches, including a shorter runtime, the need for less data storage, and the absence of a requirement for prior expert knowledge. ML begins by extracting the features at the beginning of the process, and then uses those characteristics to achieve various tasks such as crop prediction, categorization, or weed detection.

In present study, we experimented using remote sensing, and climate data using machine learning algorithms for prediction of crop yield. Results of the experiment are presented and discussed in this chapter. To predict the crop yield using satellite data in the Rajasthan region and analyze vegetation categories using a calculated vegetation index. In the research work, the authors implemented machine learning techniques, Linear Regression, Random Forest, Decision Tree Support Vector Machine, and K-Nearest Neighbor, along with remote sensing data for feature extraction. For vegetation index calculation, the satellite images were processed in the form of raster data with band wise and with the assistance of coordinate reference system we transformed the raster data into a matrix. The bands were further converted into float value to get the numerical value which produces the pixel wise matrix. We have used 5 X 5 block and the maximum value of the pixel from block was selected. In case of invalid data warning were generated. The process has been briefed in the Figure 3.

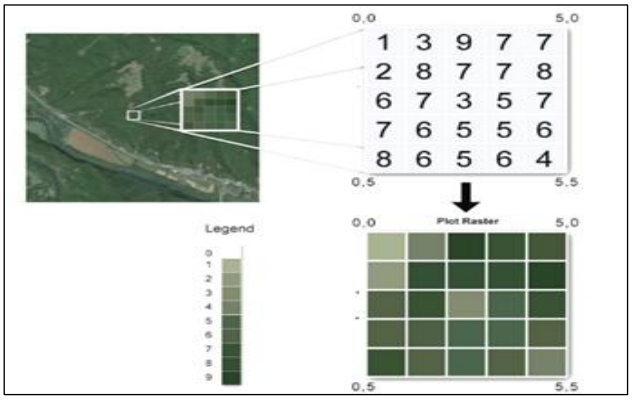

Figure 3: Grid Wise Representation of Raster Data Processing (50)

The data frames of vegetation indices were analysed by processing the satellite images for each month individually. The vegetation indices values for each month were compiled into a single data frame, followed by data classification to gain insights into the distribution of crop health.

Proposed Framework

The proposed framework indicating the process of the study, the first step determines the data collection from multiple sources. The satellite images are used to extract the vegetation indices; the extracted vegetation indices are normalized. Identified machine learning algorithms were implemented on the various sets of input combinations of climate and satellite data. The execution of the model was evaluated using MAE and RMSE metrics. The top performing model amongst the entire implemented model is identified. The Proposed framework predicts crop yield which is based remote sensing data of Rajasthan state in India.

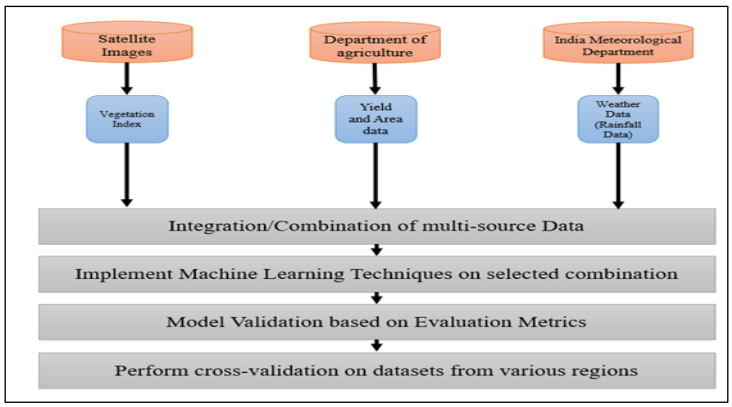

Figure 4: Flow Diagram to Predict Crop Yield Using Multisource Data

The designed framework is trained using support vector machine and generates improved results than other machine learning algorithms. To investigate the ability of framework Haryana state of India has been identified as the validation region. The data of Haryana region was collected from 2013 to 2017 from http://data.icrisat.org/dld/src/crops.html for validating the proposed framework (51). Satellite images were used to extract the vegetation indices, such as the NDVI, EVI, SAVI CVI, and NDWI, and all the VIs were combined with the rainfall data of the same region. Figure 4 describes the process of crop yield prediction using multi source data. In this study Linear regression, Decision tree, random forest, K NN and support vector machine methods are implemented. The performance of all the models was compared with the help of evaluation metrics RMSE and MAE. MAE has no preference for either little or large faults. RMSE is utilised in situations in which it is permissible to accept little errors, but significant errors must be penalised and reduced to the greatest extent possible. While implementing the models the data was split for testing and training the model, 70% of the was used for training and 30% data for testing. In decision tree the maximum depth was selected as three whereas three in random forest and number of trees were hundred. Support vector regression used radial basis function kernel, the tuneable parameter, epsilon, was set to 0.1, and the Regularization parameter was ten. Table 3 shows the comparison of all the models performance. Decision tree regression has the lowest performance with 0.30 t/ha RMSE and 0.23 t/ha MAE, whereas support vector machine performs better than all the other model with 0.26 t/ha RMSE and 0.19 t/ha MAE.

The Figure 5 represents the comparison of all the models using bar graph which discovers that random forest and KNN have performed very well and the RMSE and MAE values are very close to SVM. Figure 6 is the line graph representation of the results obtained on vegetation indices extracted from satellite imagery. RMSE and MAE are represented using blue and orange color.

Table 3: Performance Comparison of Machine Learning Algorithms for Crop Yield Prediction

Algorithm	RMSE	MAE
Linear Regression	0.272	0.216
Decision Tree	0.303	0.234
Random Forest	0.273	0.213
K-Nearest Neighbour	0.276	0.205
Support Vector Machine	0.268	0.195

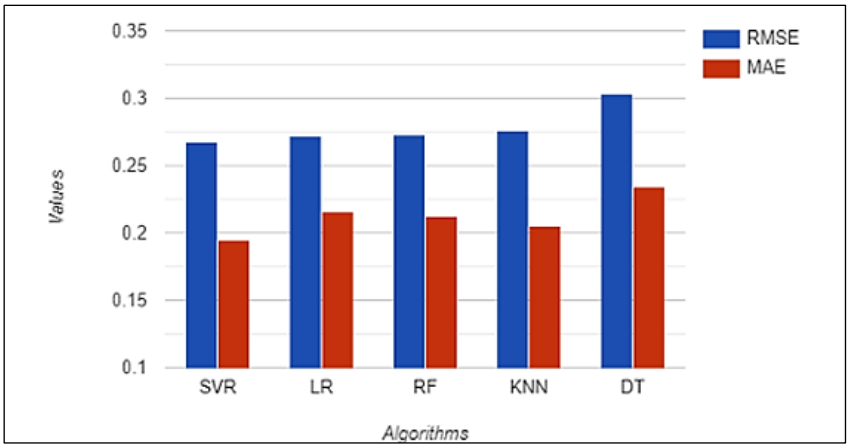

Figure 5: The RMSE and MAE Evaluation Metrics Value of All the Machine Learning Models

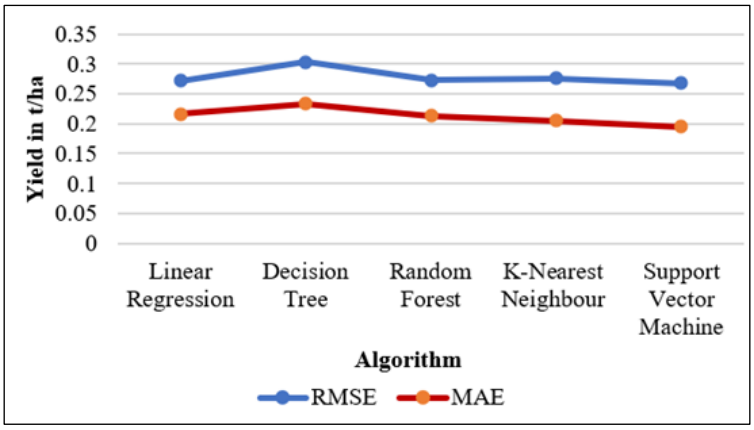

Figure 6: Representation of the Results Obtained on Vegetation Indices Extracted from Satellite Imagery

As using satellite images that are of great size, more computational time is required, Overall, Support vector regression outperforms all of the other machine learning methods with RMSE of 0.26 t/ha. Figure 6 illustrates the comparison of all the implemented algorithms. Five machine learning algorithms were implemented to predict the crop yield of Haryana region. Random forest and support vector machine showed better results than other algorithms, however overall support vector machine gave promising results in Haryana state dataset. Random forest presents RMSE as 0.31 t/ha, and MAE as 0.25 t/ha whereas support vector machine provides RMSE as 0.28 t/ha and MAE as 0.24 t/ha. So overall, support vector machine performs better than all the other techniques. The experiment shows that as we add more data the performance of the model can be increased, but the climatic condition of each state/region is different which affects the bands value and overall predict. Table 4 shows the framework performance on Haryana data.

Table 4: The Machine Learning Model Performance on Haryana Dataset

Algorithm	RMSE	MAE
Linear Regression	0.47	0.38
Decision Tree	0.32	0.26
Random Forest	0.31	0.25
K-Nearest Neighbour	0.34	0.29
Support Vector Machine	0.28	0.24

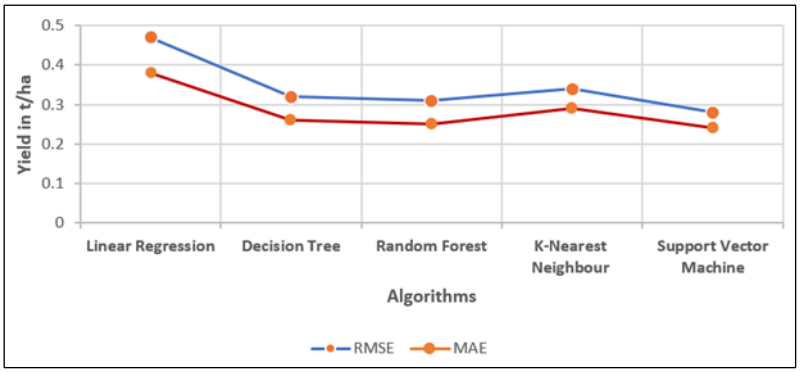

Figure 7: Line Graph of Performance Comparison of Machine Learning Algorithms

The Figure 7 shows the implementation of all the applied machine learning models. To assess the efficacy of the proposed machine learning models, the study region was changed, Rajasthan to Haryana State for result validation. The varied climatic and soil conditions of Rajasthan offered a strong basis for training, whereas the agricultural environment of Haryana acted as a testing ground for model generalization. The findings demonstrated that the Support Vector Machine (SVM) consistently surpassed other techniques in both states. The current study may support data

driven decision for the various stakeholders in agriculture. The farmers can take advantage from the predictive insights of the model for planting planning and harvesting schedule. The agribusinesses may utilise the yield prediction for supply chain management and additionally policymakers can incorporate this approach into food security frameworks to enhance resource allocation. Although, machine learning algorithms have potential and has given promising results in the agriculture domain it still faces challenges in practical implementation. The unavailability of high-quality datasets especially on small scale farms may impact the model's ability to predict.

Conclusion

In India, agriculture employs over half of the population and contributes 15–16% of the GDP. In the study, crop yield prediction was performed using remote sensing data using five machine learning techniques, including, linear regression, decision tree, random forest, K nearest neighbour and support vector machine. The authors focused on distinguishing regions of a satellite image and then finding normalised difference vegetation index and improved vegetation index values. They found that vegetation of the same crop field fluctuates throughout the months and that we may anticipate crop development in that area by computing. The results report that support vector machine performs better compared to other techniques. The support vector machine performed better with 0.268 t/ha RMSE and 0.195 t/ha MAE. The results showed that machine learning methods can use data from more than one source to make accurate yield estimates. Machine learning models can efficiently extract data, however due to their internal black-box, model uncertainty increases. So, in the future, researchers might be able to make more accurate predictions by combining machine learning with a model of crop growth. It has also been suggested that a more extensive collection of characteristics might be learned during the process of modelling; for instance, in addition to remote sensing data and climate data, soil data can be employed to train the machine learning model in order to achieve a higher level of precision. The present work can be extended in the future, and additional data such as soil properties, pesticides data, satellite extracted sun induced chlorophyll fluorescence data may enhance the model's performance.

Abbreviations

ANN: Artificial Neural Network, CNN: Convolutional Neural Network, CVI: Chlorophyll Vegetation Index, EVI: Enhanced Vegetation Index, GDP: Gross Domestic Product, K-NN: K-Nearest Neighbor, LAI: Leaf Area Index, LST: Land Surface Temperature, LSTMs: Long Short-Term Memory, MAE: Mean Absolute Error, ML: Machine Learning, MLR: Multiple Linear Regression, MSE: Mean Squared Error, NDVI: Normalized Difference Vegetation Index, NDWI: Normalized Difference Water Index, NIR: Near-Infrared, RMSE: Root Mean Squared Error, SAVI: Soil-Adjusted Vegetation Index, SMO: Sequential Minimal Optimization, SVR: Support Vector Regression, VI: Vegetation indices.

Acknowledgment

We extend our gratitude to Manipal Jaipur University for generously providing the resources that made this study possible.

Author Contributions

Kavita Jhajharia: Conceptualization, Methodology, Formal Analysis, writing—original draft preparation, Visualization, Pratistha Mathur: Validation, writing—review and editing, Supervision, Neha V Sharma: Data curation, writing—review and editing.

Conflict of Interest

The authors declare that there is no conflict of interest.

Ethics Approval

Not Applicable.

Funding

None.

References

1. Siddiqui K. India's economic reforms and challenges for industrialisation. JES. 2018;6:1–20.
2 Duraisamy V, Bendapudi R, Jadhav A. Identifying hotspots in land use land cover change and the drivers in a semi-arid region of India. Environ Monit Assess. 2018;190(9):535.
3. Parmar I, Soni P, Kuwornu J, Salin K. Evaluating Farmers' Access to Agricultural Information: Evidence from Semi-Arid Region of Rajasthan State, India. Agriculture. 2019;9:60.

4. Singh AK, M, Bhatt B, Singh K, Upaadhaya A. An Analysis of Oilseeds and Pulses Scenario in Eastern India during 2050-51. JAS. 2012;5:p241.
5. Sridhar A, Balakrishnan A, Jacob MM, Sillanpää M, Dayanandan N. Global impact of COVID-19 on agriculture: role of sustainable agriculture and digital farming. Environ Sci Pollut Res. 2023;30:42509–25.
6. Takahashi K, Muraoka R, Otsuka K. Technology adoption, impact, and extension in developing countries' agriculture: A review of the recent literature. Agricultural Economics. 2020;51:31–45.
7. Xia L, Ti C, Li B, Xia Y, Yan X. Greenhouse gas emissions and reactive nitrogen releases during the life-cycles of staple food production in China and their mitigation potential. Science of The Total Environment. 2016;556:116–25.
8. Tsouros DC, Bibi S, Sarigiannidis PG. A Review on UAV-Based Applications for Precision Agriculture. Information. 2019;10:349.
9. Durai SKS, Shamili MD. Smart farming using Machine Learning and Deep Learning techniques. Decision Analytics Journal. 2022;3:100041.
10. Thilakarathne NN, Bakar MSA, Abas PE, Yassin H. A Cloud Enabled Crop Recommendation Platform for Machine Learning-Driven Precision Farming. Sensors. 2022;22:6299.
11. Cai Y, Guan K, Peng J, Wang S, Seifert C, Wardlow B, Li Z. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sensing of Environment. 2018;210:35–47.
12. Mishra S, Tyagi AK. The Role of Machine Learning Techniques in Internet of Things-Based Cloud Applications. Artificial Intelligence-based Internet of Things Systems. 2022; 1(1):105–135.
13. Jhajharia K, Mathur P. Prediction of crop yield using satellite vegetation indices combined with machine learning approaches. Advances in Space Research. 2023;72(9):3998-4007.
14. Ludwig C, Walli A, Schleicher C, Weichselbaum J, Riffler M. A highly automated algorithm for wetland detection using multi-temporal optical satellite data. Remote Sensing of Environment. 2019;224:333–51.
15. Rouse JW, Haas RH, Schell JA, Deering DW. Monitoring vegetation systems in the Great Plains with ERTS. NASA Spec Publ. 1974; 351:309.
16. Liu HQ, Huete A. A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. IEEE Transactions on Geoscience and Remote Sensing. 1995;33:457–65.
17. Kimball BA, Idso SB. Increasing Atmospheric CO2: Effects on Crop Yield, Water use and Climate. In: Stone JF, Willis WO, Developments in Agricultural and Managed Forest Ecology. 1983; 12(1-3): 55–72.
18. NASA. Global climate change impact on crops expected within 10 years, NASA study finds. Washington (DC): National Aeronautics and Space Administration. 2021; https://www.nasa.gov/earth-and-climate/global-climate-change-impact-on-crops-expected-within-10-years-nasa-study-finds/
19. Huete AR. A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment 1988;25:295–309.
20. Vincini M, Frazzi E, D'Alessio P. A broad-band leaf chlorophyll vegetation index at the canopy scale. Precision Agric. 2008;9:303–19.
21. Gao B. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment. 1996;58:257–66.
22. Fei S, Hassan MA, Xiao Y, Su X, Chen Z, Cheng Q, Duan F, Chen R, Ma Y. UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. Precision Agric. 2023;24:187–212.
23. Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, Butler EE, Timlin DJ, Shim K-M, Gerber JS, Reddy VR, Kim S-H. Random Forests for Global and Regional Crop Yield Predictions. PLoS One. 2016;11:e0156571.
24. Sellam V, Poovammal E. Prediction of Crop Yield using Regression Analysis. Indian Journal of Science and Technology. 2016;9(38):1-5.
25. Gandhi N, Armstrong LJ, Petkar O, Tripathy AK. Rice crop yield prediction in India using support vector machines. International Joint Conference on Computer Science and Software Engineering (JCSSE). 2016; 2016: p. 1–5. https://doi.org/10.1109/JCSSE.2016.7748856
26. Nigam A, Garg S, Agrawal A, Agrawal P. Crop yield prediction using machine learning algorithms. In: Proceedings of the Fifth International Conference on Image Information Processing (ICIIP); 2019; Shimla, India. IEEE; 2019. p. 125–130. https://doi.org/10.1109/ICIIP47207.2019.8985951
27. Khosla E, Dharavath R, Priya R. Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. Environment, Development and Sustainability: A Multidisciplinary Approach to the Theory and Practice of Sustainable Development. 2020;22:5687–708.
28. Kumar YJN, Spandana V, Vaishnavi VS, Neha K, Devi VGRR. Supervised machine learning approach for crop yield prediction in agriculture sector. In: Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES); 2020; Coimbatore, India. IEEE; 2020. p. 736–41. https://doi.org/10.1109/ICCES48766.2020.9137868
29. Bondre DA, Mahagaonkar S. Prediction Of Crop Yield And Fertilizer Recommendation Using Machine Learning Algorithms. Ijeast. 2019;04:371–6.
30. Batool D, Shahbaz M, Shahzad Asif H, Shaukat K, Alam TM, Hameed IA, Ramzan Z, Waheed A, Aljuaid H, Luo S. A Hybrid Approach to Tea Crop Yield Prediction Using Simulation Models and Machine Learning. Plants. 2022;11:1925.
31. Pathak TB, Dara SK, Biscaro A. Evaluating Correlations and Development of Meteorology Based Yield Forecasting Model for Strawberry. Advances in Meteorology. 2016;2016:1–7.
32. Beulah R, Punithavalli M. Prediction of sugarcane diseases using data mining techniques. In: Proceedings of the IEEE International Conference on Advances in Computer Applications (ICACA); 2016; Coimbatore, India. IEEE; 2016. p. 393–6. https://doi.org/10.1109/ICACA.2016.7887987

33. Wei M, Wang H, Zhang Y, Li Q, Du X, Shi G, Ren Y. Investigating the Potential of Crop Discrimination in Early Growing Stage of Change Analysis in Remote Sensing Crop Profiles. Remote Sensing. 2023;15:853.
34. Elmetwalli AH, Mazrou YSA, Tyler AN, Hunter PD, Elsherbiny O, Yaseen ZM, Elsayed S. Assessing the Efficiency of Remote Sensing and Machine Learning Algorithms to Quantify Wheat Characteristics in the Nile Delta Region of Egypt. Agriculture. 2022;12:332.
35. Saranya A, Subhashini R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decision Analytics Journal. 2023;7:100230.
36. Schwalbert RA, Amado T, Corassa G, Pott LP, Prasad PVV, Ciampitti IA. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agricultural and Forest Meteorology. 2020;284:107886.
37. Paudel D, Boogaard H, de Wit A, Janssen S, Osinga S, Pylianidis C, Athanasiadis IN. Machine learning for large-scale crop yield forecasting. Agricultural Systems. 2021;187:103016.
38. Ali I, Cawkwell F, Dwyer E, Green S. Modeling managed grassland biomass estimation by using multitemporal remote sensing data—A machine learning approach. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2016; 10(7):3254-64.
39. Wang AX, Tran C, Desai N, Lobell D, Ermon S. Deep transfer learning for crop yield prediction with remote sensing data. In: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS). 2018; Menlo Park, CA, USA. ACM; 2018. p. 1–5. https://doi.org/10.1145/3209811.3212707
40. Goldstein A, Fink L, Meitin A, Bohadana S, Lutenberg O, Ravid G. Applying machine learning on sensor data for irrigation recommendations: revealing the agronomist's tacit knowledge. Precision Agriculture. 2018;19:421–44.
41. Gastli MS, Nassar L, Karray F. Satellite images and deep learning tools for crop yield prediction and price forecasting. Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2021; Shenzhen. IEEE; 2021: p. 1–8. https://doi.org/10.1109/IJCNN52387.2021.9534388
42. Li Z, Ding L, Xu D. Exploring the potential role of environmental and multi-source satellite data in crop yield prediction across Northeast China. Science of The Total Environment. 2022;815:152880.
43. Rajasthan Map 2023. Rajasthan district map PDF. https://www.importantpdfdownload.in/rajasthan-district-map-pdf/
44. NITI Aayog. New Delhi: Government of India. https://www.niti.gov.in/.
45. India Meteorological Department. Satellite images. New Delhi: Ministry of Earth Sciences, Government of India. https://mausam.imd.gov.in/imd_latest/contents/satellite.php.
46. Poursanidis D, Chrysoulakis N. Remote Sensing, Natural Hazards and the contribution of ESA Sentinels missions. Remote Sensing Applications: Society and Environment. 2017;6:25-38.
47. Feng P, Wang B, Liu DL, Xing H, Ji F, Macadam I, Ruan H, Yu Q. Impacts of rainfall extremes on wheat yield in semi-arid cropping systems in eastern Australia. Climatic Change. 2018;147:555–69.
48. Challinor AJ, Watson J, Lobell DB, Howden SM, Smith DR, Chhetri N. A meta-analysis of crop yield under climate change and adaptation. Nature Clim Change. 2014;4:287–91.
49. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?. Geosci Model Dev Discuss. 2014;7(2):1525–34.
50. National Ecological Observatory Network (NEON). Boulder (CO): Battelle. https://www.neonscience.org/
51. International Crops Research Institute for the Semi-Arid Tropics (ICRISAT). District Level Database (DLD). Hyderabad: ICRISAT. http://data.icrisat.org/dld/