INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Improving Crop Yield Prediction Models with Optimization-Based Feature Selection and Filtering Approaches

Anuj Mehla[1], **Sukhvinder Singh Deora**[2]*, **Sandeep Dalal**[3]

1 Research Scholar, Department of Computer Science & Application, Maharshi Dayanand University, Rohtak, Haryana, India
2 Assistant Professor, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India
3 Associate Professor, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India

OPEN ACCESS

Received: 26-06-2023

Accepted: 05-10-2023

Published: 22-12-2023

Citation: Mehla A, Singh Deora S, Dalal S (2023) Improving Crop Yield Prediction Models with Optimization-Based Feature Selection and Filtering Approaches. Indian Journal of Science and Technology 16(47): 4512-4524. https://doi.org/10.17485/IJST/v16i47.1602

* **Corresponding author**.

sukhvinder.singh.deora@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: To analyze the impact of various factors on crop yield and provide insights for improving crop production in the region. **Methods:** This research employs feature selection algorithms, machine learning models, and feature extraction algorithm Principal Component Analysis (PCA) technique to identify the key factors affecting crop yield in India. Data from the Indian Meteorological, Statistical, and Agriculture Departments spanning five decades are analyzed to provide valuable insights to policymakers and farmers. This research analyzed 20 factors in determining their impact on crop yield in the Indian economy. Three feature selection algorithms were used to identify the essential factors: forward feature selection, backward feature selection, and recursive feature elimination. These three algorithms were used to select the most important factors from the Twenty selected factors, and then three ML models were used to estimate the accuracy of the feature selection algorithms: Random Forest, XGBoost, and Multiple Linear Regression. Principal Component Analysis (PCA) was used for the dimensionality reduction of the features. RMSE, MAPE, MAE, and R2 were used to measure the feature selection method's performance. **Findings:** Out of the three machine learning algorithms, the Random Forest algorithm with the forward feature selection algorithm provided the highest model accuracy of 98.415 percent. Moreover, compare the combination of three machine learning algorithms and different feature selection algorithms. **Novelty:** Our approach to predicting crop yield is based on a combination of Feature Selection, PCA, and Machine Learning algorithms. This proposed research utilizes Feature Selection algorithms to identify the most crucial features among 20 available options and then apply Machine Learning models to make accurate predictions based on these features.

Keywords: Recursive Feature Elimination; Forward Feature Selection; Principal Component Analysis; Random Forest Regression; XGBoost; Crop yield prediction; Multiple Linear Regression

1 Introduction

Researchers employ self-learning strategies to establish relationships between historical agricultural yield produce and influencing data to improve crop yield prediction. Even though machine learning algorithms have outperformed biophysical models, selecting appropriate algorithms and identifying crucial dataset characteristics to enhance the learning algorithm remains difficult[1,2]. However, some obstacles still exist to overcome in developing high-performance predictive models. The selection of suitable machine learning algorithms is a difficulty. Identifying crucial dataset characteristics to enhance the learning algorithm is a further obstacle. For instance, Klompenburg et al. (2016) analyzed 50 yield prediction studies using various machine learning algorithms and found that models with more features occasionally attained superior performance.

To determine the model with better performance, it is necessary to evaluate models with varied subsets of features. This is coherent with the significance of dimensionality reduction in machine learning, especially when dealing with datasets containing many attributes[3] . Dimensionality reduction enables faster training, reduces complexity, facilitates interpretation, and maximizes accuracy by selecting suitable subsets while avoiding overfitting. The selection of features is essential for developing accurate and robust ML models[4]. It aims to identify the most significant factors influencing crop yield to develop targeted and interpretable predictive models. Feature selection and feature extraction are two common techniques for dimensionality reduction[5]. Excluding certain features may result in data loss, although feature selection preserves vital information pertinent to individual components. Furthermore, feature extraction reduces the feature space size without sacrificing essential information, but it is frequently necessary to recover individual feature contributions. Feature selection algorithms are categorized as (1) employing feature selection without assessing its impact, (2) determining the optimal feature selection technique, and (3) assessing the efficacy of combining feature selection and feature extraction. Table 1 compares features, feature selection methods, ML algorithms, and performance metrics for predicting the accuracy of related work.

A similar approach is applied in predicting optimal crop yield productivity. Additional empirical research and exhaustive evaluations are required to compare the efficacy of these approaches and investigate the advantages of combining them. Additional research is required to understand their synergistic effects and determine the most effective combination of these techniques for enhancing crop yield prediction models.

Utilizing feature selection without evaluating its value: This group of studies has utilized a variety of feature selection algorithms to develop crop yield forecasting. In 2019, Lingwal et al.[6] utilized various methods to select the ten relevant attributes out of 18 agricultural and weather-related factors that accurately predict paddy yield in Punjab State, India.

Two research studies conducted by P. S. Maya Gopal and R. Bhargavi[7] examined various approaches to forecast paddy crop yield. The techniques assessed were backward feature elimination, forward feature selection, variance inflation factor, correlation-based feature selection, and random forest. These studies provide valuable insights into the different methods for predicting crop yield, which could help farmers make more informed decisions and maximize their harvest[5,7].

Effectiveness of combining feature selection and extraction examined for agricultural yield prediction.

For example, In one study, Corrales et al. (2018)[8] improved soybean yield prediction by combining the wrapper method, support vector machines, and linear regression. They also carefully selected subsets of the most representative variables to enhance their results.

In the Indian state of Tamil Nadu, P. S. Maya Gopal and R. Bhargavi[5,7] developed crop yield prediction models with a modified R2 of 85% (or 84%) by utilizing selected features (or all features). Even with basic datasets, Whitmire et al.[9] showed that integrating machine learning with feature selection enhanced alfalfa yield prediction. Srivastava et al.[10] analyzed the significance of winter wheat yield characteristics in Germany and built models based on subsets of the most significant components. The accuracy of these models remained comparable to that of models based on comprehensive features.

The findings of this study indicate that feature selection and feature extraction may be valuable techniques for crop yield prediction. Nonetheless, additional research is required to comprehensively compare the efficacy of these approaches and investigate the potential synergies that may result from their combination. Several crucial factors in India inspired the decision to pursue this research: food safety, agricultural planning, market stability, climate change resilience, and nurturing economic development.

Table 1. Comparative analysis of related work

	Features used	**Feature Selection methods used**	**Machine Learning algorithms used**	**Performance metrics used**
(5)	• cultivation (hectare), • tanks(nos.), • canal length (m), • tube wells (nos.), • production (tons), • open wells (nos.), • temperature (max, min, average), • rain fall (mm), • solar radiation(W/m2), • potassium, • phosphorus, • nitrogen, • seed quantity (kg).	• Sequential Forward FS, • Sequential Backward Elimination FS, • Correlation based FS, • Random Forest Variable Importance, • Variance Inflation Factor	• Multiple Linear Regression, • Artificial Neural Network, • M5Prime	• RMSE, • MAE, • R • RRMSE • R2
(6)	• Chem fert, • Temperature(maximum, minimum), • mean evaporation in mm, • mean sunshine duration in hours, • mean wind speed in km/h • precipitation number of days(0.1-0.2,>= 0.3mm), • Total rainfall per month in mm.	• Random Forest • Correlation-based feature selection • Recursive feature elimination algorithm	• RaNN • Multiple Linear Regression • Random Forest • Decision Tree • Boosting Regression • Support Vector Machine Regression • Ensemble Learner • Artificial Neural Network	• RMSE, • R^2 • MAE
(8)	• Nbgrmax, Stlevdrp, • Stflodrp, • Stdrpmat, • Masec(n), • Lai(n), • Qnplante, • Qfix, • Zrac, • Jul • ,Raint, Etpp(n), • Precip, • Ep, • AvgTemp, • MinTemp, • MaxTemp, • Swfac, • Inn, • Mafruit	• Pearson coefficient, • Spearman coefficient, • Entropy-based information gain, • Random Forest, • M5 decision tree, • Least Absolute Shrinkage and Selection Operator, • Recursive feature elimination	• Linear Regression(LR), • Support vector regression(SVR), • Back Propagation neural network (BPNN), • Random Forest(RF), • Least Absolute Shrinkage and Selection Operator (LASSO), • M5 decision tree.	• R2, • RMSE • MSE • MAE

Continued on next page

Table 1 continued

[9]	• Soil moisture(average), • Day length(hrs), • Percent cover, • Solar radiation, • Julian day, • Rainfall, • Temperature(minimum, maximum), • No. of days time since sown, since last harvest, • Average Air Temperature	• Correlation-based method, • ReliefF method, • ZeroR classifier	• Linear Regression, • Regression trees, • Support vector machines, • Neural networks, Bayesian regression, • Nearest neighbors	• R, • R2, • MAE
[10]	• Temperature(maximum, minimum), • radiation, • precipitation, • relative humidity, • wind speed data, • Volumetric (%), • Permanent wilting point (LL), • crop available warer at field capacity (DUL), • saturation point (SAT), • bulk density (BD) with a depth of 1.3 m, • Wheat sowing, • harvest, • flowering	• Correlation coefficient	• Proposed CNN, • Deep Neural Network(DNN), • XGBoost	• RMSE, • MAE, • Correlation coefficient metrics

In summary, previous research has used feature selection and extraction techniques to reduce the dimensionality of crop yield prediction models. Using a combination of feature selection and feature extraction, this study examines the effectiveness of feature reduction techniques. It shall compare the predictive performance of three machine learning algorithms (Random Forest Regression, XGBoost, and MLR) to identify the optimal method for predicting maximal crop yield production. It also intends to reveal the underlying relationships between particular characteristics and crop yield, thereby shedding light on the critical factors influencing agricultural productivity. It eliminates a novel method for predicting crop productivity using machine learning algorithms and feature selection techniques. Feature selection is performed using Random Forest Variable Importance (RFVARIMP), Recursive Feature Elimination (RFE), and Forward Feature Selection (FFS). At the same time, A feature extraction technique, Principal Component Analysis (PCA), is implemented to improve model interpretability and performance. Python is used for data analysis due to its simplicity, adaptability, and large user and developer community. Table 2 List of acronyms used in this study.

2 Methodology

This section briefly explains the optimization-based feature selection and filtering approaches for predicting crop yield.

Assuming the availability of data on both crop yield and factors that affect crop yield are available during the same timeframe, It is possible to create models for predicting crop yield using machine learning (ML). This can be achieved using the general Equation:

$$\mathrm{PY} = \mathrm{F}(\mathrm{px1}, \mathrm{px2}, \mathrm{px3}, \ldots\ldots, \mathrm{pxn}) \tag{1}$$

In Equation (1), PY(Yield Productivity) represents the yield productivity of wheat. The function F refers to the learner or algorithm employed for predicting crop yield based on the provided features, where px1, px2, px3,…., and pxn represent the individual features, such as maximum/minimum temperature, rainfall, humidity, and so on.

To address these challenges, researchers have developed feature dimension reduction techniques that eradicate interferences from the original feature dataset and generate a lower-dimensional feature space while preserving essential information. Although this issue has garnered substantial attention worldwide, it must be addressed. Therefore, this study proposes an effective solution by building upon previous achievements. Dimensionality reduction is achieved through Feature selection

Table 2. Acronyms

Acronym	Definition
PCA	Principal Component Analysis
FFS	Forward Feature Selection
RFVARIMP	Random Forest Variable Importance
RFE	Recursive Feature Elimination
MLR	Multiple Linear Regression
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
R2	R-squared value
PC	Principal Component
FS	Feature Selection
FE	Feature Extraction
FSE	Feature Selection and Extraction
PY	Yield Productivity
AIC	Akaike's Information Criterion

(FS) and Feature Extraction (FE). The former seeks to select a subset of relevant features, while the latter transforms the original elements into a lower-dimensional feature space. To further enhance performance, the two techniques can be combined, resulting in what is referred to as FSE (feature selection and extraction). When dealing with high-dimensional datasets containing numerous redundant or extraneous components, the learning algorithm can adversely impact(11,12), leading to unstable model training, loss of precision(13), overfitting, increased memory requirements, and computational costs.

The paper presents a methodology employing FSE to develop crop yield prediction models using machine learning algorithms such as Random Forest, XGBoost, and MLR. The feature subset is obtained through feature selection techniques such as FFS, RFE, and RFVARIMP. Additionally, FE using PCA is applied. The best models are then selected based on metrics such as root-mean-square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and R-squared value (R2). The best model is determined by comparing the performance of the FSE model to that of other models.

Numerous ML algorithms have been utilized to construct agricultural yield prediction models, yet a definitive conclusion has yet to be drawn on the overall model. Our research addresses this issue by investigating multiple ML algorithms, including MLR, XGBoost, and Random Forest, References(14,15) ML methods.

2.1 Data collection and pre-processing

This stage describes the data used in the study, including its sources and the pre-processing steps taken to clean and prepare the data for analysis.

2.1.1 Data set collection

Data about the crop was gathered to predict crop yield. It was sourced from two entities: the Indian Meteorological Department in the Ministry of Earth Sciences(16) and the Directorate of Economics and Statistics in the Department of Agriculture and Farmers' Welfare, Ministry of Agriculture and Farmers' Welfare, Government of India(17) spanning the years 1967 to 2018. A new dataset was created by merging the relevant information from the original dataset specific to India. It encompasses 52 years and 1092 samples, including planted area, productivity, yield, minimum/maximum temperature, humidity, monthly precipitation, and rainfall (mm). Table 3 in the research article provides a comprehensive list of all the features considered for this study and their descriptions.

2.1.2 Data pre-processing

In this step, the dataset is cleaned and prepared using Python libraries for analysis, including missing value imputation, normalization, and data scaling(18).

Table 3. Dataset Description

Feature ID	Type of Feature	Description
Year	Forecaster	The total years 1967 to 2018
Area(000 Hectares)	Forecaster	The total area in hectares in which wheat was grown that year
Production(000 Tonnes)	Forecaster	Total production of the year in a ton
Yield(Kg./Hectares)	Target /Predicted feature	Total yield production per hectare of the year in kg
rainfall(OND)	Forecaster	An Average of the rainfall recorded in October, November, and December in mm
rainfall(JF)	Forecaster	An Average of the rainfall recorded for January and February in mm
rainfall(MAM)	Forecaster	An Average of the rainfall recorded for March, April, and May in mm
AMT(OND)	Forecaster	An average of the maximum daily temperatures recorded in October, November, and December.
AMT(JF)	Forecaster	An average of the maximum daily temperatures recorded in January and February
AMT(MAM)	Forecaster	An average of the maximum daily temperatures recorded in March, April, and may
mint(OND)	Forecaster	An average of the minimum daily temperatures recorded in October, November, and December
mint(JF)	Forecaster	An average of the minimum daily temperatures recorded in January and February
mint(MAM)	Forecaster	An average of the minimum daily temperatures recorded in March, April, and May
Precoat	Forecaster	October's daily average precipitation
prenoon	Forecaster	November's daily average precipitation
predoc	Forecaster	December's daily average precipitation
Prejean	Forecaster	January's daily average precipitation
prefab	Forecaster	February's daily average precipitation
preach	Forecaster	March daily average precipitation
prepare	Forecaster	April daily average precipitation
prepay	Forecaster	May daily average precipitation

2.2 Feature Engineering

During this process, the mining of available data for additional elements can increase the accuracy of the predictive model. This may involve combining or changing data sets, changing variables that have already been measured, or finding factors that have not yet been measured.

2.2.1 Feature Selection Algorithm

A unique subset of features can be selected using FS methods to increase crop yield productivity. This approach produces more precise results and requires less computational effort. Feature selection presents three primary benefits: (i) increased training speed for the ML algorithm, (ii) simplified model complexity, and (iii) clear model interpretation. Filter, wrapper, and embedded methods are three types of FS algorithms (8,19,20). The filter approach selects relevant features based on individual features during data processing and integrates them during model training (3). This method is computationally efficient and independent of the learning method employed.

In contrast, wrapper algorithms add or remove features and assess their impact on the model's performance. These methods consider feature dependencies and their contribution to model generation. While wrapper techniques outperform filter methods, they can be computationally intensive and lead to overfitting. On the other hand, Embedded methods are a hybrid approach that combines the benefits of filter and wrapper methods. These methods determine feature importance during training and are often optimized for specific learning machines, thereby reducing computational expenses and enhancing feature selection efficiency. Careful selection of an appropriate subset of features can enhance model accuracy and reduce overfitting. Table 4 provides a comprehensive list of features selected by various feature selection algorithms.

Table 4. Selected number of features with feature selection methods

FSA/ Features	FFS	RFE	RFVariableImp
Year	●	●	●
Area (000 Hectares)	●	●	●
Production (000 Tonnes)	●		●
rainfall (OND)		●	
rainfall (JF)		●	●
rainfall (MAM)	●		
AMT(OND)		●	●
AMT(JF)		●	●
AMT(MAM)	●		●
mint (OND)		●	●
mint (JF)	●		
mint (MAM)		●	
Preoct		●	●
Prenov	●	●	●
Predec		●	●
Prejan	●	●	
Prefeb	●		
premarch		●	
Preapril		●	
Premay			●

Description: FSA- Feature Selection Algorithms, FFS- Forward Feature Selection algorithm, RFE:-Recursive Feature Elimination, RFVariableImp- Random Forest Variable Importance.

2.2.2 Feature Extraction

PCA, which has been extensively used, is one of the most frequently used methodologies for feature extraction. This research uses PCA for feature extraction because it can reduce high-dimensional data to a lower-dimensional representation. PCA generates Principal Components (PCs) that have two crucial characteristics. First, each PC is derived as a linear combination of input features. Second, the PCs are independent of one another, which eliminates any redundant features. In addition, by combining PCA with Random Forest, this proposed research addresses issues such as overfitting, accelerating the training process, and eradicating irrelevant or redundant data features. Ultimately, these measures contribute to improved model efficacy[21]. Table 5 provides a list of features and their importance.

Algorithm

```
Input: Number of parameters include climate data and crop data PX: = {px1,px2,px3,.......pxn}
Output: the value of the predicted yield PY.
1. Input the total number of features correlated with crop yield.
PX:={px1,px2,px3,.......pxn}
2. Use the feature optimization or selection algorithm (FOA).
3. If FOA: = Forward feature selection (FFS), go to step 7.
Else if FOA: = Recursive feature elimination (RFE), then go to step 8.
Else FOA: = Random Forest Feature Importance(RFFeatureImp), then go to step 9.
4. Apply PCA for dimensionality reduction on each feature selection algorithm.
5. Select the appropriate Machine Learning Regression Algorithms (MLRA).
If MLRA:= MLR(Multiple Linear Regression), apply the MLR model and go to step 6.
Elseif MLRA:= XGBoost Regression, apply XGBoost Model algorithm and go to step 6.
Elseif MLRA:= Random Forest Regression algorithm, apply Random Forest Regression Model algorithm and go to step 6.
6. Determine the yield forecast.
{
```

```
6.1.Apply distinct feature dimensionality reduction subsets.
6.2. Calculate the Predicted yield Value PY.
6.3. Determine the precision using performance metrics.
6.3.1. RMSE (Root Mean Square Error) = $\sqrt{\frac{1}{N}\sum_{j=1}^{N}}\,(py\_pred - py\_true)^{2}$
6.3.2. MAE (Mean absolute error) = $\frac{1}{N}\sum_{j=1}^{N}\,(|py\_pred - py\_tru\,e|)$
6.3.3. MAPE (Mean absolute percentage error) = $\frac{1}{N}\sum_{j=1}^{N}\left(\frac{\text{py\_pred-py\_true}}{\text{py\_true}}\right) * 100$
6.3.4. R2 (R-squared) = $1 - \frac{\sum_{\mathrm{j=1}}^{\mathrm{N}}(|\ \text{py\_true-py\_pred}\ |)2}{\sum_{\mathrm{j=1}}^{\mathrm{N}}(|\ \text{py\_true-mean(py\_true}\ )|\big)2}$
}
7. Procedure FFS
PX: = {px1,px2,px3,…….pxn}
a) Initialize an empty set of features.
b) Add the best feature to the set. Train the regression model and calculate AIC (Akaike's Information Criterion) for the trained model using the AIC=2k-2ln(L), ln(L) is the natural log of the function for the model.
c) Select the best feature by calculating the AIC for the model trained on the current set of features plus the candidate feature and select the feature that results in the lowest AIC.
d) Add the selected feature to the set of features.
e) Repeat (steps 2 to 4) the process of selecting the best feature, adding it to the set, and calculating the AIC until all features have been considered or a stopping criterion has been reached.
f) Evaluate the final set of features using the performance metric of choice, such as accuracy, F-score, or mean squared error.
8. Procedure RFE
PX:={px1,px2,px3,…….pxn}
a) Initialize the random forest model.
b) Create an initial set of features in the dataset.
c) Train the model on the initial set of features and evaluate model performance using performance metrics.
d) Identify the model's weakest feature with the lowest weight or coefficient and remove it from the feature set.
e) Repeat steps 3 and 4 based on the number of features or the model's performance.
f) Select the final set of features after removing all the weakest features.
9. Procedure RFFeatureImp
PX:={px1,px2,px3,…….pxn}
9.1. Determine the significance of each feature using the random forest regression algorithm.
9.2. Locate the purity level of the node.
9.3. If the node's purity level > the median of the node's purity
9.3.1. Select the feature.
Return PXsubset : = { px1,px2,px3,…….pxn}
10. STOP
```

Table 5. Random Forests Variable Importance

Feature	Importance
Production(000 Tonnes)	0.47491
Year	0.316405
Area(000 Hectares)	0.171004
AMT(MAM)	0.005789
premay	0.00354
preapril	0.003497
Preoct	0.002737
mint(MAM)	0.002598
mint(OND)	0.002426
mint(JF)	0.002025
prenov	0.001988
AMT(JF)	0.001911
predec	0.001838
rainfall(JF)	0.001636

Continued on next page

Table 5 continued

premarch	0.001589
AMT(OND)	0.001566
rainfall(OND)	0.001413
prefeb	0.001252
rainfall(MAM)	0.001132
prejan	0.000745

2.3 Model Selection and training

2.3.1 Model Selection

Model selection is done by comparing three models (RF, XGBoost, and MLR)[22–26]. It is discussed that the random forest regression model is the best machine learning algorithm for predicting the accurate crop yield and can be used for training, testing, and validation.

2.3.2 Model training

All three ML models are trained using the processed data and engineered features in this phase. Optimal parameters and hyperparameters are selected during this phase, following PCA-based dimension reduction. The following modifications were made to the existing algorithms:

1. The forward feature selection algorithm was used to select the optimal subset of features.
2. The Recursive Feature Elimination algorithm was used to iteratively remove features until the desired number of features was reached.
3. The Random Forest Feature Importance algorithm was used to select the most important features based on the yield value of the dependent feature.
4. PCA-based dimension reduction was used to improve the performance of the model.

The training and testing of the models were done using a 70-30 split of the dataset. The models were evaluated using RMSE, MAE, R2, and MAPE as performance metrics. The comparison with gold standards was done by comparing the results of the models with the actual crop yield. Figures 1, 2 and 3 display the RMSE, MAE, MAPE, and R^2 values of all three algorithms- namely, Random forest, XGBoost, and MLR - with distinct feature selections.

During this phase, the effectiveness of the models was assessed using RMSE, MAE, and MAPE as performance metrics. Finally, As a result, it was found that the Hybrid Random Forest algorithm with Forward Feature Selection outperforms the other algorithms.

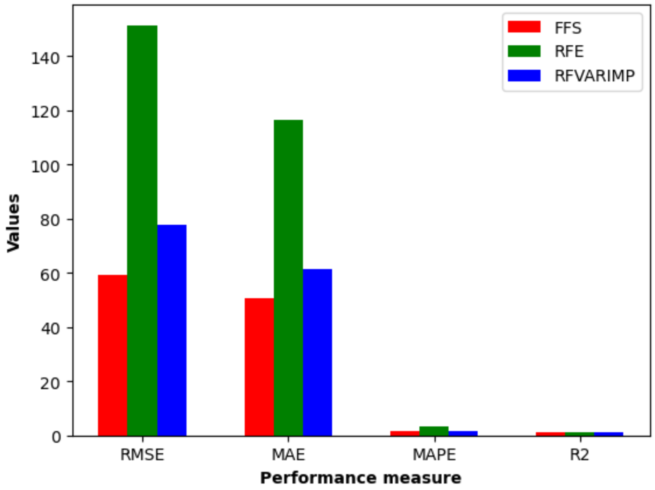

Fig 1. The performance metrics of the Random Forest model

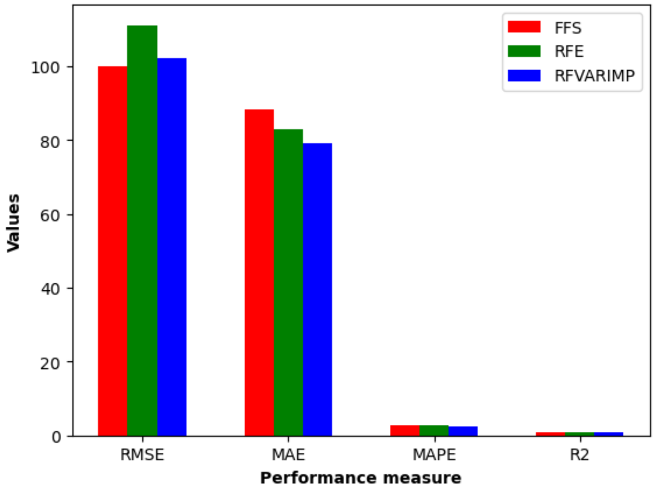

Fig 2. The performance metrics of the XGBoost model

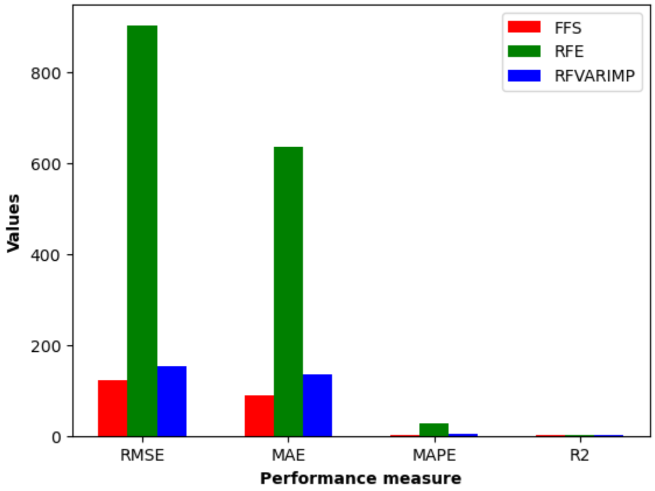

Fig 3. The performance metrics of the MLR model

3 Results and Discussion

The efficiency of three various ML methods- MLR, XGBoost, and Random Forest – for predicting crop yield is evaluated and compared based on different evaluation metrics for regression performance. It used weather and crop parameters details from the Directorate of Economics and Statistics in the Department of Agriculture and Farmers' Welfare, Ministry of Agriculture and Farmers' Welfare, and the Government of India[(17)], respectively. The datasets contain parameters such as planted area, productivity, yield, minimum/maximum temperature, humidity, monthly precipitation, and rainfall (mm). Seven hundred sixty-four instances were acquired for training, and 328 were kept for testing. This proposed research applied three feature selections for optimum feature selection (FFE, RFE, and RFVImportance) for each ML algorithm, and Table 2 lists all the selected features of the three algorithms. PCA algorithm is then applied to reduce high-dimensional data to low-dimensional data. The experiment was conducted on a core i7 or higher hardware with a minimum of 4GB of RAM and 500 GB of hard drive using Python 3. The effectiveness of the models based on the adjusted RMSE, MAE, MAPE, and R2 values show the variance that the model explains.

A. RMSE

The term "RMSE" refers to root-mean-square error, a way to gauge the variance between predicted and actual values while detecting any anomalies or outliers within the data.

RMSE: $\sqrt{\frac{1}{N}\sum_{j=1}^{N}}\left(py\_pred - py\_true\right)^2$

B. MAE

The mean absolute error (MAE) measure determines the average absolute difference between predicted and actual values. This helps to provide an understanding of the typical magnitude of prediction errors.

MAE: $\frac{1}{N}\sum_{j=1}^{N}\left(|py\_pred - py\_tru\,e|\right)$

C. MAPE

The MAPE(mean absolute percentage error) calculates the average percentage difference between predicted and observed values.

MAPE: $\frac{1}{N}\sum_{j=1}^{N}\left(\frac{py\_pred-py\_true}{py\_true}\right)*100$

In the given context, ”py_pred” represents the predicted value, ”py_true” represents the actual value, and ”N” represents the number of data points.

A lower MAPE value indicates a more accurate forecast. A MAPE value of 0% represents a perfect forecast, but it is uncommon to achieve this value in practice. A 10% or less MAPE value is generally regarded as good accuracy.

D. R2

R2 performance measure measures how well a regression model fits the data.

$\mathrm{R}^2{:}1-\frac{\sum_{j=1}^{N}(|py\_true-py\_pred|)2}{\sum_{j=1}^{N}(|py\_true-mean(py\_true)|)2}$

Here, $\sum_{j=1}^{N}(|py\_true-py\_pred|)2$ is the sum of squared differences between predicted and actual values.

Figure 4 compares RMSE, MAE, MAPE, and R2 values of all three feature selection algorithms (FFS, RFE, and RandomForestVariable Importance) for all three ML models (Random Forest, XGBoost, and MLR).

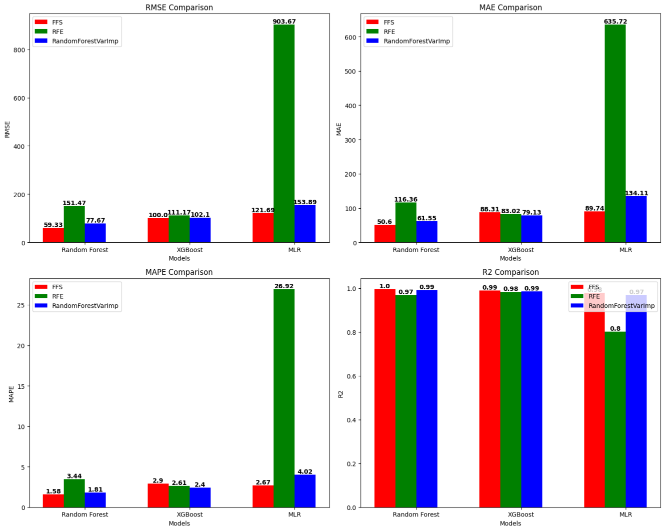

Fig 4. Comparison of RMSE, MAE, MAPE, and R2 for different Feature Selection andthree ML models (Random Forest, XGBoost,and MLR)

The formula for calculating R2 is the sum of squared differences between the actual values and the mean of the actual values. R2 can range from 0 to 1, where a score of 0 indicates that the model does not fit the data, and a score of 1 means that the model perfectly fits the data. A higher R2 score indicates a better fit between the model and the data.

Table 6. RMSE, MAE, and MAPE value of Random Forest with different feature selection algorithms

Feature Selection	RMSE	MAE	MAPE	$\mathbf{R^2}$
FFS	59.33	50.60	1.58	0.9953
RFE	151.47	116.36	3.44	0.969738
RFVariableImp	77.67	61.55	1.81	0.99204

Tables 6, 7 and 8 represents RMSE, MAE, MAPE, and R2 values of the Random Forest algorithm, XGBoost, and MLR with three different feature selection algorithms.

Table 7. RMSE, MAE, and MAPE value of XGBoost algorithm with different feature selection algorithms

Feature Selection	RMSE	MAE	MAPE	R^2
FFS	100.00	88.31	2.90%	0.98924
RFE	111.17	83.02	2.61%	0.9836
RFVariableImp	102.10	79.13	2.40%	0.9862

Table 8. RMSE, MAE, and MAPE value of MLR algorithm with different feature selection algorithms

Feature Selection	RMSE	MAE	MAPE	R^2
FFS	121.69	89.74	2.67%	0.9804
RFE	903.67	635.72	26.92%	0.80194
RFVariableImp	153.89	134.11	4.02%	0.9687

Figure 1 shows the accuracy of RMSE, MAE, MAPE, and R2 values of three different feature selection algorithms for predicting crop yield using the Random Forest Algorithm. Figures 2 and 3 show the accuracy of various performance metrics with XGBoost and MLR algorithms. The RMSE, MAE, MAPE, and R2 of the Random Forest algorithm with FFS are 59.33,50.60, 1.58, and 0.9953, respectively, which are much better than other ML (XGBoost and MLR) algorithms with feature selection (FFS, RFE, RFVARIMP).

The FFS method has a computational time complexity of O(n), where n represents the number of features the algorithm selects. This makes it more efficient compared to other algorithms. Table 9 compares the results of this research with those of the existing literature, emphasizing the novel aspects and advancements of this research in comparison to previous publications.

Table 9. Comparative Analysis of Obtained Results and Originality in Relation to Prior Publications

Algorithm	Accuracy
MLR with SFFS algorithm	85% accuracy.
Hybrid RaNN model (6)	98% accuracy
RFE-SVR (8)	R^2- 0.710, and RMSE-0.645
Random forest method (9)	R-0.933
CNN model (10)	R- 0.81, RMSE-0.64, MAE-0.50
Proposed Random Forest with FFS algorithm	RMSE-59.33 and 99.53% accuracy.

Limitations

This research has some limitations, such as:

1. The assumption of feature independence, which may not always hold true in real-world agricultural systems;
2. It's efficacy depends on the input data's availability and quality.
3. Incomplete or inaccurate data can introduce biases and have a negative impact on the model's predictions.
4. It works best for predicting statistical data, but results may be less precise when applied to image data.

The proposed research can be improved by using other feature selection algorithms like LASSO(least absolute shrinkage and selection attribute) and CFS (correlation-based feature selection) with neural networks, and it can be applied to multiple crops.

4 Conclusion

Five decades of wheat crop data are analyzed to identify the most influential factors in crop yield prediction. Seasonal precipitation from October to January is the most significant factor. It is discovered that performance has minor differences in metrics such as MAE, RMSE, MAPE, and R2 values. However, the Random Forest algorithm, which utilized the Forward Feature Selection technique, attained the maximum level of accuracy, 98.415%. The study provides a reliable subset of characteristics that can accurately predict wheat crop yield in India. This information can be used to construct an India-specific wheat crop yield prediction model. It might help in planning India's agricultural policies and practices.

References

1) Van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*. 2020;177:105709. Available from: https://doi.org/10.1016/j.compag.2020.105709.

2) Yang Q, Shi L, Han J, Zha Y, Zhu P. Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. *Field Crops Research*. 2019;235:142–153. Available from: https://doi.org/10.1016/j.fcr.2019.02.022.
3) Jia W, Sun M, Lian J, Hou S. Feature dimensionality reduction: a review. *Complex & Intelligent Systems*. 2022;8(3):2663–2693. Available from: https://doi.org/10.1007/s40747-021-00637-x.
4) Raja SP, Sawicka B, Stamenkovic Z, Mariammal G. Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers. *IEEE Access*. 2022;10:23625–23641. Available from: https://doi.org/10.1109/ACCESS.2022.3154350.
5) Gopal PSM, Bhargavi R. Selection of Important Features for Optimizing Crop Yield Prediction. *International Journal of Agricultural and Environmental Information Systems*. 2019;10(3):54–71. Available from: https://doi.org/10.4018/IJAEIS.2019070104.
6) Lingwal S, Bhatia KK, Singh M. A novel machine learning approach for rice yield estimation. *Journal of Experimental & Theoretical Artificial Intelligence*. 2022;p. 1–20. Available from: https://doi.org/10.1080/0952813X.2022.2062458.
7) Gopal PSM, Bhargavi R. Optimum Feature Subset for Optimizing Crop Yield Prediction Using Filter and Wrapper Approaches. *Applied Engineering in Agriculture*. 2019;35(1):9–14. Available from: https://doi.org/10.13031/aea.12938.
8) Corrales DC, Schoving C, Raynal H, Debaeke P, Journet EPP, Constantin J. A surrogate model based on feature selection techniques and regression learners to improve soybean yield prediction in southern France. *Computers and Electronics in Agriculture*. 2022;192:106578. Available from: https://doi.org/10.1016/j.compag.2021.106578.
9) Whitmire CD, Vance JM, Rasheed HK, Missaoui A, Rasheed KM, Maier FW. Using Machine Learning and Feature Selection for Alfalfa Yield Prediction. *AI*. 2021;2(1):71–88. Available from: https://doi.org/10.3390/ai2010006.
10) Srivastava AK, Safaei N, Khaki S, Lopez G, Zeng W, Ewert F, et al. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports*. 2022;12(1). Available from: https://doi.org/10.1038/s41598-022-06249-w.
11) Liu Z, Japkowicz N, Wang R, Cai Y, Tang D, Cai X. A statistical pattern based feature extraction method on system call traces for anomaly detection. *Information and Software Technology*. 2020;126:106348. Available from: https://doi.org/10.1016/j.infsof.2020.106348.
12) Pham HT, Awange J, Kuhn M, Van Nguyen B, Bui LK. Enhancing Crop Yield Prediction Utilizing Machine Learning on Satellite-Based Vegetation Health Indices. *Sensors*. 2022;22(3):719. Available from: https://doi.org/10.3390/s22030719.
13) Barbosa BDS, Ferraz GAES, Costa L, Ampatzidis Y, Vijayakumar V, Santos LMD. UAV-based coffee yield prediction utilizing feature selection and deep learning. *Smart Agricultural Technology*. 2021;1:100010. Available from: https://doi.org/10.1016/j.atech.2021.100010.
14) Mehla A, Deora SS. Use of Machine Learning and IoT in Agriculture. *IoT Based Smart Applications*. 2023;p. 277–293. Available from: https://doi.org/10.1007/978-3-031-04524-0_16.
15) Ramos APM, Osco LP, Furuya DEG, Gonçalves WN, Santana DC, Teodoro LPR, et al. A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices. *Computers and Electronics in Agriculture*. 2020;178:105791. Available from: https://doi.org/10.1016/j.compag.2020.105791.
16) India Meteorological Department Ministry Of Earth Sciences Government Of India. 2022. Available from: https://mausam.imd.gov.in/.
17) Directorate of Economics And Statistics,Ministry Of Agriculture,Government Of India. 2022. Available from: https://eands.dacnet.nic.in/.
18) Python. 2023. Available from: https://docs.python.org/3/library/.
19) Suruliandi A, Mariammal G, Raja SP. Crop prediction based on soil and environmental characteristics using feature selection techniques. *Mathematical and Computer Modelling of Dynamical Systems*. 2021;27(1):117–140. Available from: https://doi.org/10.1080/13873954.2021.1882505.
20) Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J. A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends*;1(2):56–70. Available from: https://www.researchgate.net/publication/341413445_A_Comprehensive_Review_of_Dimensionality_Reduction_Techniques_for_Feature_Selection_and_Feature_Extraction.
21) Nayana BM, Kumar KR, Chesneau C. Wheat Yield Prediction in India Using Principal Component Analysis-Multivariate Adaptive Regression Splines (PCA-MARS). *Agri Engineering*. 2022;4(2):461–474. Available from: https://doi.org/10.3390/agriengineering4020030.
22) Aworka R, Cedric LS, Adoni WYH, Zoueu JT, Mutombo FK, Kimpolo CLM, et al. Agricultural decision system based on advanced machine learning models for yield prediction: Case of East African countries. *Smart Agricultural Technology*. 2022;2:100048. Available from: https://doi.org/10.1016/j.atech.2022.100048.
23) Obsie EY, Qu H, Drummond F. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. *Computers and Electronics in Agriculture*. 2020;178:105778. Available from: https://doi.org/10.1016/j.compag.2020.105778.
24) Manivasagam MA, Sumalatha P, Likitha A, Pravallika V, Satish KV, Sreeram S. An Efficient Crop Yield Prediction Using Machine Learning. *International Journal of Research in Engineering*. 2022;5(3). Available from: https://journal.ijresm.com/index.php/ijresm/article/view/1862.
25) Elavarasan D, Vincent PMDR. A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. *Journal of Ambient Intelligence and Humanized Computing*. 2021;12(11):10009–10022. Available from: https://doi.org/10.1007/s12652-020-02752-y.
26) Elavarasan D, Vincent DR. Reinforced XGBoost machine learning model for sustainable intelligent agrarian applications. *Journal of Intelligent & Fuzzy Systems*. 2020;39(5):7605–7620. Available from: https://doi.org/10.3233/JIFS-200862.