EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

# An approach for crop yield prediction using hybrid XGBoost, SVM and C4.5 classifier algorithms

Renzón Daniel Cosme Pecho*[1], K. Sri Vijaya[2], Neelam Sharma[3], Hemant Pal[4] and Bibin K. Jose[5]

[1]Department of Chemistry and Biology, Salutem Diagnostic Imaging Center, Lima, 15498, Peru
[2]Department of IT, PVP Siddhartha Institute of Technology, Vijayawada, 520007, India
[3]School of Physical Education, Lovely Professional University Punjab, Phagwara, Punjab, 144001, India
[4]Department of Computer Science, Medi-Caps University, Indore, Madhya Pradesh, 453331, India
[5]Department of mathematics, S.D. College Alappuzha, Kerala, 688003, India

Received 24 June 2023
Revised 10 February 2024
Accepted 16 February 2024

**Abstract**

Science and technological knowledge advancements have resulted in a vast number of data for the agricultural industry. Crop yield prediction (CYP) is a problematic issue in the agricultural field. The proposed work constructs a hybridization of the XGBoost-SVM-C4.5 framework to forecast crop yield. Also, the proposed methodology is employed to predict different crop yields based on temperature, rainfall, and soil parameters. The experimental setup was based on data gathered from the Indian Meteorological Department for various crops. Five metrics were analyzed to determine the performance of each approach under research: Determination Coefficient, Root Mean Squared Error, Correlation coefficient, Mean Absolute Error and Mean Squared Error. Experiments have been conducted to determine the most effective approach for predicting various crop yields. Additionally, the results of this research give an appropriate method for forecasting the future of food production, allowing for a more precise plan for agricultural production.

**Keywords:** Crop yield prediction, C4.5, Feature selection, Machine learning, SVM, XGBoost

## 1. Introduction

Accurate and timely forecasting of agricultural statistics, such as crop output, contributes greatly to a nation's development. Predictions made early and accurately boost future yield and profit. Prediction of Crop yield is a complex subject that is essential for resource efficiency along with sustainable growth [1, 2]. Crop yield forecasts are helpful to many different parties involved in the agri-food chain, including farmers, agronomists, policymakers, and commodities traders. Numerous factors, such as soil quality, climate, and fertilizer application, have an impact on the yield of crops [3, 4].

For the Crop Yield Prediction (CYP), there are two major feature sets: one type of data contains details on how the land is used, which fertilizer is applied, and how the cultivators water it. The other comprises elements of the natural world, like temperature, solar radiation, and rainfall [5]. Nevertheless, gathering information relevant to the CYP is a difficult and drawn-out process [5-7].

To produce precise forecasts based on available data, researchers are creating data-driven [6, 7]. To maximize reliability in data-driven systems, machine learning (ML) techniques are essential [8]. Despite significant advancements in machine learning and its applications across various domains, machine learning techniques have inherent limitations when applied in a purely data-driven manner [8-10]. The degree of reliance on the input and the variables of interest in the obtained datasets, the system's capability, and the quality of the information all affect how accurate the forecasts and reservations produced by machine learning algorithms are [9, 10]. The predictive capacity of systems can be significantly reduced by high noise, inaccurate data, biases, outliers, and insufficient datasets [11, 12].

The artificial neural network (ANN) is the machine learning method most commonly utilized for CYP. The human brain is replicated in a neural network. One advantage of artificial neural networks (ANNs) is their ability to learn from cases directly, without statistical methods to estimate parameters [13, 14]. In recent research, ANNs have been used extensively to calculate agricultural yields for various crops, including paddy, wheat, basil, and kiwifruit. Elavarasan and Vincent [11] investigate the way tillage practices affect crop productivity and soil properties using MLR and ANN. The data are validated using the correlation coefficient of determination ($R2$) and the root mean square error (RMSE). To estimate yields from agriculture based on farm conditions, social variables, and meteorological inputs, a modular neural network with two hidden layers was established [15, 16]. Shidnal et al. [12] and Sivanandhini and Prakash [13] evaluated the minimal input parameters required to calculate the lint production using the performance assessment statistics $R^2$ and MAE in ANN. Numerous studies analyzed agricultural data using ANN models. Among these works, Suganya et al. [14] used climatic and soil data to forecast rain-fed wheat production, Gulati and Jha [15] used a multilayer feed-forward neural network model to predict turmeric essential oil output and Obsie et al. [16] developed an internet platform to predict apricot yield. Ma et al.

*Corresponding author.
Email address: rdaniel3200@gmail.com
doi: 10.14456/easr.2024.29

[17] and Paudel et al. [18] used the Bayesian ensemble model (BM) to analyze historical crop yield data and forecast crop yield, while Hu et al. [19] and Morales and Villalobos [20] used synthetic datasets from biophysical models of crops to investigate the impact of data amount, data partitioning strategies, and predictive algorithm choice on predictive performance.

Climate and soil conditions are regarded to be the primary environmental factors influencing the development and variability of agricultural yields. The phonological cycle of plants is also influenced by climate variables. Numerous studies on the yield prediction of food crops such as rice and grain maize have been conducted in the southern part of the Korean Peninsula. The same experiment based on meteorological data were conducted on plants such as apple, Chinese cabbage, whole crop maize, whole crop rye and Italian ryegrass, [19, 20].

The research preprocessed meteorological data to address current issues and put it into the suggested XGBoost-SVM-C4.5 architecture. The proposed study is also utilized to build a link between these three variables and determine their effect on crop production. Thus, it is clear from the survey that the new machine-learning approach provided here will be utilized to estimate crop yields efficiently.

The key contributions of this paper are as shown:

- Initially, the climate-based data and a large amount of historical agricultural production are acquired and preprocessed.
- We designed the Simulated Annealing (SA) for Feature Selection to obtain an optimal feature set for accurate classification.
- To explore the performance of hybrid XGBoost-SVM-C4.5 models and the accuracy of crop yield prediction.
- To optimize the parameters of the C4.5 decision tree to obtain accurate crop prediction results.
- We designed a hybrid machine learning-based CYP model and evaluated their performance on data collected from the Indian Meteorological Department and various Maharashtra government websites.
- To develop an effective model for forecasting food production based on an analysis of the prediction performance.

The balance of this study is partitioned into the subsequent sections: Sec 2 discusses related work, and Sec 3 has the problem statement. Sec 4 summarizes the proposed model for agricultural production forecasting. Section 5 discusses the experimental observations and settings and the proposed model's effectiveness in contrast to existing approaches. Sec 6 concludes with a discussion of the conclusion and future efforts.

## 2. Materials and methods

Here, the hybrid techniques for our proposed agricultural information forecasting research (crop yield and climatic temperature) are discussed. To begin, preprocessing of input data is carried out to identify missing values, remove duplicates, normalize the dataset, and convert target variables to factor attributes. Utilizing an optimization approach, essential properties are retrieved from preprocessed data. Prior to collecting data, the optimal features are subjected to categorization algorithms. Then Simulated Annealing (SA) is proposed to reduce the feature subset and tune the parameters of the C4.5 decision tree to improve the prediction result. Finally, a crop yield forecast is developed and the outcomes are analyzed using various performance criteria. The research highlighted the most effective methodology for selecting features in combination with effective classification algorithms.

### *2.1 Selection of area*

The current study focuses on one of Nashik's most significant agriculture-based districts. This research will be conducted in the Maharashtra district of Nashik. Nashik is located at 74° 56' E longitude, and 73° 16' E, 20° 52' latitude and 73° 35' N, with a total area of 15,582 kilometers square and a population of 6,109,05, making it Maharashtra's third-biggest district. Dhule district borders it to the north, Jalgaon to the east, and Aurangabad to the southeast [21, 22]. Nashik is famous for wine production. The Godavari River runs through Nashik. Malegaon is the biggest district in Nashik, while Peth is the smallest. Other districts in Nashik include Trimbak, Deola, Niphand, Dindori, Buglan, Yeola, Surgana, Sinner, Kalwan, Nandgaon, Chandwad and Igatpuri [23, 24]. Figure 1 is a map of the Maharashtra district of Nashik.

**Figure 1** Map of Nashik district, Maharashtra, India

However, the temperature peaks between March and May. This is also the time of year when the state experiences heat waves. The monsoon season in Nasik begins in the second week of June and lasts until August. October through February are excellent weather months. Nashik's weather prediction is based on the current temperature, rainfall, wind speed, humidity, and air quality.

*2.2 Dataset*

The data for this research was gathered from the Nasik Meteorological Department, the Maharashtra Statistical Department, and the website (https://data.gov.in). Government entities provide agricultural datasets incorporating data from past years to these websites. The data set of Maharashtra's districts has a huge record of various crops. Variables in the dataset include temperature, rainfall, agricultural yield, soil, etc. Historical records, remote sensing datasets, GPS-based datasets, and social datasets are generally employed in agricultural yield prediction.

Data on food and agriculture is provided by FAOSTAT (Food and Agriculture Organization of the UN). This resource has data from 200 nations. The final dataset comprises the following input fields: Item collected, Country, Area, Production, Yield, Rainfall, soil pH, and temperature (in degrees Celsius) [25].

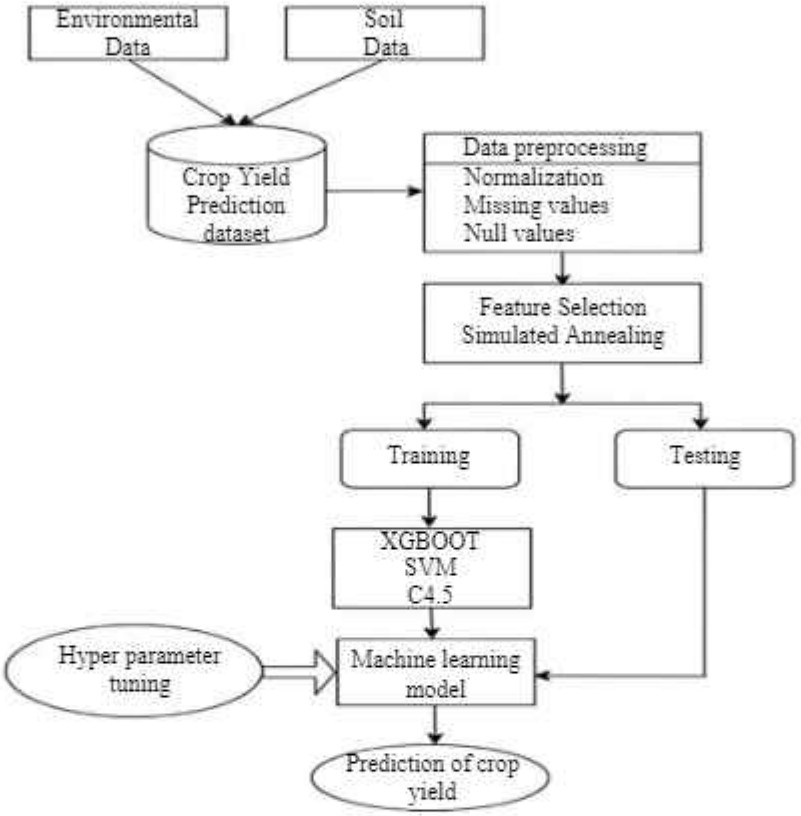


**Figure 2** Proposed Methodology

The FAO data repository has yield statistics for six major crops in Maharashtra, India. The gathered data includes nation, item, year, and yield value from 2000 to 2020. Environmental and soil conditions impact crop output. Rainfall and soil quality also influence the growth of the crop. World Data Bank provides annual rainfall statistics for India. Each item's soil pH is obtained from the FAO data collection. The global data bank has compiled the average temperature for India. To get an appropriate forecast, the acquired data is cleaned and rescaled between 0 and 1. The proposed architecture is shown in Figure 2.

*2.3 Data preprocessing*

Initially, in preprocessing, missing values are handled, and detection of outliers and data normalization are performed.

*2.3.1 Handling missing values*

Ignoring records is the simplest method to handle missing values, but this approach isn't practical for all data sets. When preparing the information, the data set is examined to see if any characteristic values are missing. The missing data for numerical characteristics was estimated using the statistical technique of mean imputation. Employing the mode method, the absent nominal characteristics were replaced.

*2.3.2 Outlier detection*

An outlier is a single observation point that stands out from the rest of the data. Measuring variation can reveal a measurement error or be the cause of an outlier. An outlier can potentially distort and fool the DL algorithm's learning process. In the end, this results in longer training times, lower model accuracy, and worse performance. This study uses an Interquartile Range (IQR)-based method to eliminate outliers from the data before sending it to the learning method.

*2.3.3 Data normalization*

Data normalization usually has values of different magnitudes. A commonly used technique for converting considerable data value ranges into smaller range values is normalization. As in this study, the min-max method is widely used for data normalization because of its higher accuracy and short learning curve. The distribution of the data set is unaffected by min-max normalization. The expressions max(x) and min(x) represent the lowest and highest values of a measure x.

$$\tilde{x}_i = \frac{x_i - min(x)}{max(x) - min(x)} \quad (1)$$

*2.4 Feature selection*

Datasets include duplicate information, which complicates categorization. Selecting features is a significant issue, mainly when datasets have many features. Feature selection is primarily used to improve the training of the ML algorithm, (ii) reduce the model's complexity, and (iii) facilitate interpretation. Additionally, it optimizes the model's accuracy by selecting the optimal subset and avoids over-fitting. This study focuses on developing an optimization technique for extracting the most valuable features from raw data. Simulated Annealing (SA) extracts the best features from the soil and environment.

- The feature selection approach initially identifies the critical characteristics that influence crop productivity (target).
- By using feature selection, noisy and irrelevant data will be eliminated.
- Due to this combination, the ML model's prediction accuracy increases.

Section 4.4.1 discusses the Simulated Annealing (SA) method for feature selection.

*2.4.1 Simulated Annealing (SA)*

SA was proposed by Kirkpatrick et al. based on the solid annealing concept. It is a stochastic optimization approach that is based on the hill-climbing technique. Since a poor outcome is approved with a fixed probability during every iteration, SA avoids the issue of local optima. The algorithm starts with a random initial solution. In each cycle, a new solution *(Newsol)* is found and analyzed about the existing solution *(Cursol)*. When a new solution's fitness is better than the existing best solution's *(Bestsol)*, the new solution replaces the existing solution. Otherwise, the Boltzmann probability indicates that the new solution is acceptable. Adoption of the new solution is thus shown as

$$P = \begin{cases} 1 & \delta(Newsol) \geq \delta(Bestsol) \\ e^{-\frac{\theta}{T}} & \delta(Newsol) < \delta(Bestsol) \end{cases} \quad (2)$$

Where θ is the difference between $\delta$ *(Bestsol) and* $\delta$ *(Newsol),* the fitness of a solution is denoted as δ. The temperature T drops with each cycle. The starting temperature *(T0)* is set at 0.1 in this paper. The actual temp is computed using the formula $T = 0.99\ T$. The probability of choosing a poor solution decreases as the number of iterations increases. Additionally, the number of cycles permitted in the SA algorithm is set to 30 in this paper.

This approach extracts the best characteristics from the soil and environment.

After cleaning and selecting significant characteristics from the dataset, the final dataset including all of the features that will be utilized in the forecasting process is shown in the table below (Table 1). The data set's final characteristics are as follows: state, district, year, area, production, yield, crops, rainfall, temperature, humidity, potassium, phosphorus, and nitrogen, as well as soil pH.

**Table 1** Sample dataset

| State | District | Year | Crops | Area (ha.) | Productivity (Qtl/ha) | Rainfall | Temperature °C | | N | P | K | pH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Maximum | Minimum | | | | |
| MAHARASHTRA | NASHIK | 2019 | Oilseed | 75700 | 754000 | 84 | 25.8 | 15.1 | 90 | 42 | 43 | 6.50 |
| MAHARASHTRA | NASHIK | 2019 | Wheat | 55600 | 813000 | 23 | 27 | 17.4 | 85 | 58 | 41 | 7.03 |
| MAHARASHTRA | NASHIK | 2019 | Pulses | 90000 | 609000 | - | 28.6 | 11.6 | 60 | 55 | 44 | 7.84 |
| MAHARASHTRA | NASHIK | 2019 | Sugarcane | 12100 | 6729000 | - | 28.8 | 9.1 | 74 | 35 | 40 | 6.98 |
| MAHARASHTRA | NASHIK | 2019 | Cotton | 14100 | 200000 | - | 30.1 | 24.7 | 78 | 42 | 42 | 7.62 |
| MAHARASHTRA | NASHIK | 2019 | Rice | 50800 | 702000 | - | 33.6 | 13.1 | 69 | 37 | 42 | 7.07 |
| MAHARASHTRA | NASHIK | 2020 | Oilseed | 74600 | 735000 | 34.6 | 38.5 | 18.1 | 69 | 55 | 38 | 5.70 |
| MAHARASHTRA | NASHIK | 2020 | Wheat | 45600 | 714000 | 31.1 | 36.2 | 16 | 94 | 53 | 40 | 5.71 |
| MAHARASHTRA | NASHIK | 2020 | Pulses | 80000 | 704000 | 366.5 | 31.9 | 22.6 | 89 | 54 | 38 | 6.68 |
| MAHARASHTRA | NASHIK | 2020 | Sugarcane | 11100 | 5630000 | 199.5 | 28.9 | 22.1 | 68 | 58 | 38 | 6.33 |
| MAHARASHTRA | NASHIK | 2020 | Cotton | 12100 | 100000 | 171.1 | 27.7 | 21.5 | 91 | 53 | 40 | 5.38 |
| MAHARASHTRA | NASHIK | 2020 | Rice | 40700 | 601000 | 155 | 29.7 | 20.9 | 90 | 46 | 42 | 7.50 |

*2.4.2 Classifier*

After extracting the most relevant features, the classification technique uses hybrid approaches on the reduced dataset, such as XGBoost, SVM, and C4.5. Then, it is further divided into training and testing prior to using the classifier method. The classifier algorithm is trained using the training dataset, which is carried out in the testing phase.

The resulting value is used to forecast agricultural yields for a particular area. The subdivisions that follow provide details on the classifiers employed in this research.

*2.4.3 Extreme gradient boosting (XGBoost)*

XGBoost is used for regression and classification and is considered one of the best performing algorithms for supervised learning known as gradient boosting machines (gbm). The operation of this approach is as follows: Consider the dataset (DS) with features $m$ and $n$ examples, then $DS = \{(x_i, y_i): i = 1 \ldots n, x_i \in \Re^m, Y_i \in \Re\}$. Let $\hat{y}_i$ be the ensemble tree model output predicted by the following equations:

$$\overset{0}{A}_{\cdot i} = \phi(X_i) = \sum_{k=1}^{K} f_k(X_i), f_k \in T \tag{3}$$

Where $K$ is the tree count, $f_k$ is the ($k$-$th$ tree). We must minimize the loss and regularization objectives to address the given equation.

$$\ell(\phi) = \sum_i l(y_i, \overset{0}{A}_{\cdot i}) + \sum_k \Omega(f_k) \tag{4}$$

Where $l$ is the loss function, the difference between the expected $\hat{y}_i$ and $y_i$ actual outputs. Ω is a model's complexity measure that helps prevent over-fitting.

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||W||^2 \tag{5}$$

$W$ is the weight of each leaf, and $T$ is the number of leaves in the tree. Boosting is used in decision trees to make the model as accurate as possible. This is done by including a new function $f$as as the model is being trained. With each cycle, we add a new function (tree):

$$\ell^{(t)} = \sum_{i=1}^{n} l\left((y_i, \overset{0}{A}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_k) \tag{6}$$

$$\ell_{split} = \frac{1}{2}\left[\frac{\left(\Sigma_{i\epsilon I_L} g_i\right)^2}{\Sigma_{i\epsilon I_L} h_i + \lambda} + \frac{\left(\Sigma_{i\epsilon I_R} g_i\right)^2}{\Sigma_{i\epsilon I_R} h_i + \lambda} - \frac{\left(\Sigma_{i\epsilon I_R} g_i\right)^2}{\Sigma_{i\epsilon I_R} h_i + \lambda}\right] - \gamma \tag{7}$$

$$g_i = \partial_{\overset{0}{A}_{\cdot(t-1)}} l\left(y_i, \overset{0}{A}_{\cdot}^{(t-1)}\right) \tag{8}$$

$$h_i = \partial^2_{\overset{0}{A}_{\cdot\ (t-1)}} l\left(y_i, \overset{0}{A}_{\cdot}^{(t-1)}\right) \tag{9}$$

*2.4.4 Support vector machine*

A stable and effective multivariate non-linear regression method, support vector machines. Simple non-linear relationships between variables may be attributed to linear relationships in higher-dimensional spaces. Linear optimization is performed on the variables of interest in a high-dimensional space, and the resultant regression analysis is returned to the observed variables' low-dimensional phase space. SVMs are currently widely employed as very efficient non-linear regression models.

First, the inputs are fixed (nonlinear) mapped into high dimensional feature space, and then a linear model is built in this feature space. ($x$, $y$) are two random variables, ($y$) being the dependent variable and ($x$) a collection of $R^d$ training data points. SVM seeks to predict $y_i$ given $x$. Support vector regression (SVR) is given as

$$y = f(x) + \epsilon \tag{10}$$

The function $f(x)$ is estimated as

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b, \tag{11}$$

Where the error term is denoted as $\epsilon$, a constant scalar is represented as $b$ and the kernel function is indicated by $K(x_i, x_j)$. The above equation's parameters are determined to solve the optimization equation.

$$\min_{\alpha,b} \sum_{i=1}^{n} \alpha_i + \frac{1}{2}\sum_{i,j=1}^{n} y_i \alpha_i y_j \alpha_j K(x_i, x) \tag{12}$$

$$subjectto \begin{cases} 0 \le \alpha_i \le C, i = 1,2,\ldots,n \\ \sum_i y_i \alpha_i = 0 \end{cases} \tag{13}$$

The support vector parameter C determines the opportunity cost between error minimization and margin maximization.

But, as with any advanced regression technique, overfitting is possible. The study utilized 20 years of yield data. 33 predictor variables were found, about equal to the amount of data. Using a flexible non-linear regression approach like SVM, any random combination of independent variables may achieve outstanding performance. It's like fitting a polynomial of degree n to a data set of n + 1 observations. So the predictor variables have to be reduced.

*2.4.5 C4.5 Approach*

Quinlan introduced Commercial Version 4.5 (C4.5), a commonly used machine learning approach for building decision trees. The method selects entropy-based measures that yield high classification accuracy in a short time. It is an expansion of the ID3 algorithm. The method's key benefits include pruning and handling numeric characteristics, missing values, and noisy data. Preparing the decision tree and creating the rules are two steps in the C4.5 algorithm. After that, entropy is measured and information gain is computed.

The following is the entropy formula: the output of the class proportion is denoted as p and entropy is denoted as S.

$$Entropy(S) = \sum_{i-1}^{n} -p_i \times \log_2 p_i \tag{14}$$

The root attribute is also the attribute with the highest gain. Gain (A, S) is expressed in Equation (15) where the case set is denoted as S; the case attribute is represented as A; |Si| is several cases to i and |S| are several cases in the set.

$$Gain(S, A) = Entropy(S) \sum_{i-1}^{n} \frac{|S_i|}{|S|} \times Entropy(S) \tag{15}$$

The C4.5 method constructs a DT using a recursive-splitting and top-down approach. C4.5 has leaf nodes, internal nodes and root nodes. The root node is the container for all input data. Typically, an internal node contains multiple branches and includes a decision function. Meanwhile, the leaf node represents the result of the specified input parameters.

*2.5 Crop prediction procedure*

The procedure for CYP is described as follows: Environmental conditions and soil parameters are provided as input and output to predict crop yield.

**Step 1:** Data collected from different websites are given as input, and then loaded the data set.

**Step 2:** The features in the data set are modified into a specific range, getting the data set into a uniform condition therefore eliminating irregularities. Any missing values are deleted, and the data is standardized via normalization. Once redundancy is eliminated, the dataset is organized and gathers efficient information for prediction analysis.

**Step 3:** After preprocessing, the feature selection is carried out.

**Step 4:** Training and testing are performed on the reduced dataset.

**Step 5:** First, 70% of the reduced dataset samples are used for training.

**Step 6:** As testing samples, 30% of the samples are taken from the normalized dataset.

**Step 7:** The classification algorithm is applied to the training and the testing dataset for CYP.

**Step 8:** Finally, generate various reports for the yield prediction of each crop.

## 3. Result and discussions

The predicted outcomes of rice, wheat, pulse, oil-seeds, cotton, and sugarcane crops in Nashik District, Maharashtra, India, in the following years are addressed in this section. In the current research, 30% is used for testing and 70% of the data is used for training to provide yield estimates for all crops mentioned above. The model's inputs include environmental and soil parameters; the output is the yield achieved. The hybrid machine learning algorithms are implemented in Python using the Tensor Flow open-source software package to assess the proposed methodology.

*3.1 Dataset description*

In this paper, the data are obtained from various sources and are combined, resulting in a dataset large enough for the research. From 2000 to 2020, soil and environmental parameters were obtained from the Maharashtra government website (data.gov.in). pH is among the soil characteristics. Rainfall and temperature are the environmental parameters. Similarly, data.gov.in is used to gather crop-related data for wheat, rice, pulses, oilseeds, cotton, and sugarcane crops [26]. The crop yield is calculated using the area farmed (in hectares), the yield obtained (in kg/hectare) and production output (in tonnes) as indicators. The data obtained comprises crop information and includes 10,000 rows, which are recorded in a CSV format as shown in Table 1 and described in Table 2. It is required to clean the data before analyzing it since the raw data is unstructured, noisy, and contains missing data. As a result, the initial stage in machine learning models is data preparation. The experimental findings achieved for forecasting crop production using the XGBoost-SVM-C4.5 model with optimization technique and comparisons with existing models are presented in the next section.

*3.2 Performance metrics*

The effective parameters used to evaluate the performance of the proposed model are Mean Squared Error, Mean Absolute Error, Determination Coefficient, Correlation coefficient and Root Mean Squared Error.

*3.2.1 Root Mean Squared Error (RMSE)*

The RMSE measures the difference between the actual and estimates, exaggerating the presence of outliers.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{16}$$

*3.2.2 Correlation coefficient (R)*

R evaluates the linear relationship between actual values and regression model predictions.

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i(\hat{y}_i-\bar{y})^2}} \quad (17)$$

*3.2.3 Determination Coefficient ($R^2$)*

This parameter is used to evaluate the goodness of the fitting equation among the explicatory factors and the crop yield.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N} y_i^2} \quad (18)$$

*3.2.4 Mean Absolute Error (MAE)*

MAE is the average of differences in estimations (in physical units).

$$MAE = \left(\frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{(n)(\bar{y})}\right) \quad (19)$$

*3.2.5 Mean Squared Error (MSE)*

The MSE is the squared difference among a variable's observed and predicted values divided by the number of values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (20)$$

**Table 2** Dataset description

| | **Description** |
|---|---|
| State | State of production |
| District | Area of production |
| Crops | Type of crop |
| Year | Year of production |
| Hg/ha_yield | Agricultural production for the year in the country |
| $T_{min}$ | Daily minimum temperature average for the year |
| $T_{max}$ | Daily maximum temperature average for the year |
| Pesticides_tonnes | Pesticides utilized on the crop that year in terms of quantity |
| Rainfall | Average rainfall for the year |
| N (Nitrogen) | The overall quantity of nitrogen utilized in agriculture during the year |
| P (Phosphorous) | The overall quantity of phosphate utilized in agriculture during the year |
| K (Potassium) | The overall quantity of potash utilized in agriculture during the year |

**Table 3** Parameter setting of the proposed SA algorithm

| **Description** | **Value** |
|---|---|
| Maximum iterations per temperature | 3000 |
| Max. no. of successful reconfigurations | 100 |
| Cooling rate | 10% |
| Convergence criteria (T < convergence) | $10^{-4}$ |
| Total iterations before reverting to the optimal configuration | 500 |

*3.3 Analysis*

The parameter setting of the proposed SA algorithm is shown in Table 3. Tenfold cross-validation is used to validate the proposed hybrid model. The XGBoost can normalize and choose the most critical characteristics on its own. As a result, the XGBoost model is loaded with the whole dataset and obtained a nearly identical fit but with a much poorer evaluation performance. Simulated Annealing is proposed to simplify and optimize our technique by lowering the number of imported variables. Using this feature selection approach, attributes are chosen to determine accurate environmental and soil factors to predict the crop yield [27, 28].

*3.3.1 Comparison based on soil and environmental physiognomies*

The performance analysis in Table 4 is based on environmental and soil parameters such as pH, K, P, N, rainfall and avg. temp. The results indicate that the hybrid XGBoost-SVM-C4.5 obtained an accuracy of 80.50, 81.25%, 76.12%, 77.15%, 79.50% and 83.40% for rice, wheat, pulses, oilseeds, cotton, and sugarcane crops, respectively. Since the performance of the proposed classifier is far from satisfactory, the SA-hybrid ML approach was employed using the reduced dataset. Interestingly, the accuracy of the classifier algorithm with an optimal subset improves performance significantly compared to the original feature set with a classifier algorithm for various crops.

**Table 4** Comparison of accuracy with and without SA-hybrid ML approach for various crops

| Crops | without SA-hybrid ML | | | with SA-hybrid ML | | |
|---|---|---|---|---|---|---|
| | Original set of features | Reduced feature subset | Accuracy | Original set of features | Reduced feature subset | Accuracy |
| Rice | 35 | - | 80.50% | 35 | 8 | 96.12% |
| Wheat | 22 | - | 81.25% | 22 | 7 | 97.25% |
| Pulse | 12 | - | 76.12% | 12 | 6 | 95.14% |
| Oil-seeds | 14 | - | 77.15% | 14 | 8 | 98.10% |
| Cotton | 30 | - | 79.50% | 30 | 7 | 97.35% |
| Sugarcane | 15 | - | 83.40% | 15 | 5 | 98.25% |

Then, the XGBoost algorithm is coupled with SVM and C4.5 classifiers. These ensemble approaches produce sequential predictions and seek to integrate weak predictive tree systems to learn from their flaws. Generally, decision trees have reduced bias and are more resistant to overfitting. As a result, ensemble approaches involving the execution of several trees are suited for making accurate predictions. Using an optimization technique, the proposed approach effectively decreases bias and variation. The few occurrences per leaf (M) and pruning confidence (C) are the two most essential hyper-parameters to select for the developed framework. The value of 'C' is set to 80, while the value of 'M' is set to 20. The proposed scheme achieves a good outcome with this parameter value.

*3.4 Experimental section*

The outcomes of our work are discussed in this section. The various evaluation metrics RMSE, MSE, MAE, $R^2$ and R metrics were applied to test the proposed models accuracy with four other existing techniques. Table 5, presents the RMSE, MSE and MAE of the existing are slightly higher than the proposed model.

These results proved that the performance of optimized XGBoost-SVM-C4.5 in the CYP has been better than Hybrid MLR-ANN, Gradient Boosting, ANN and RF. Besides, Table 5 also shows the results for the metric of R and $R^2$. Since the average metric of R and $R^2$ is greater for the proposed method than existing it is interpreted that the performance of optimized XGBoost-SVM-C4.5 in these metrics has been better. According to Table 5, the results reveal that the proposed model gets most of the best-correlated models.

**Table 5** Metrics for evaluating the proposed hybrid model's performance in Comparison to existing models

| Algorithm | RMSE | MAE | MAP | R | R2 |
|---|---|---|---|---|---|
| Hybrid MLR-ANN | 0.051 | 0.041 | - | 0.99 | - |
| Gradient Boosting | 0.57 | 0.49 | 0.53 | - | 0.61 |
| ANN | 0.37 | 0.22 | 0.09 | - | 0.62 |
| RF | 0.72 | 0.62 | 0.51 | - | 0.37 |
| Proposed | 0.042 | 0.035 | 0.08 | 0.99 | 0.78 |

Accuracy measures the ratio of accurate predictions made by the model. It is a term that refers to the degree to which the forecasted value is near the actual value. The accuracy metric for the forecasted data predicted by the developed hybrid model is shown in Figure 3.

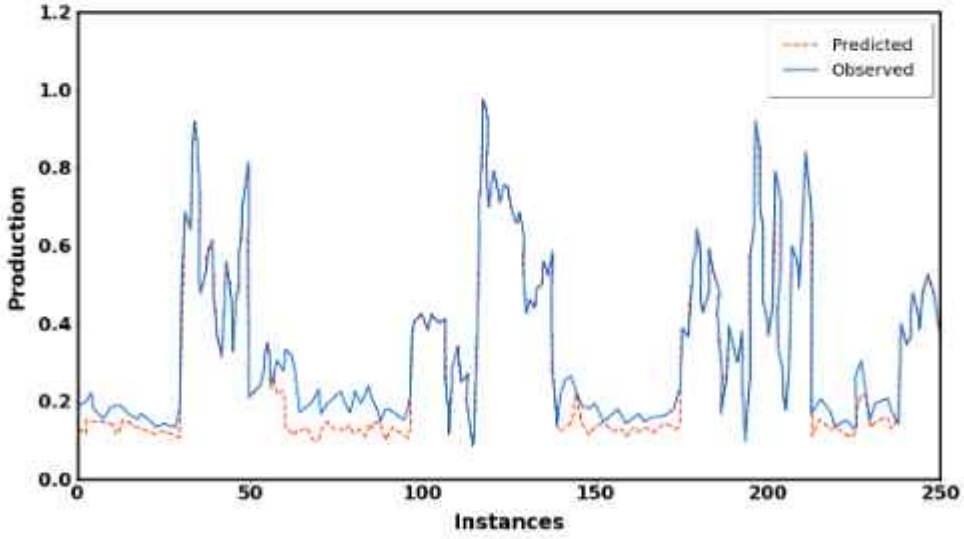


**Figure 3** Proposed model for accuracy measure

*3.4.1 Predicting the yield of crops in Nasik, Maharashtra*

Figures 4-9 shows the predicted production of Rice, wheat, oilseeds, cotton, Sugarcane and cotton in Nashik district, Mumbai by the proposed model in the years 2000 and 2020. The Figures show that the difference between the actual and the predicted yield is less, which tends to reveal that the error rate is lesser when implemented by the proposed model.

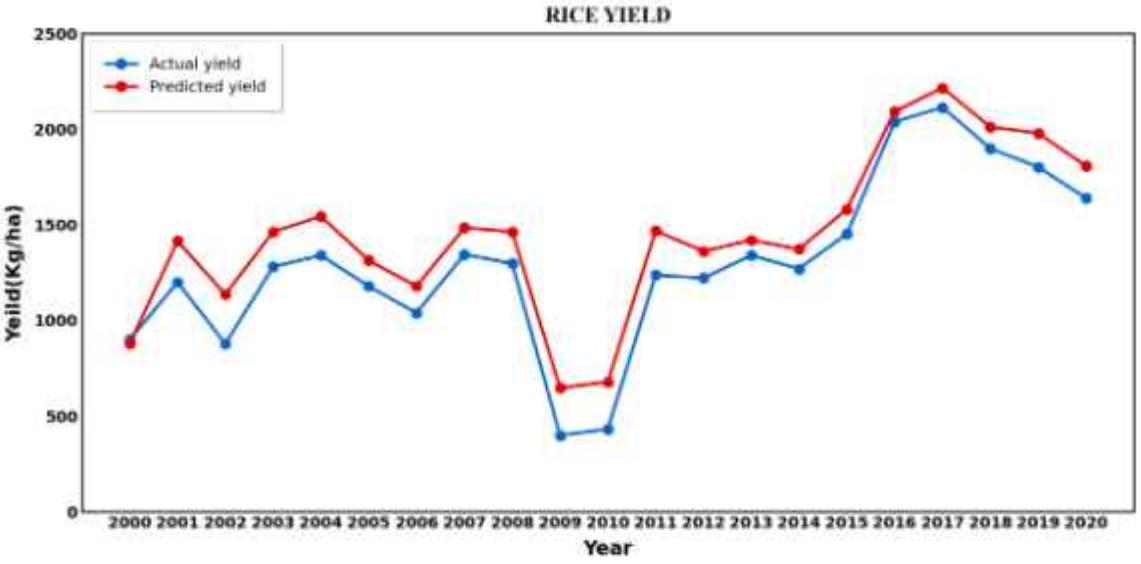


**Figure 4** Predicting yield of Rice

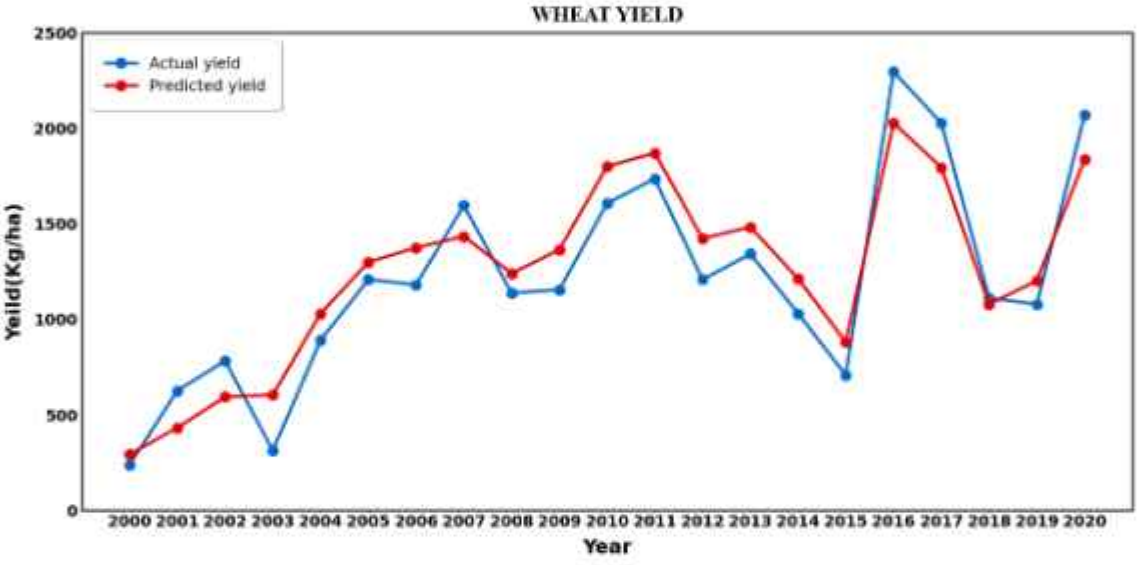


**Figure 5** Predicting yield of Wheat

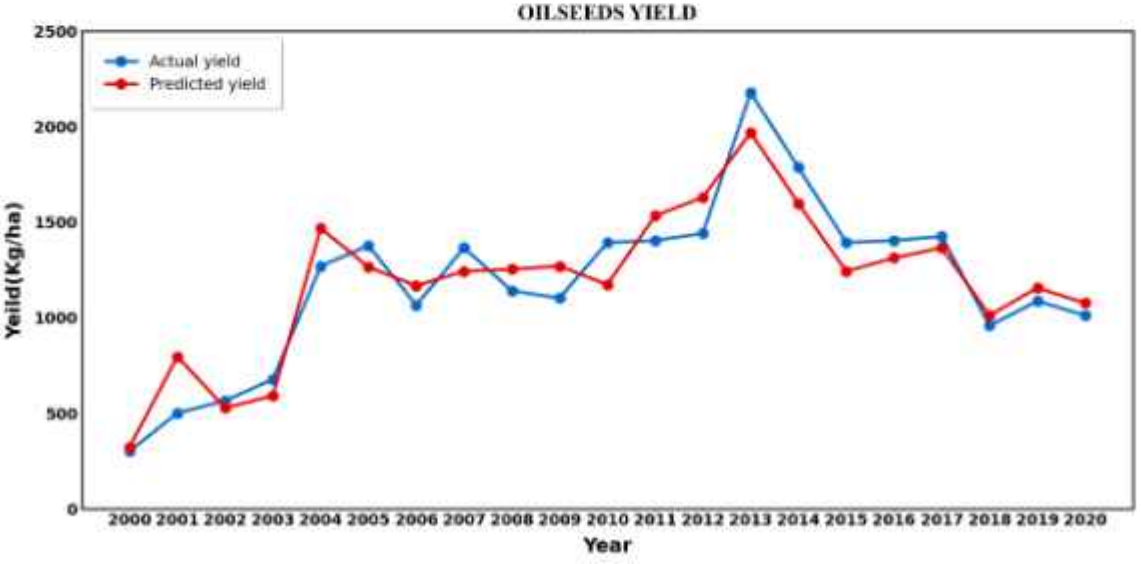


**Figure 6** Predicting yield of Oilseeds

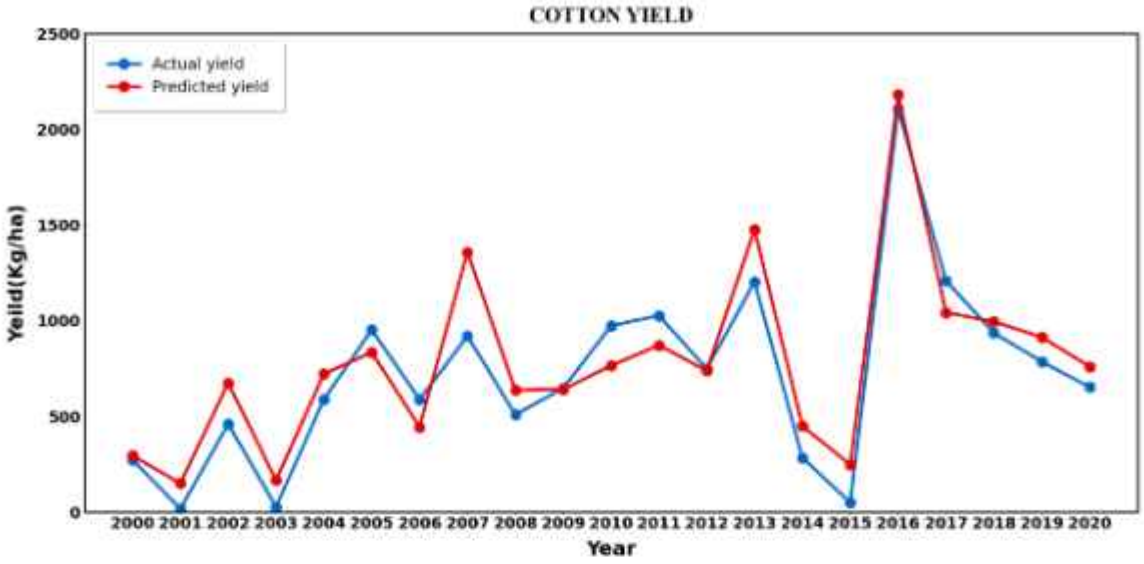


**Figure 7** Predicting yield of Cotton

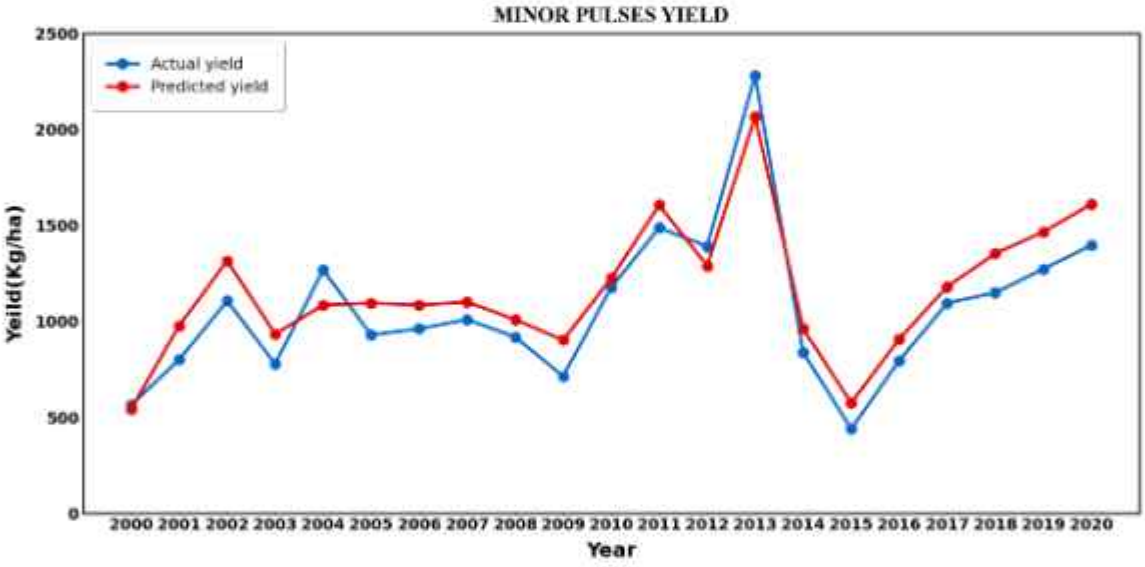


**Figure 8** Predicting yield of Minor pulses

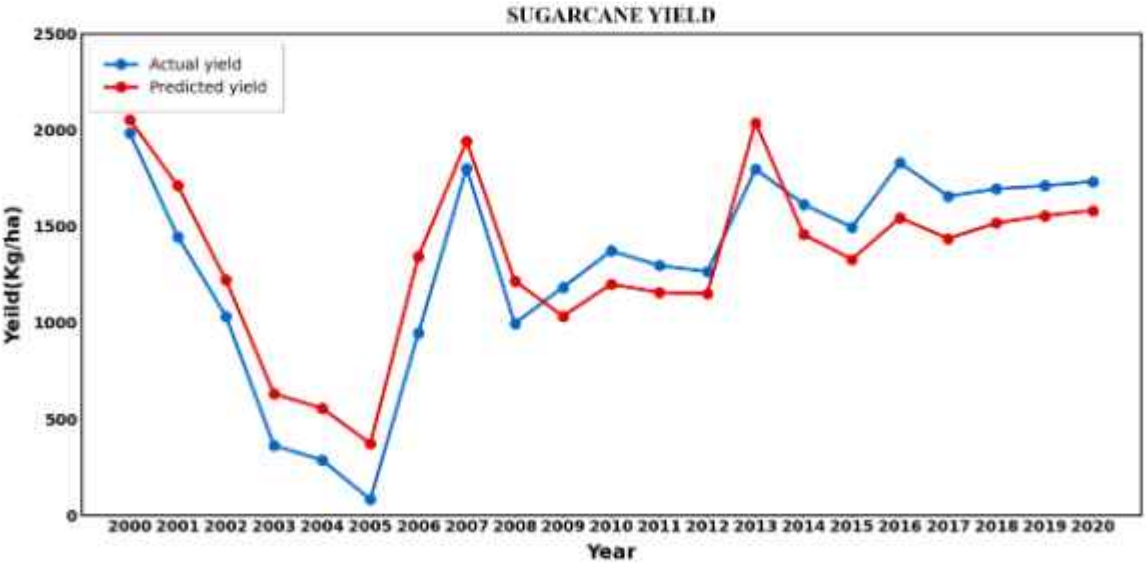


**Figure 9** Predicting yield of Sugarcane

Additionally, the results reveal that the implemented method gives higher prediction accuracy in agricultural production forecasts on current information compared to the existing approaches due to its decreased root mean square error. As a result, this model was chosen to forecast India's agricultural production. The results of the prediction of Indian agricultural products for 2020–2050 using the optimized technique are presented in Figure 10. It demonstrates that agricultural products in India are likely to have a similar rising tendency. This is because the prediction model used for this research indicates that food production in India would rise over the next decade, based on time-series data.

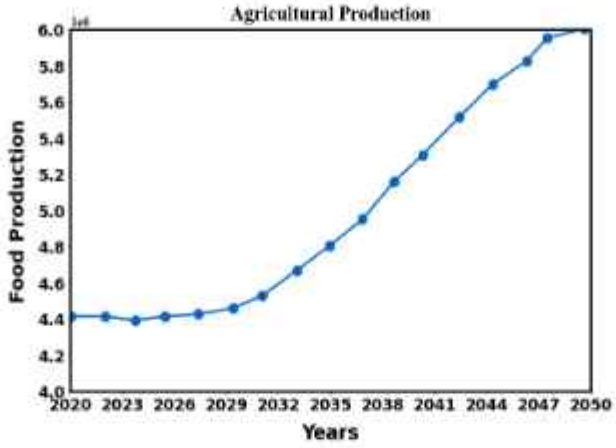


**Figure 10** The result of predicting agricultural production for the next 30 years in India

The results of the experiments show that the proposed hybrid XGBoost-SVM-C4.5 with optimization is better than the other machine learning model. The analysis revealed improved model performance by minimizing complexity and testing losses. Additionally, the predicting accuracy of the model is determined by its performance with minimized errors i.e., RMSE, MSE and MAE error measures and a greater value of correlation coefficient and determination coefficient.

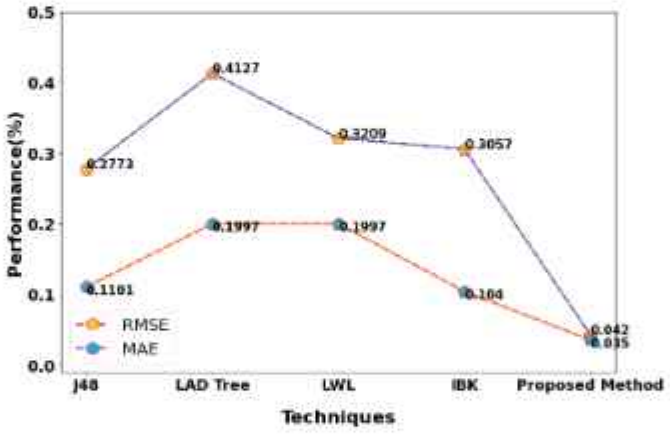


**Figure 11** Differentiation of various ML approaches with proposed

The metrics like RMSE and MAE are employed to compare our proposed approach. Compared with existing approaches, the presented approach yields a better solution. The differentiation between the existing approach and the proposed one is shown in Figure 11. Similarly, compared with prior approaches, the proposed approach takes less time to execute, which is shown in Figure 12.

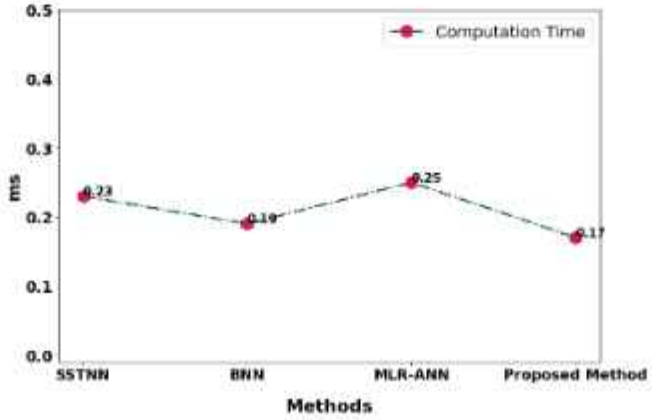


**Figure 12** Comparison of Computational complexity

Hybridizing the XGBoost, SVM, and C4.5 classifier algorithms offers several benefits for crop yield prediction. Firstly, the ensemble framework of a hybrid framework allows it to leverage the benefits of each unique algorithm. SVM excels in managing nonlinear relationships and highly dimensional feature spaces, while C4.5 decision trees offer interpretability and perform well with categorical data. When it comes to spotting complex relationships and patterns in data, XGBoost excels. By combining these algorithms, the hybrid model can provide a more comprehensive and dependable way to capture the different factors influencing crop yield. However, there are challenges with this hybrid approach. One significant drawback of combining multiple algorithms is the increased computational intensity and complexity. The training and inference procedures might need a large amount of computational power, which might limit the model's application in low-resource scenarios. Moreover, the complexity of the hybrid model may make it harder to understand. When considering the combined effects of XGBoost, SVM, and C4.5 on predictions, explaining the model's decisions to stakeholders or end users becomes challenging. It's crucial to strike the correct balance between complexity and interpretability.

## 4. Conclusion

This study suggests a hybrid XGBoost-SVM-C4.5 model to precisely estimate crop production. It raises prediction accuracy and finds the near-optimal error minimum. Moreover, Simulated Annealing is used to fine-tune the C4.5 classifier parameter and extract pertinent features from the dataset. Common metrics for performance were used to evaluate the improved hybrid methods' prediction outcomes. This study was conducted primarily in Nashik, Maharashtra, India, but it could be extended to other agrarian areas. Predictions about India's future years may be provided to the government, which can help plan and implement policies that will meet the nation's food needs. Future studies may also look into alternative machine learning models to improve the analysis and efficacy.

## 5. Acknowledgements

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

## 6. Reference

[1] Elavarasan D, Vincent PMDR. Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access. 2020;8(1):86886-901.
[2] Shahhosseini M, Hu G, Huber I, Archontoulis SV. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Sci Rep. 2021;11(1):1606.
[3] Hara P, Piekutowska M, Niedbała G. Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. Land. 2021;10(6):609.
[4] Bhojani SH, Bhatt N. Wheat crop yield prediction using new activation functions in neural network. Neural Comput Applic. 2020;32(17):13941-51.
[5] Oikonomidis A, Catal C, Kassahun A. Hybrid deep learning-based models for crop yield prediction. Appl Artif Intell. 2022;36(1):2031822.
[6] Elavarasan D, Vincent PMDR. A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. J Ambient Intell Human Comput. 2021;12(11):10009-22.
[7] Shook J, Gangopadhyay T, Wu L, Ganapathysubramanian B, Sarkar S, Singh AK. Crop yield prediction integrating genotype and weather variables using deep learning. PLoS One. 2021;16(6):e0252402.
[8] Iniyan S, Varma VA, Naidu CT. Crop yield prediction using machine learning techniques. Adv Eng Softw. 2023;175:103326.
[9] Champaneri M, Chachpara D, Chandvidkar C, Rathod M. Crop yield prediction using machine learning. Int J Sci Res. 2020;9(4):645-8.
[10] Elavarasan D, Vincent PMDR, Srinivasan K, Chang CY. A hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling. Agriculture. 2020;10(9):400.
[11] Elavarasan D, Vincent PMDR. Fuzzy deep learning-based crop yield prediction model for sustainable agronomical frameworks. Neural Comput Applic. 2021;33(20):13205-24.
[12] Shidnal S, Latte MV, Kapoor A. Crop yield prediction: two-tiered machine learning model approach. Int J Inf Technol. 2021;13(5):1983-91.
[13] Sivanandhini P, Prakash J. Crop yield prediction analysis using feed forward and recurrent neural network. Int J Innov Sci Res Technol. 2020;5(5):1092-6.
[14] Suganya M, Dayana R, Revathi R. Crop yield prediction using supervised learning techniques. Int J Comput Eng Technol. 2020;11(2):9-20.
[15] Gulati P, Jha SK. Efficient crop yield prediction in India using machine learning techniques. Int J Eng Res Technol. 2020;8(10):24-6.
[16] Obsie EY, Qu H, Drummond F. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. Comput Electron Agric. 2020;178:105778.
[17] Ma Y, Zhang Z, Kang Y, Özdoğan M. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. Remote Sens Environ. 2021;259:112408.
[18] Paudel D, Boogaard H, de Wit A, Janssen S, Osinga S, Pylianidis C, et al. Machine learning for large-scale crop yield forecasting. Agric Syst. 2021;187:103016.
[19] Hu T, Zhang X, Bohrer G, Liu Y, Zhou Y, Martin J, et al. Crop yield prediction via explainable AI and interpretable machine learning: dangers of black box models for evaluating climate change impacts on crop yield. Agric For Meteorol. 2023;336:109458.
[20] Morales A, Villalobos FJ. Using machine learning for crop yield prediction in the past or the future. Front Plant Sci. 2023;14:1128388.
[21] Suresh G, Kumar AS, Lekashri S, Manikandan R. Efficient crop yield recommendation system using machine learning for digital farming. Int J Mod Agric. 2021;10(1):906-14.

[22] Hammer RG, Sentelhas PC, Mariano JCQ. Sugarcane yield prediction through data mining and crop simulation models. Sugar Tech. 2020;22(2):216-25.
[23] Feng P, Wang B, Liu DL, Waters C, Xiao D, Shi L, et al. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agric For Meteorol. 2020;285-286:107922.
[24] Shastry KA, Sanjay HA. Hybrid prediction strategy to predict agricultural information. Appl Soft Comput. 2021;98:106811.
[25] Murali P, Revathy R, Balamurali S, Tayade AS. Integration of RNN with GARCH refined by whale optimization algorithm for yield forecasting: a hybrid machine learning approach. J Ambient Intell Humaniz Comput. 2020:1-13.
[26] Khosla E, Dharavath R, Priya R. Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. Environ Dev Sustain. 2020;22(6):5687-708.
[27] Wei MCF, Maldaner LF, Ottoni PMN, Molin JP. Carrot yield mapping: a precision agriculture approach based on machine learning. AI. 2020;1(2):229-41.
[28] Prasad NR, Patel NR, Danodia A. Crop yield prediction in cotton for regional level using random forest approach. Spat Inf Res. 2021;29(2):195-206.