ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Optimizing Crop Yield Predictions through Satellite Data Fusion and Machine Learning

Shilpa Naresh Vatkar[1] and Sujata S. Kulkarni[2]

ABSTRACT

Accurate crop yield estimation is crucial for sustainable agriculture and food security, especially in Maharashtra, where climate variability significantly impacts crop growth. This study utilizes satellite data from MODIS, Landsat, Sentinel-1, and Sentinel-2 to predict the yields of 22 crops across 36 districts. Machine learning models, including Random Forest, Gradient Boosting, and SVM, were evaluated using RMSE, MAE, and R2 metrics. Random Forest outperformed the others, achieving R2 values above 0.70 for all crops, with a peak R2 of 0.93. Incorporating seasonal and permuted feature data further enhanced predictions, demonstrating the efficacy of integrating satellite data and machine learning for agriculture. Keywords: Machine learning, MODIS, Landsat-8, Sentinel-2, Sentinel-1, crop yield, features, vegetation indices.

Article information:

Keywords: Machine Learning, MODIS, Landsat-8, Sentinel-2, Sentinel-1, Crop Yield, Features, Vegetation Indices

Article history:
Received: December 21, 2024
Revised: June 24, 2025
Accepted: September 4, 2025
Published: September 20, 2025
(Online)

DOI: 10.37936/ecti-cit.2025194.260193

1. INTRODUCTION

Analyzing crop yield and production plays a crucial role in agricultural research, as it provides key insights into food security, economic planning, and sustainable development. In this context, “yield” refers to the quantity of crops per unit area, while “production” refers to the total crop volume harvested. Accurately predicting and estimating crop yield [1, 2] has become increasingly important due to the growing global population and the corresponding rise in food demand. Advanced technologies like remote sensing and machine learning [3,4] have significantly enhanced yield forecasting accuracy. Numerous studies have demonstrated the effectiveness of satellite remote sensing data, such as MODIS, Landsat 8 [5-7], Sentinel-1, and Sentinel-2 [7], in monitoring crop growth and assessing yield variability across various regions.

Research highlights the value of integrating climate data, soil properties, and remote sensing information to improve the accuracy of yield predictions, as factors like rainfall, temperature, and extreme weather events can heavily influence agricultural output. Machine learning algorithms are effective for yield estimation as they can model complex, nonlinear crop growth patterns. Various studies used different machine learning models, such as Random Forest (RF), Support Vector Machines (SVM), and neural network architectures like Artificial Neural Networks (ANN) and Deep Neural Networks (DNN). Most of these studies have focused on regional analyses, primarily, mainly due to the availability of regional yield data for model training and evaluation. However, there remains a strong need for more precise yield estimates at the field management level, which is essential for making well-informed crop management decisions and supporting stakeholders such as farmers and policymakers.

Accurate data collection and analysis are fundamental to the success of machine learning algorithms, as both the quality and quantity of data directly affect prediction accuracy. The advent of big data, characterized by its volume, velocity, and variety, reduces variability and offers more detailed insights. The use of multiple datasets from sources such as sensors, social media, and healthcare can enhance the effectiveness of data analyses. These datasets, accessible through APIs or web scraping, can include

[1] The author is with the Department of Electronics and Telecommunication Engineering, K J Somaiya School of Engineering (formerly known as KJ Somaiya College of Engineering), Somaiya Vidyavihar University, Vidyavihar, Mumbai- 400077, India, Email: shilpavatkar@somaiya.edu

[2] The author is with the Department of Electronics and Communications Engineering, Sardar Patel Institute of Technology, Mumbai University, Andheri, Mumbai-400058, India, Email: sujata_kulkarni@spit.ac.in.

static or real-time data. The integration of data from various platforms underscores the importance of data cleaning and preprocessing to ensure high-quality inputs for machine learning models [11]. This study has two main objectives: (i) to evaluate different algorithms for estimating crop yield and production in Maharashtra across different seasons, including Kharif, Rabi, and the entire year [12], and (ii) to perform a selective feature analysis of the machine learning models used. Two additional references [7, 9] are included, where [7] presents seasonal wheat estimation using the vegetation index LAI, and [9] reports wheat yield estimation based on Earth observation, meteorological data, and biophysical models. In contrast, [9] estimates wheat yield using Earth observation data, meteorological variables, and biophysical models. While prior works [7] showed seasonal correlations in winter wheat or rice yield prediction, these studies often focus on a single crop or a limited region. Our study extends this approach by incorporating 22 crops across multiple seasons through the fusion of MODIS, Landsat, Sentinel-1, and Sentinel-2 data.

The structure of the paper is as follows: Section I provides an introduction and a review of the relevant literature. Section II discusses the materials and methods used in the study. Section III delves into crop analysis and prediction, highlighting both the advantages and challenges involved. Section IV outlines the proposed methodologies, followed by a presentation of the experimental results and their discussion. Finally, Section V offers conclusions and recommendations for future research.

1.1 Related Work

Recent studies demonstrate the effectiveness of machine learning models—particularly CNN, RF, and SVR—for tasks such as vegetation cover estimation, crop classification, and yield prediction using multi-source satellite data. These works highlight the increasing accuracy and scalability of ML approaches in both micro-level and regional agricultural monitoring. Although RF and SVM have been widely applied [11, 20], their performance often varies depending on feature dimensionality and crop diversity. Unlike [11], which relied solely on NDVI, our model incorporates up to nine vegetation indices and examines feature permutations (e.g., NDVI+SAVI, SAVI+MSAVI) to address challenges such as sparse vegetation and soil brightness. However, relatively few studies have critically investigated how different feature combinations influence multi-crop, multi-season yield prediction, as demonstrated in Table 1.

Table 1: *Comparison of Recent Studies utilizing Machine Learning Techniques for Agricultural Monitoring .*

References	Methodology Used	Remarks
[4]	Compared ML models (RF, SVR, etc.) for estimating vegetation fractional cover using Sentinel-2 and drone data	Demonstrated higher accuracy when combining drone and satellite data; highlighted the suitability of ML for sub-field level vegetation monitoring
[8]	CNN-based crop type classification using combined Sentinel-1A (SAR) and Sentinel-2 (MSI) data	Achieved high classification accuracy; emphasized advantage of fusing SAR and optical data for in-season crop detection
[13]	Multiple ML algorithms (RF, SVM, ANN) used for generalized crop prediction	Found RF to outperform other models; underlined importance of algorithm selection and preprocessing in diverse agricultural settings
[14]	Multi-factorial analysis with ML for vegetation dynamics using satellite data	Offered scalable insights into regional vegetation trends; validated usefulness of ML for macro-level crop monitoring under environmental variations

2. MATERIALS AND METHODS

Maharashtra, a key agricultural state in western India, comprises 36 districts, including Pune, Nagpur, Nashik, and Kolhapur. The state's diverse agro-climatic zones ranging from the coastal Konkan region to the semi-arid areas of Marathwada and Vidarbha support the cultivation of a wide variety of crops. During the Kharif season (June to September), farmers primarily grow rain-fed crops such as rice, soybean, cotton, jowar, and tur. The Vidarbha region is particularly known for cotton and soybeans cultivation, whereas the Konkan area is known for rice production, which benefits from abundant monsoon rainfall.

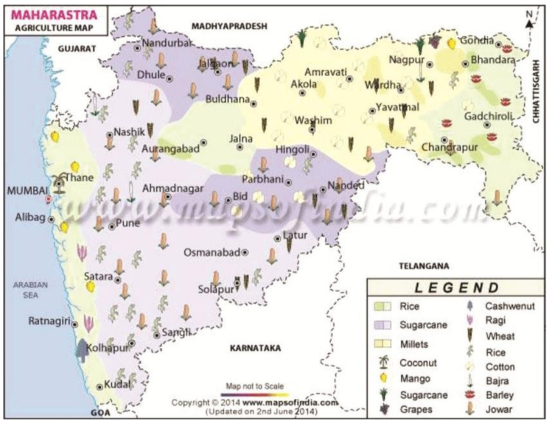


Fig.1: *Major crops in Maharashtra (Source: Maps of India) [32].*

Data on crop distribution and agro-climatic zones have been collected from official resources such as [32] and [12] is demonstrated in Fig.1. During the Rabi season, which spans from October to March, crops like wheat, gram, barley, and safflower are grown, taking advantage of the residual soil moisture from the monsoon. Key districts for wheat cultivation include Nashik, Pune, and Ahmednagar. In addition, sugarcane, a crop that is cultivated year-round, holds significant economic importance for the state, with primary production concentrated in districts such as Kolhapur and Ahmednagar.

Crop yield prediction in Maharashtra is challenged by diverse agro-climatic zones, heavy dependence on the monsoon, and various socio-economic constraints. Yield variability is further influenced by unpredictable rainfall, frequent extreme weather events, and fragmented landholdings. Sparse vegetation and soil brightness complicate the application of vegetation indices such as NDVI, while temporal gaps in satellite imagery—caused by cloud cover or revisit intervals—lead to incomplete datasets. Although socio-economic and climatic factors such as rainfall and market dynamics are recognized as important challenges, this paper focuses on remotely sensed data due to its availability and resolution consistency. Future studies should integrate climate and socio-economic data layers to achieve more holistic modeling.

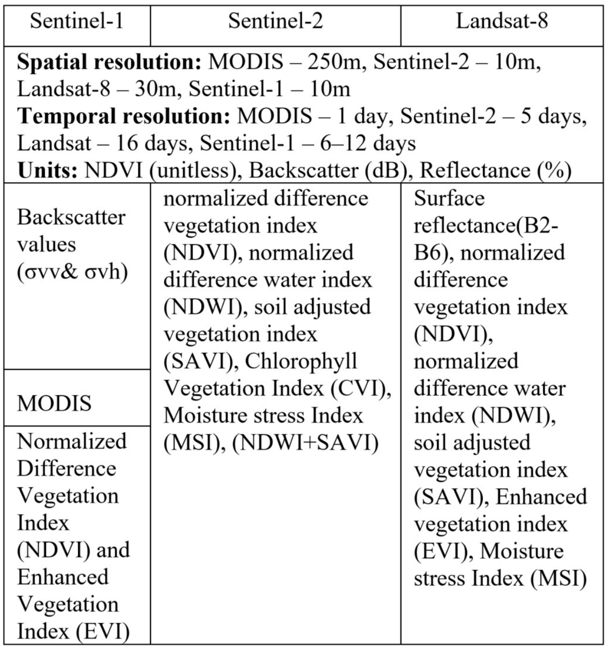
Moreover, the high diversity of crops each requiring tailored approaches along with socio-economic factors such as limited irrigation and frequent pest outbreaks, further increases the complexity of yield prediction. Market and policy influences, including minimum support prices, also shape cropping patterns but remain difficult to model. These challenges highlight the need for advanced machine learning models and the integration of multi-source satellite data to address spatial, temporal, and environmental variability, thereby enabling more accurate and actionable predictions for agriculture in Maharashtra.

2.1 Dataset Description

The proposed research utilizes satellite images acquired from sources such as MODIS, Landsat 8, Sentinel-1, and Sentinel-2. MODIS [13] provides extensive daily data on crop growth, while Landsat 8 offers high-resolution imagery every 16 days, which is valuable for monitoring crop stress and irrigation practices. Sentinel-1, equipped with radar technology, and Sentinel-2, which uses multispectral imaging, enable continuous observation of crop health, soil moisture, and vegetation conditions across key districts. Satellite-based data (e.g., MODIS, Landsat, Sentinel-1, Sentinel-2) provided consistent and reliable inputs, such as vegetation indices and soil moisture metrics. These data sources are freely available, making them accessible and cost-effective for researchers, governments, and organizations globally. Their varying spatial, spectral, and temporal characteristics complement each other, enabling comprehensive crop monitoring. MODIS for large-scale trends and frequent updates. Landsat and Sentinel-2 for high-resolution field-level analysis. These satellites have been extensively validated in crop yield prediction models and agricultural studies, making them a reliable choice. These satellite images have significantly enhanced yield prediction and crop management strategies for key crops such as rice, cotton, soybean, and sugarcane across the state. The collected datasets are summarized in Table 2, with imagery from Sentinel-1, Sentinel-2, MODIS, and Landsat 8 accessed via the Google Earth Engine code editor [14–16]. While commercial satellites such as WorldView or PlanetScope may provide higher resolution or additional features, their higher costs and limited availability make them less suitable for large-scale or budget-constrained applications.

In this research, the proposed approach considers a wide range of performance indicators, including crop season, crop type, the Normalized Difference Vegetation Index (NDVI) [14] , Normalized Difference Water Index (NDWI), Soil Adjusted Vegetation Index (SAVI), Modified Soil Adjusted Vegetation Index (MSAVI), Chlorophyll Vegetation Index (CVI), Moisture Stress Index (MSI), surface reflectance values [15,16], Enhanced Vegetation Index (EVI) [4,18], and backscatter values [17,19]. These factors are crucial for improving crop yield, although additional elements such as temperature, humidity, and soil moisture also play essential roles.

Table 2: *Vegetation indices and other performance parameters extracted from Sentinel-1, Sentinel-2, Landsat-8, and MODIS for crop yield prediction.*



Sentinel-1	Sentinel-2	Landsat-8
Spatial resolution: MODIS – 250m, Sentinel-2 – 10m, Landsat-8 – 30m, Sentinel-1 – 10m **Temporal resolution:** MODIS – 1 day, Sentinel-2 – 5 days, Landsat – 16 days, Sentinel-1 – 6–12 days **Units:** NDVI (unitless), Backscatter (dB), Reflectance (%)		
Backscatter values (σvv& σvh)	normalized difference vegetation index (NDVI), normalized difference water index (NDWI), soil adjusted vegetation index (SAVI), Chlorophyll Vegetation Index (CVI), Moisture stress Index (MSI), (NDWI+SAVI)	Surface reflectance(B2-B6), normalized difference vegetation index (NDVI), normalized difference water index (NDWI), soil adjusted vegetation index (SAVI), Enhanced vegetation index (EVI), Moisture stress Index (MSI)
MODIS		
Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI)		

1) **Normalized Difference Vegetation Index (NDVI):**
The NDVI measures vegetation health and density by comparing reflectance in the near-infrared and red bands. Higher NDVI values indicate healthier and denser vegetation, which often correlates with higher crop yields [21]. It is particularly useful for monitoring dense-canopy crops such as rice, wheat, and maize, providing a direct measure of vegetative vigor that is essential for yield assessment during peak growth stages.

$$\mathrm{NDVI} = (R_{nir} - R_{red})/(R_{nir} + R_{red}) \tag{1}$$

R_{nir} and R_{red} are the averaged reflectance among the waveband range to match MODIS data in the near-infrared (841–876 nm) and red (620–670 nm) wavelengths, respectively.

2) **Normalized Difference Water Index (NDWI):**
The NDWI assesses vegetation water content by comparing reflectance in the near-infrared and shortwave infrared bands. It is particularly valuable for monitoring water stress, which can significantly affect crop growth and yield [21]. NDWI is especially important for water-sensitive crops such as sugarcane and soybean, as it helps detect stress conditions that directly influence biomass accumulation and yield formation.

$$\mathrm{NDWI} = (R_{nir} - R_{swir})/(R_{nir} + R_{swir}) \tag{2}$$

3) **Soil Adjusted Vegetation Index (SAVI):**
The SAVI, similar to NDVI, adjusts for soil brightness, making it effective in sparsely vegetated areas. It improves yield predictions for crops such as tur, groundnut, and bajra under dryland and early growth conditions [21]. By reducing soil background effects, SAVI provides a more accurate measure of vegetation health.

$$\mathrm{SAVI} = (1{+}0.5)(R_{nir}{-}R_{red})/(R_{nir}{-}R_{red}{+}0.5) \tag{3}$$

4) **Chlorophyll Vegetation Index (CVI):**
The CVI measures chlorophyll content in plants, which is directly linked to photosynthetic activity and overall plant health. Higher CVI values indicate better crop condition and potentially higher yields [21]. It is beneficial, handy for crops such as cotton, maize, and sunflower, as it captures chlorophyll levels that reflect photosynthetic efficiency and crop vigor.

$$\mathrm{CVI} = R_{nir} * R_{red}/R_{Green}^2 \tag{4}$$

5) **Moisture Stress Index (MSI):**
Indicates moisture stress levels by comparing different spectral bands. It helps in understanding how moisture stress affects crop growth and yields [21]. MSI supports prediction accuracy in crops like safflower, jowar, and pulses, which are often grown in semi-arid regions and are highly sensitive to moisture fluctuations.

$$\mathrm{MSI} = R_{swir}/R_{nir} \tag{5}$$

6) **Enhanced Vegetation Index (EVI):**
The EVI is a robust vegetation index that provides greater precision in dense canopy areas by accounting for atmospheric and soil variability, thereby improving yield predictions [21]. It is particularly well suited for high-biomass crops such as sugarcane, where NDVI tends to saturate.

$$\text{EVI=2.5*}((R_{nir}R_{red})/(R_{nir} + 6 * R_{red} - 7.5 * R_{blue} + 1)) \tag{6}$$

7) **(NDWI + SAVI):**
Combining NDWI and SAVI [4] provides a more comprehensive assessment of vegetation health and moisture status by integrating both water and soil adjustments. This approach is particularly effective for detecting water stress in sparsely vegetated fields.

8) **Surface reflectance:**
Surface reflectance values from Landsat 8 Bands B2 to B6 are essential for crop yield estimation [15,16]. B2 (Blue) detects crop stress and water bodies, B3 (Green) tracks vegetation health, and B4 (Red) is critical for NDVI-based biomass monitoring. Band B5 (NIR) assesses canopy density, while B6 (SWIR1) detects soil moisture and water stress. Together, these bands provide key insights into crop growth and stress, supporting more accurate yield prediction and efficient resource management. They are particularly effective for monitoring crops such as groundnut, sesame, and maize, enabling stress detection and phenological stage analysis.

9) **Backscatter values:**
Sentinel-1 backscatter values [17,19] measure the intensity of radar signals reflected from the Earth's surface and are crucial for monitoring land use, vegetation, soil moisture, and crop growth. As a radar satellite, Sentinel-1 provides consistent data in all weather conditions, including during cloud cover and at night. Higher backscatter values typically indicate rough surfaces such as forests or built-up areas, while lower values correspond to smoother surfaces like water bodies or bare soil. In agriculture, backscatter data are widely used to monitor crop health, estimate biomass, and assess soil conditions. This is especially important for crops such as sugarcane, cotton, and soybean in cloudy or monsoon-prone regions, where radar ensures reliable data availability.

Together, these indices play a crucial role in monitoring crop health, assessing water availability, and evaluating soil conditions, all of which are essential for accurate yield prediction and efficient crop management.

2.2 Preprocessing

The preprocessing strategy aims to prepare the data in a way that enhances model accuracy and efficiency. Essential steps include gaining a thorough understanding of the dataset, addressing missing data, eliminating irrelevant or redundant features, and normalizing or standardizing numerical data. The objective is to ensure the model is trained on clean, relevant, and well-structured data, thereby improving its overall performance. In this study, the integrated dataset contained various attributes, with less important ones filtered out during the preprocessing stage. After removing irrelevant parameters, key performance metrics including NDVI, NDWI, SAVI, MSAVI, CVI, MSI [4, 18], and backscatter values were selected for analysis. The dataset comprises 2,822 labeled samples from 2015 to 2023, covering 22 crop types across 34 districts in the Maharashtra region. Ground truth yield data were obtained from State Agriculture Department reports, which provide districtwise annual production figures. The dataset was subsequently divided, with 70% used for training and 30% for testing.

2.3 Crop Analysis and Prediction Benefits and Challenges

As noted earlier, although machine learning is being implemented across numerous industries, its application in agriculture remains a challenging and continually developing area of research. This section highlights the key benefits and challenges of utilizing ML for crop analysis and forecasting, based on insights from recent studies [22-24].

A) Benefits

Machine learning (ML) offers transformative advantages in agriculture, enhancing efficiency, crop yields, and sustainability. By analyzing large datasets from weather, soil, and crop conditions, ML allows data-driven decisions that optimize resource use, reduce input costs, and improve crop management. It also supports early disease detection and real-time monitoring of crop health, leading to timely interventions and increased profitability.

B) Challenges

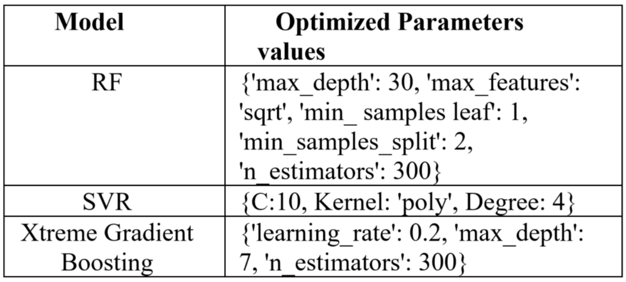
Despite its potential, ML adoption in agriculture faces key challenges such as poor data quality, model complexity, and limited interpretability of results. Many farmers lack access to the necessary infrastructure and technical knowledge, especially in remote areas. Additionally, concerns about data privacy and the need for user-friendly tools can hinder implementation. Overcoming these issues requires collaborative efforts among farmers, researchers, and tech developers.

3. METHODOLOGY

In this study, various machine learning (ML) algorithms are employed to estimate crop yield and production. Algorithms such as Extreme Gradient Boosting (XGB), Random Forest (RF), and Support Vector Machines (SVM) with polynomial and radial basis function (RBF) kernels [20], along with KSTAR, AdaBoost, Hoeffding Tree, and Decision Tree [14], are pivotal in agriculture, particularly for predicting crop yields. These algorithms analyze complex agricultural data, including soil quality, weather patterns, irrigation levels, and crop health. By processing this data, the models can identify trends, predict yields, and offer insights to enhance farming practices. Decision trees and random forests handle nonlinear data effectively, while boosting methods like AdaBoost enhance accuracy by combining weak models. SVMs with polynomial and RBF kernels [20] excel at multi-dimensional data for classification and regression, aiding farmers in data-driven decisions, optimizing resources, and improving crop yields to support food security.

The GridSearchCV API from the scikit-learn library in Python was used to optimize the various model parameter listed in Table 3 [28].

Table 3: *Parameters used in different models.*



Model	Optimized Parameters values
RF	{'max_depth': 30, 'max_features': 'sqrt', 'min_ samples leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
SVR	{C:10, Kernel: 'poly', Degree: 4}
Xtreme Gradient Boosting	{'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 300}

Fine-tuning these parameters using cross-validation ensures models achieve their best performance for crop yield forecasting. Randomly selecting the hyperparameter combinations within defined ranges, delivers faster outcomes, particularly for large datasets.

Compared to traditional methods, these ML approaches offer several advantages, including the ability to efficiently process smaller datasets and deliver more accurate predictions. The proposed framework is illustrated in Fig 2.

A total of 2,822 sample images were collected over the course of the year, covering 22 distinct crop types, including Tur, Bajra, Castor Seed, Cotton, Gram, Groundnut, Jowar, Maize, Moong, Rabi Pulses, Cereals, Oilseeds, Summer Pulses, Ragi, Rice, Safflower, Sesame, Soybean, Sugarcane, Sunflower, Urad, Wheat, and Castor Seed. Data from January 1, 2015, to December 30, 2023, was used to develop and

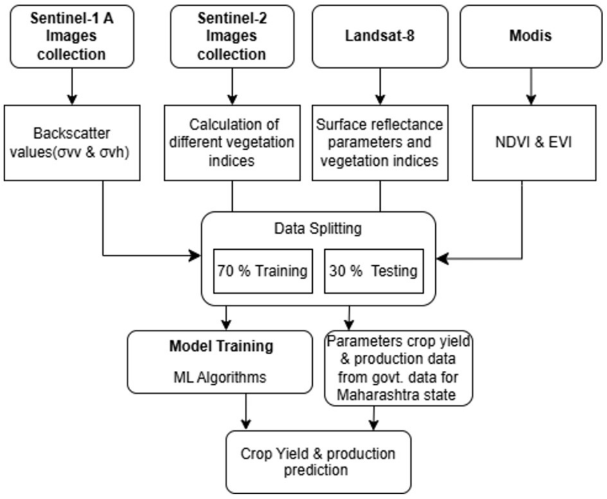

Sentinel-1 A Images collection
Sentinel-2 Images collection
Landsat-8
Modis
Backscatter values(σvv & σvh)
Calculation of different vegetation indices
Surface reflectance parameters and vegetation indices
NDVI & EVI
Data Splitting
70 % Training
30 % Testing
Model Training
ML Algorithms
Parameters crop yield & production data from govt. data for Maharashtra state
Crop Yield & production prediction

Fig.2: *Framework of proposed methodology.*

validate machine learning algorithms. The dataset was divided, with 70% allocated for training and 30% for testing. The model's predictive performance was evaluated using the test data, based on parameters such as crop year, season, crop type, crop yield, crop production, vegetation indices (VI), backscatter values, and surface reflectance values. Performance metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2), were employed to evaluate the accuracy of the machine learning models.

4. RESULTS AND DISCUSSIONS

able 4. presents performance metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) values [11] for various machine learning algorithms including KSTAR, AdaBoost, Hoeffding Tree, Decision Tree, Extreme Gradient Boosting (XGB), Random Forest (RF), and Support Vector Machines (SVM) with polynomial and radial basis function (RBF) kernels in crop yield prediction. The results indicate that the Random Forest algorithm outperforms the others, achieving the best scores for MSE, MAE, and R^2.

Table 4. showcases crop yield prediction results for the Maharashtra region, encompassing 34 districts and multiple crops during Kharif, Rabi, and full-year periods. Performance varied among algorithms based on MSE, MAE, and R^2 metrics. The Random Forest model demonstrated the highest accuracy, with a lowest MSE of 41.91, an MAE of 1.30, and a strong R^2 of 0.70, indicating its effectiveness in explaining data variance. The SVM with RBF kernel [20] also performed well, recording an MSE of 43.30, an MAE of 1.42, and an R^2 of 0.69, slightly better than the polynomial kernel. Extreme Gradient Boosting [26] showed reasonable predictive ability, while Decision Tree and Hoeffding Tree models performed less effectively. Overall, Random Forest and SVM models emerged as the most reliable for predicting crop yield in this study.

Table 4: *ML Algorithm Performance Metrics For 9 input Features NDVI, NDWI, SAVI, MSAVI, CVI, MSI, EVI (NDWI+SAVI), Surface Reflectance Values, and Backscatter values of 34 Districts and 22 crops in Maharashtra State for Crop Yield Prediction.*

ML ALGORITHMS	MSE	MAE	R-SQUARED
KSTAR	98.13	3.75	0.61
Adaboost	109.06	4.30	0.50
Hoeffding Tree	58.08	2.90	0.65
Decision Tree	79.27	1.68	0.45
Extreme GB	58.08	1.71	0.60
RF	**41.91**	**1.30**	**0.70**
SVM (Poly kernel)	45.55	1.45	0.67
SVM(RBF)	43.30	1.42	0.69

Combining vegetation indices such as NDVI with SAVI, NDVI with MSAVI, and SAVI with MSAVI offers a more comprehensive and accurate assessment of crop health and productivity. The combination of NDVI and SAVI is particularly beneficial, as NDVI measures plant vigor by analyzing the reflectance of red and near-infrared light, at the same time SAVI corrects for soil brightness, making it useful in areas with sparse vegetation or bare soil. This pairing enhances the accuracy of crop health evaluations in mixed environments with varying vegetation densities.

Integrating NDVI with soil-adjusted indices such as SAVI and MSAVI improves crop monitoring by minimizing soil background effects, particularly during early growth stages or in fields with sparse vegetation. SAVI and MSAVI are especially effective for rainfed and dryland crops like tur, bajra, jowar, moong, ragi, and other pulses, where soil exposure is high. For semi-arid crops such as castor seed and ground nut, combinations like SAVI+MSAVI or standalone SAVI help manage soil influence and sparse growth. Oilseeds including sesame, sunflower, and safflower also benefit from SAVI or MSAVI in early or dry conditions, at the same time summer pulses similarly require soil-adjusted indices to improve stress detection.

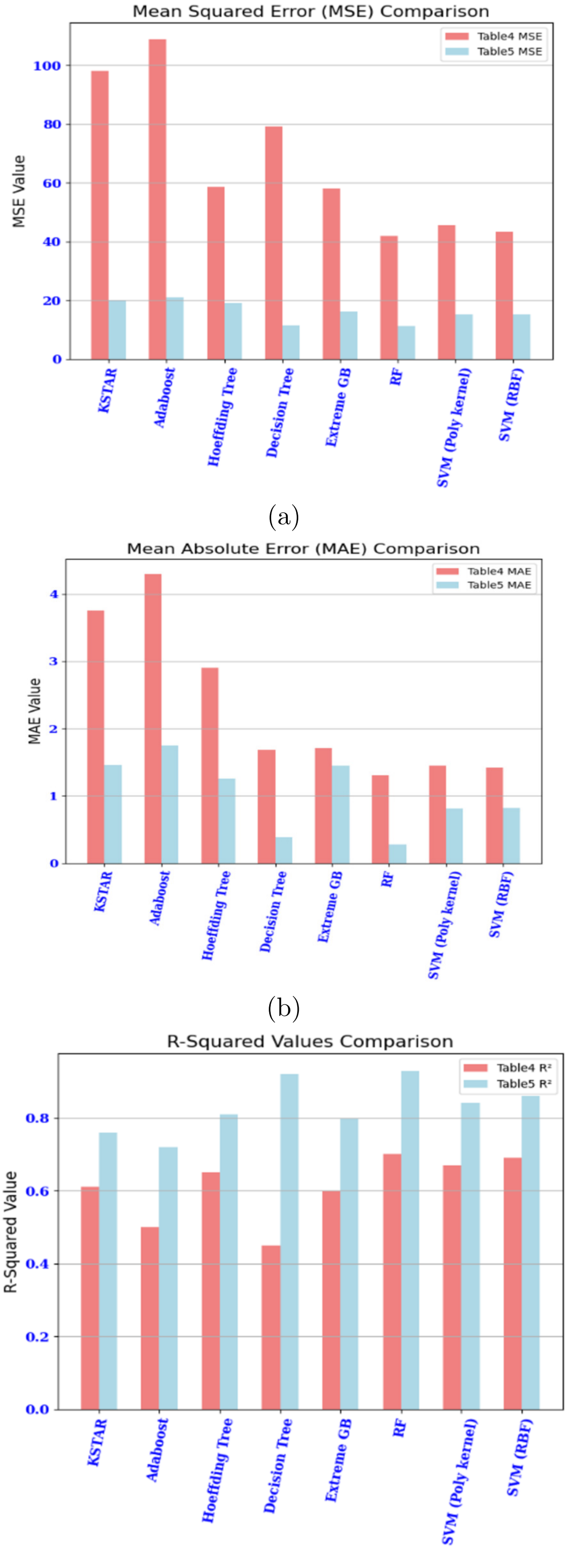
In contrast, NDVI-based combinations are better suited for irrigated or dense-canopy crops. NDVI+SAVI reliably tracks rice, wheat, and cotton, efficiently addressing canopy growth and necessary soil adjustments. Maize benefits most from NDVI+MSAVI, balancing its dense canopy with soil background correction. These targeted selections ensure indices align with crop type, growth stage, and environmental conditions, enabling more accurate assessment of vegetation health, early stress detection, and yield prediction.

To isolate seasonal effects, models were trained separately on Kharif and Rabi data subsets. Feature combinations such as (NDWI + SAVI) achieved higher R^2 for rainfed crops during Kharif, whereas (NDVI + MSAVI) performed better for irrigated Rabi crops. Feature permutation analysis resulted in a 23% reduction in MSE compared to using raw index inputs alone.

Table 5 shows that when NDVI, SAVI, and MSAVI are combined as separate input features, the performance of both Random Forest and Decision Tree algorithms [28] improves moderately. Throughout the study, results from the AdaBoost and KSTAR algorithms were relatively lower compared to those produced by the Random Forest algorithm.

Table 5: *ML algorithm performance metrics for input features SAVI, MSAVI, (NDVI+SAVI), (NDVI+MSAVI), (SAVI+ MSAVI) of 34 districts and 22 crops in Maharashtra state for crop yield prediction.*

ML ALGORITHMS	**MSE**	**MAE**	**R-SQUARED**
KSTAR	20.12	1.46	0.76
Adaboost	21.05	1.75	0.72
Hoeffding Tree	19.08	1.25	0.81
Decision Tree	**11.33**	**0.28**	**0.92**
Extreme GB	16.12	1.45	0.80
RF	**11.24**	**0.38**	**0.93**
SVM (Poly kernel)	15.16	0.81	0.84
SVM(RBF)	15.20	0.82	0.86



(a)

(b)

(c)

Fig.3: *(a) Comparative barplot of MSE values for Table 4. and Table 5 represents crop yield prediction. (b) Comparative barplot of MAE values for Table 4. and Table 5 represents crop yield prediction. (c) Comparative barplot of R^2 values for Table 4. and Table 5 represents crop yield prediction.*

Fig. 3(a) illustrates the Mean Squared Error (MSE) values for various models across two datasets (Table 4 and Table 5). Table 4 details performance metrics for nine features, including NDVI, NDWI, SAVI, MSAVI, CVI, MSI, EVI, combined features (NDWI+SAVI), surface reflectance, and backscatter values. In contrast, Table 5 includes two individual features (SAVI and MSAVI) and three combined features (NDVI+SAVI), (NDVI+MSAVI), and (SAVI+MSAVI)). The red bars represent the Mean Absolute Error (MAE) values for Table 4, while the blue bars correspond to Table 5.

Models such as KSTAR and AdaBoost exhibit the highest MSE in Table 4, whereas all models show significantly lower MSE values in Table 5, as illustrated in Fig. 3(a). SVMs with polynomial and RBF kernels demonstrate comparatively better performance across both datasets, consistently yielding lower MSE's. Overall, the use of combined features results in lower MSEs for all models, indicating enhanced predictive performance compared to the complete set of features listed in Table 4.

Challenges such as sparse vegetation were mitigated using soil-adjusted indices (SAVI, MSAVI), as reflected in the improved R^2 values (0.93) shown in Table 5. The advantage of multi-source data fusion is evident from the robust performance of RF (R^2 $>$ 0.7) across districts representing different agro-climatic zones.

The bar plot in Fig. 3(b) compares the Mean Absolute Error (MAE) values of different models for the two datasets, as detailed in Tables 4 and 5. In Table 4, models such as KSTAR and AdaBoost exhibit the highest MAE values, exceeding 3.5, while their corresponding MAEs in Table 4 drop significantly below 2. Similarly, Hoeffding Tree, Decision Tree, and Extreme Gradient Boosting exhibit higher MAE values when using combined features compared to the nine individual features, although the difference is less pronounced. In contrast, models such as Random Forest (RF), SVM with a polynomial kernel, and SVM with an RBF kernel consistently report lower MAE values across both datasets, with Table 5 highlighting their superior performance.

Overall, the use of combined features leads to lower MAE values across all models, indicating improved accuracy and reduced error compared to using the full set of features. Fig. 3(c) presents a comparison of R^2 values, illustrating the performance of various machine learning models for the two datasets detailed in Tables 4 and 5, covering 34 districts and 22 crops in Maharashtra. R^2 values reflect how effectively a model explains data variability, with values closer to 1 indicating better performance.

Random Forest (RF) achieves the highest R^2 values in both datasets, underscoring its capability to capture data patterns effectively. Decision Tree and Extreme Gradient Boosting also demonstrate strong performance, while KSTAR and SVM (Radial Basis function) show lower R^2 values, reflecting weaker explanatory power. Notably, R^2 values for the combined features are generally higher than those for all nine features, indicating improved predictive performance in the second dataset.

This highlights the importance of selecting appropriate models to achieve predictive accuracy. Table 6 presents the R^2 values for various machine learning models applied to crop production estimation. These values indicate the proportion of variance each model explains, with higher R^2 values reflecting better predictive accuracy. Among the models, Random Forest (RF) stands out with the highest R^2 value of 0.60, indicating it explains 60% of the variability in crop production, making it the most accurate model in this comparison. Support vector machine with RBF and polynomial kernels follow closely, with R^2 values of 0.59 and 0.57, respectively, demonstrating strong predictive capabilities. Extreme Gradient Boosting (GB) also performs well, achieving an R^2 of 0.51. In contrast, decision tree-based models, including the Decision Tree and Hoeffding Tree, exhibit moderate R^2 values of 0.47 and 0.46, respectively.

At the lower end, KSTAR and AdaBoost exhibit the least favorable R^2 values of 0.42 and 0.41, indicating they explain less variability in crop production compared to the other models. Overall, models such as RF and SVM (RBF) provide more reliable estimates for crop production, suggesting they are preferable for predictive analysis in agricultural forecasting.

Table 6: *R^2 values of different ML algorithms for crop production prediction of 34 districts in Maharashtra.*

Method	**R^2**	**Method**	**R^2**
KSTAR	0.42	Extreme GB	0.51
Adaboost	0.41	**RF**	**0.60**
Hoeffding Tree	0.46	SVM (Poly)	0.57
Decision Tree	0.47	SVM (RBF)	0.59

The bar graph in Fig. 4 illustrates the R^2 scores of different machine learning models Random Forest (RF), Extreme Gradient Boosting (XGB), and Support Vector Machine (SVM) across various crops and districts, providing insights into their predictive performance. Sugarcane in Sangli district exhibits high R^2 scores, with XGB achieving approximately 0.75 and SVM around 0.65, indicating strong predictive power for yield in this district. In contrast, Soybean in Pune district during the Kharif season shows moderate R^2 values of about 0.45 for XGB and 0.35 for SVM, reflecting less reliable predictions. The pseudo-R^2 (coefficient of determination) was calculated using sklearn. metrics.r2_score, which is suitable for assessing explained variance in non-linear regression models.

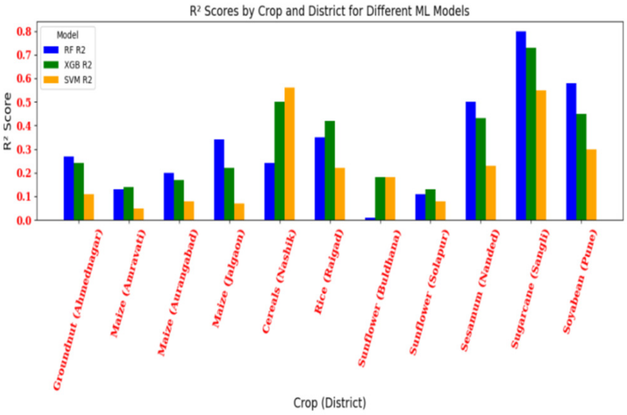


Fig.4: *R^2 values across different crop types and district for RF, XGB and SVM models.*

Groundnut in Ahmednagar district exhibits suboptimal R^2 scores across all models, with values around 0.1 for Random Forest, 0.2 for XGBoost, and 0.15 for SVM, reflecting poor model performance for this crop in this district. Maize in Aurangabad and Cereals in Nashik also show relatively low R^2 values, typically ranging between 0.1 to 0.3, suggesting that these crops have weak predictive accuracy for the models tested. In contrast, Sunflower in Solapur and Rice in Raigad show moderate to good R^2 values, ranging from 0.3 to 0.6, indicating a better fit between the model predictions and actual data. XGBoost generally provides the highest R^2 values across most crops, particularly for Sugarcane, with scores often exceeding 0.7, indicating its effectiveness for these crop district combinations. Random Forest and SVM also show good performance, but with less consistency, mainly for crops like Rice and Soyabean, where SVM has a lower R^2 value.

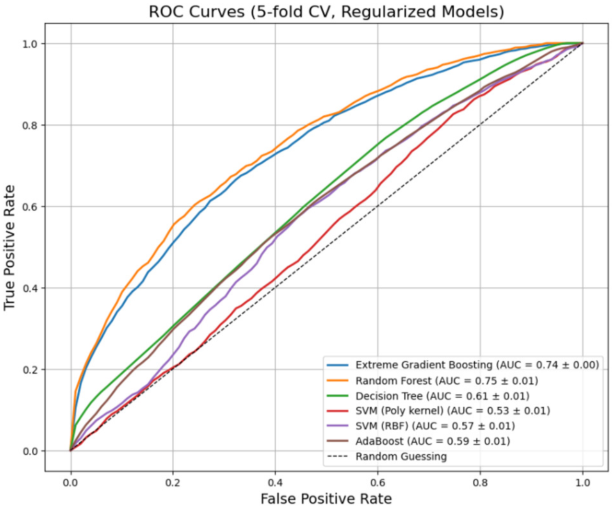


Fig.5: *ROC curve of machine-learning algorithms for crop yield prediction.*

To mitigate potential overfitting during model evaluation, 5-fold cross-validation was applied, as illustrated in Fig. 5. In this approach, the dataset is divided into five subgroups, and each model is trained and validated five times, with a different subgroup used for testing in each iteration. This approach ensures that performance metrics are not biased by a single train-test split, providing a more reliable estimate of how the models generalize to unseen data. The ROC curves [27] as demonstrated in Fig. 5. demonstrate the trade-off between true positive and false positive rates for various regularized models, with their average AUC values and standard deviations reflecting both predictive performance and stability across folds. This validation technique reduces variance in evaluation, making the comparison among models such as Random Forest (AUC = 0.75 ± 0.01) and XGBoost (AUC = 0.74 ± 0.00) more robust and trustworthy.

In Maharashtra, sugarcane, maize, and soybean exhibited distinct production trends from 2015 to 2023, as demonstrated in Fig. 6. Sugarcane production rose steadily after 2019–2020, surpassing 2.5 lakh tonnes, while maize production crossed 40,000 tonnes post-2019–2020, and soybean maintained consistent growth since 2015–2016. Across these crops, XGBoost provided smoother and more accurate predictions, whereas Random Forest captured short-term fluctuations but showed higher variability.

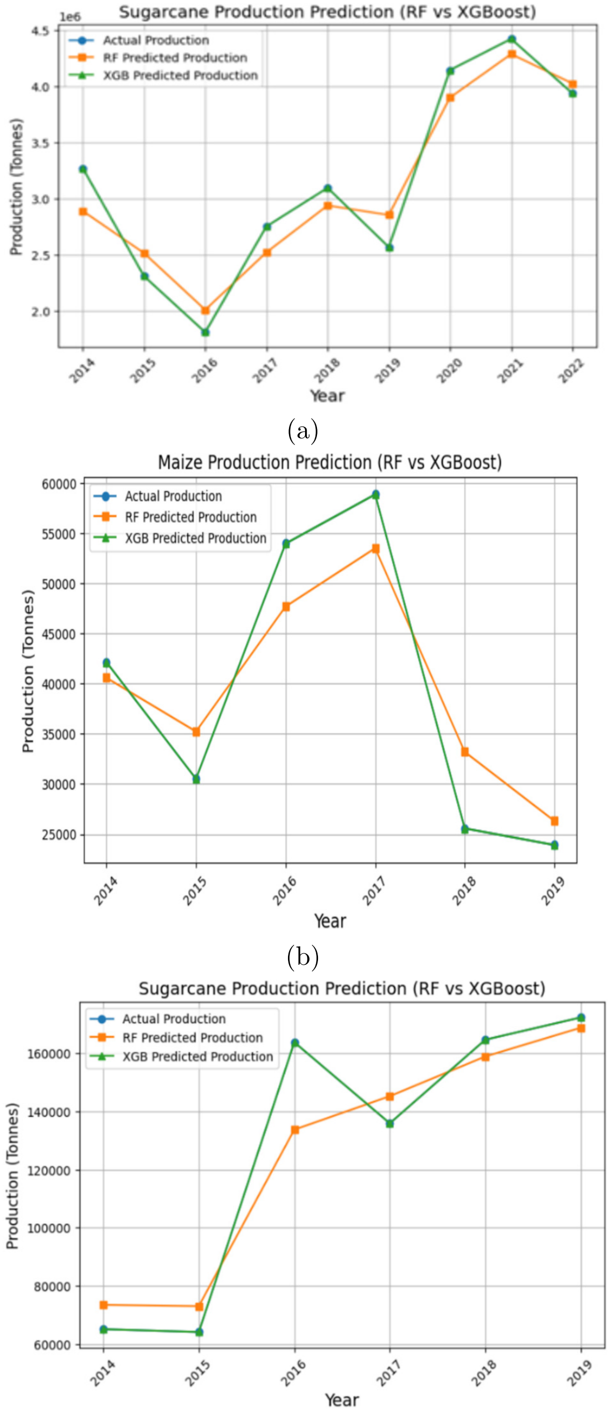
Yield patterns in Fig. 7 reflect these production trends, with sugarcane yields rising sharply after 2020–2021, maize yields improving from 2018–2019 onward, and soybean yields showing steady growth since 2015–2016. Here too, XGBoost consistently offered more reliable and generalized predictions, while Random Forest was prone to underestimation and inter-annual variability. Overall, both models performed well, but XGBoost demonstrated greater accuracy and stability, making it more suitable for long-term crop forecasting.

5. CONCLUSIONS

This study evaluated crop production and yield prediction for sugarcane, maize, and soybean in Maharashtra between 2015 and 2023 using ensemble learning models. By integrating remote sensing indices with historical data, the performance of Random Forest (RF) and Extreme Gradient Boosting (XGBoost) was analyzed. The models effectively captured production and yield trends: sugarcane production exceeded 2.5 lakh tonnes after 2019–2020 with yield rising post-2020–2021, maize production crossed 40,000 tonnes with yield improvements from 2018–2019, and soybean showed consistent growth since 2015–2016. These results highlight the ability of ML models to capture long-term crop dynamics and variability.

In terms of accuracy, both RF and XGBoost achieved strong performance, with RF capturing local fluctuations ($R^2 > 0.90$ across crops) but sometimes underestimating yield. XGBoost consistently delivered smoother and more robust estimates, with maize yield prediction achieving $R^2 = 0.95$ compared to 0.92 for RF, and soybean production reaching 0.94 compared to 0.91. Seasonal crop-wise analysis further confirmed that XGBoost performed best for both kharif crops like soybean and maize, and rabi crops such as wheat and pulses, demonstrating its adaptability across diverse seasonal conditions. RF, though effective in short-term variability, was less stable across seasons.

The feature combination analysis presented in Table 5 highlighted that optimal vegetation index combinations significantly enhance prediction accuracy. NDVI+SAVI and NDVI+MSAVI performed best for dense-canopy crops such as maize, rice, and wheat, whereas soil-adjusted indices like SAVI and MSAVI were more effective for rainfed crops including tur, ba-

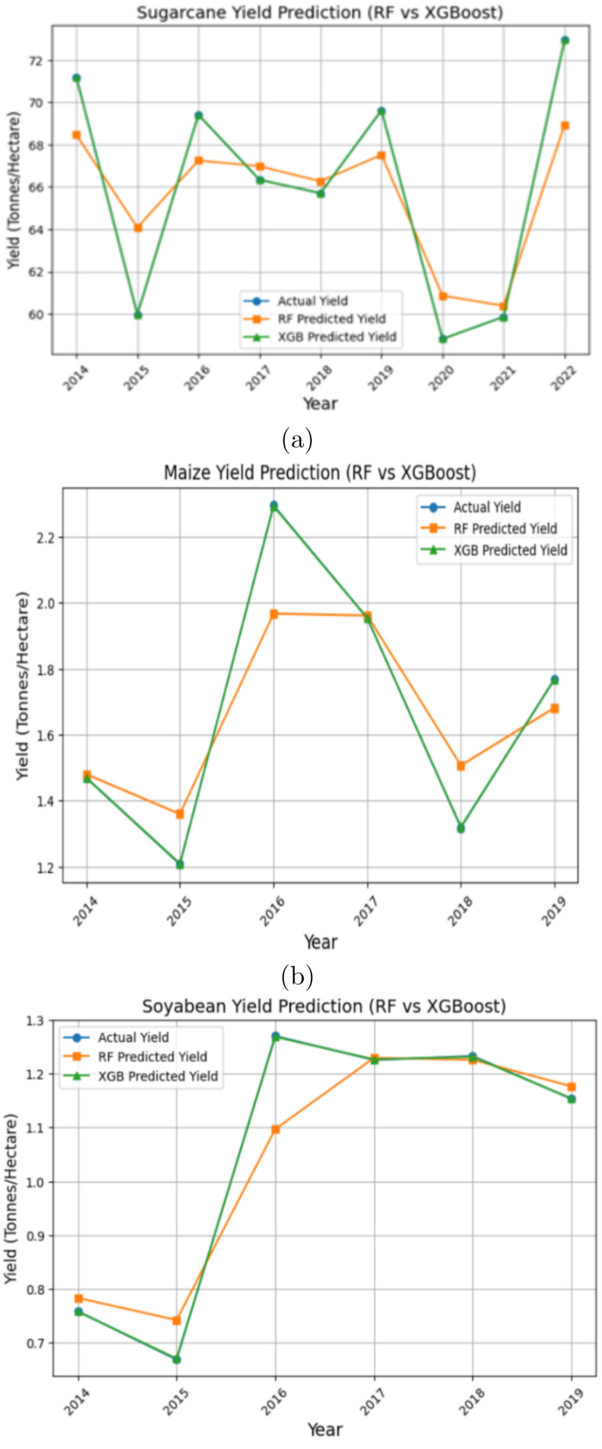


(a)

(b)

(c)

Fig.6: *Year-wise mean crop production of (a)Sugarcane, (b) Maize & (c) Soyabean from Jan 2015 to Dec 2023.*



(a)

(b)

(c)

Fig.7: *Year-wise mean crop Yield of (a) Sugarcane (b) Maize & (c) Soyabean from Jan 2015 to Dec 2023.*

jra, jowar, and pulses primarily during early growth stages. Seasonal, crop-wise results in Table 5 further reinforced these findings, demonstrating that index selection tailored to crop type and growth stage plays a critical role in improving model accuracy. Overall, the study concludes that while both models are valuable, XGBoost combined with tailored feature selection (as evidenced in Table 5) provides the most reliable framework for precision agriculture, supporting sustainable resource planning and climate resilient crop forecasting.

Future research should focus on improving data quality through advanced preprocessing techniques, multi-source satellite integration, and region-specific modeling for underperforming crop district combinations. Model performance can be enhanced using hyperparameter tuning, hybrid ensembles, and scalable algorithms such as LightGBM and CatBoost. To improve usability, explainable AI techniques (e.g., SHAP, LIME) will be applied, with results delivered through farmer-friendly dashboards and mobile apps. Additionally, lightweight, cloud-enabled, and offline-compatible tools should be developed to ensure accessibility in rural areas, while integrating pest dynamics, irrigation factors, and socio-economic variables under secure data management practices to enhance prediction reliability.

AUTHOR CONTRIBUTIONS

Conceptualization, Shilpa Vatkar, Sujata Kulkarni; methodology, Shilpa Vatkar, Sujata Kulkarni.; software, Shilpa Vatkar, Sujata Kulkarni.; validation, Shilpa Vatkar, Sujata Kulkarni; formal analysis, Shilpa Vatkar.; investigation, Sujata Kulkarni.; data curation, Shilpa Vatkar, Sujata Kulkarni.; writing—original draft preparation, Shilpa Vatkar.; writing—review and editing, Shilpa Vatkar, Sujata Kulkarni.; visualization, Shilpa Vatkar, Sujata Kulkarni.; supervision, Sujata Kulkarni.; funding acquisition, Sujata Kulkarni. All authors have read and agreed to the published version of the manuscript.

References

[1] J. Liu *et al.*, "Crop Yield Estimation Using Time-Series MODIS Data and the Effects of Cropland Masks in Ontario, Canada," *Remote Sensing*, vol. 11, no. 20, p. 2419, Jan. 2019.

[2] N. Kim, K.-J. Ha, N.-W. Park, J. Cho, S. Hong, and Y.-W. Lee, "A Comparison Between Major Artificial Intelligence Models for Crop Yield Prediction: Case Study of the Midwestern United States, 2006–2015," *ISPRS International Journal of Geo-Information*, vol. 8, no. 5, p. 240, May 2019.

[3] S. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," *Frontiers in Plant Science*, vol. 10, 2019.

[4] A. K. Maurya, M. Nadeem, D. Singh, K. P. Singh and N. S. Rajput, "Critical Analysis of Machine Learning Approaches for Vegetation Fractional Cover Estimation Using Drone and Sentinel-2 Data," *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Brussels, Belgium, pp. 343-346, 2021.

[5] F. Gao, J. Masek, M. Schwaller and F. Hall, "On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2207-2218, Aug. 2006.

[6] E. Vermote, "MOD09Q1 MODIS/Terra Surface Reflectance 8-Day L3 Global 250m SIN Grid V006," *NASA EOSDIS Land Processes DAAC*, 2015.

[7] T. Dong *et al.*, "Estimating winter wheat biomass by assimilating leaf area index derived from fusion of Landsat-8 and modis data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 49, pp. 63–74, 2016.

[8] M. Mao, H. Zhao, G. Tang and J. Ren, "In Season Crop Type Detection by Combing Sentinel-1A and Sentinel-2 Imagery Based on the CNN Model," *Agronomy*, vol. 13, no. 7, p. 1723, 2023.

[9] F. Kogan *et al.*, "Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models," *International Journal of Applied Earth Observation and Geoinformation*, vol. 23, pp. 192–203, Aug. 2013.

[10] R. Balaghi, B. Tychon, H. Eerens and M. Jlibene, "Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco," *International Journal of Applied Earth Observation and Geoinformation*, vol. 10, no. 4, pp. 438–452, 2008.

[11] M. Huihui, M. Jihua, F. Ji, Q. Zhang and H. Fang, "Comparison of Machine Learning Regression Algorithms for Cotton Leaf Area Index Retrieval Using Sentinel-2 Spectral Bands," *Applied Sciences*, vol. 9, no. 7, p. 1459, 2019.

[12] S. Kelkar and A. Kulkarni, "Impact of climate variability and change on crop production in Maharashtra, India," *Current Science*, vol. 118, no. 8, 25 April 2020.

[13] E. Ersin *et al.*, "Crop Prediction Model Using Machine Learning Algorithms," *Applied Sciences*, vol. 13, no. 16, p. 9288, 2023.

[14] A. Abdollahi, B. Pradhan and A. Alamri, "Regional-Scale Analysis of Vegetation Dynamics Using Satellite Data and Machine Learning Algorithms: A Multi-Factorial Approach," *International Journal on Smart Sensing and Intelligent Systems*, vol. 16, no. 1, 2023.

[15] C. Achahboun, M. Chikhaoui, M. Naimi and M. Bellafkih, "Crops Classification Using Machine

Learning And Google Earth Engine," *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)*, Casablanca, Morocco, pp. 1-8, 2023.

[16] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," Remote Sensing of Environment, vol. 202, pp. 18-27, 2017.

[17] A. Verma, A. Kumar and K. Lal, "Kharif crop characterization using combination of SAR and MSI Optical Sentinel Satellite datasets," *Journal of Earth System Science*, vol. 128, no. 230, 2019.

[18] F. Novelli, H. Spiegel, T. Sandén and F. Vuolo, "Assimilation of Sentinel-2 Leaf Area Index Data into a Physically-Based Crop Growth Model for Yield Estimation," *Agronomy*, vol. 9, no. 5, p. 255, May 2019.

[19] P. Tummala, M. Sobhana and S. Kakumani, "Predicting crop yield with NDVI and Backscatter values using Deep Neural Networks," *2022 International Mobile and Embedded Technology Conference (MECON)*, Noida, India, pp. 390-394, 2022.

[20] S. V. S. Prasad, T. S. Savithri and I. V. M. Krishna, "Performance Evaluation Of SVM Kernels On Multispectral Liss III Data For Object Classification," *International Journal On Smart Sensing And Intelligent Systems*, vol. 10, no. 4, 2017.

[21] A. Bannari, D. Morin, F. Bonn and A. R. Huete, "A review of vegetation indices," *Remote Sensing Reviews*, vol 13, no.1-2, pp. 95-120, 1995.

[22] S. P. N and H. P. M. Kumar, "Soil Quality Identifying and Monitoring Approach for Sugarcane Using Machine Learning Techniques," *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, Mandya, India, pp. 1-5, 2022.

[23] F. Xiao, H. Wang, Y. Xu and R. Zhang, "Fruit Detection and Recognition Based on Deep Learning for Automatic Harvesting: AnOverview and Review," *Agronomy*, vol. 13, no. 6, p. 1625, 2023.

[24] A. L'Heureux, K. Grolinger, H. F. Elyamany and M. A. M. Capretz, "Machine Learning With Big Data: Challenges and Approaches," in *IEEE Access*, vol. 5, pp. 7776-7797, 2017.

[25] S. D. Río, V. López, J. M. Benítez and F. Herrera, "On the use of Map Reduce for imbalanced Big Data using Random Forest," *Information Sciences*, vol. 285, pp. 112–137, 2014.

[26] X. Li and R. Bai, "Freight Vehicle Travel Time Prediction Using Gradient Boosting Regression Tree," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, CA, USA, pp. 1010-1015, 2016.

[27] L. Lavazza, S. Morasca and G. Rotoloni and , "On the Reliability of the Area Under the ROC Curve in Empirical Software Engineering," in *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, pp. 93-100, 2023.

[28] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *The Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

Shilpa Naresh Vatkar is faculty of Department of Electronics and Telecommunication Engg with K.J. Somaiya School of Engg, Somaiya Vidyavihar University. Her research interest includes image processing, remote sensing, VLSI signal processing, VLSI design, artificial intelligence, machine learning and neural networks.

Sujata S. Kulkarni is working as Associate Professor, in the Department of Computer Science Engineering, Bhartiya Vidya Bhavan's Sardar Patel Institute of Technology, Mumbai University, India. Her research interest includes image processing Pattern Recognition, Biometric identification, Security, Cryptography, artificial intelligence, machine learning, Communication and Networking, Wireless Communication Networks, and Embedded System.