RESEARCH PAPER

Machine Learning-Based Crop Yield Prediction in India: A Comparative Evaluation of Individual and Hybrid Ensemble Models

Gitanjali Sanjay Patil[1] and Dr. Hemant Kumar Gupta[2]
Research Scholar, Department of CSE, SAGE University, Indore, MP 452020, India.
E-mail: Gitanjalipatil.in@gmail.com
Associate Professor, Department of CSE, SAGE University, Indore, MP 452020, India.
E-mail: hemugupta3131@gmail.com

Abstract:
In a country like India where the agricultural productivity is increasingly at risk due to climate variability, prediction of crop yield plays a crucial role in enhancing the rural economy and food security. This study creates a machine learning model using 19,689 records from the Open Government Data Platform, FAO, and IMD for various crops, states, seasons and years ranging from 1997 to 2020. A carefully designed feature engineering process resulted in the creation of 114 features: ratios, log transforms, interaction terms, polynomial features, and seasonal-temporal encoding. Six models (Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM, and Support Vector Regression) and four hybrid ensemble models were tested with RMSE, MAE, R^2, and adjusted R^2. All ensemble models performed worse than Random Forest which had the lowest RMSE of 60.99, the lowest MAE of 4.22, and the highest R^2 value of 0.9954. The results demonstrate the importance of careful design of the features and model selection.

Keywords: crop yield prediction, machine learning (ML), feature engineering, Random Forest (RF), ensemble learning

How to cite this article: Gitanjali Sanjay Patil1 and Dr. Hemant Kumar Gupta2. Machine Learning-Based Crop Yield Prediction in India: A Comparative Evaluation of Individual and Hybrid Ensemble Models. Int J Drug Deliv Technol. 2026;16(89s):2140-2148. DOI: 10.25258/ijddt.16.89s.202.

Introduction

Since time immemorial agriculture has been playing a significant role in the Indian economy which has also helped in the socio-economic development of the rural area and food security of the nation. Its production creates jobs and income for many people and also serves a number of related industries such as food processing farming machinery and textiles. Thus increased productivity is crucial to boost rural economies and food security. Climate change and erratic weather patterns are increasingly impacting Indian agriculture. Unusual precipitation, warmer temperatures and the growing occurrence of extreme weather events can disrupt the crop growing season and diminish crop production. Small and marginal farmers who have access to less financial resources, modern technologies and effective adaptation strategies are facing more severe challenges. Farm productivity and sustainability are also impacted by other issues including reduced soil fertility, drought and shifts in pest and disease profiles.

The challenge of predicting crop yields in this context can be met with a promising solution: machine learning, Crop yield prediction using machine learning is a promising solution to address these challenges. Machine learning models can be used to identify patterns in historical agricultural data and weather conditions within the soil that can affect crop production and estimate crop yields. The forecasts can be useful for farmers to make decisions regarding crop selection, sowing timing and resource utilisation. The application of machine learning in precision agriculture can thus play a significant role in optimizing resource use, enhancing agricultural decision-making, and promoting sustainability.

The present study proposes a machine learning model to predict crop production using the agricultural data obtained from three sources i.e. Open Government Data (OGD) Platform of India, Food and Agriculture Organization (FAO) and India Meteorological Department (IMD). The data is consolidated and has a total of 19,689 records across various crop categories, Indian states, cropping seasons, and years ranging from 1997 to 2020. A comprehensive feature engineering pipeline is built to capture the interactions between the agricultural variables, which include features such as ratios, logs, interactions, polynomials, time encoding and state grouping of yields. Following categorical encoding, there are 114 input variables for the model development.

The following machine learning regression algorithms are evaluated: Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM and Support Vector Regression. In addition, four hybrid ensemble approaches are discussed (Ridge-based stacking, weighted voting, advanced stacking with ElasticNet and artificial neural network-based stacking). RMSD, MAE, R^2, and adjusted R^2 are used to evaluate the model performance. The errors in prediction and their capability in generalizing are also assessed using residual analysis, actual versus

predicted yield plots, and learning curves. This comparative evaluation provides an insight about the comparative effectiveness of individual regression algorithms, and ensemble learning strategies, for making the crop yield prediction. Experimental results demonstrate that Random Forest is the best model of the ten models used with an RMSE of 60.99, an MAE of 4.22, an R^2 of 0.9954, and an adjusted R^2 of 0.9952 on the test set. The results show that the stand-alone Random Forest model outperforms the ensemble models used in this work for this dataset. It is therefore imperative that feature engineering is carefully designed, model selection is judicious, and rigorous evaluation of model performance is carried out in prediction of agricultural yield. The proposed framework can be extended and investigated to estimate crop yields for various crops, states and crop seasons in India by using data.

Literature Survey

Fares and colleagues examined the use of machine learning and deep learning methods—such as random forest (RF), gradient boosting machines (GBM), artificial neural networks (ANN), convolutional neural networks (CNN), and long short-term memory (LSTM) networks for forecasting crop performance under water stress caused by drought. Rather than relying on a single standardized dataset, their review draws together evidence from multiple previously published studies that used weather, soil, crop, irrigation, and remote-sensing data. They note that RF and ANN are among the most frequently adopted techniques and emphasize persistent issues related to data availability, spatial variability, and model generalization [1]. Kshatri et al. surveyed machine learning and deep learning approaches to crop-yield prediction and input optimization, including decision trees, random forests, support vector machines, k-nearest neighbors, CNNs, and LSTMs. They report that recent deep-learning and hybrid approaches often achieve above 95%; however, no common evaluation metric or benchmark dataset is specified in the accessible abstract. The review identifies multimodal integration and real-time deployment as remaining challenges [2]. Nirmaladevi et al. developed a hybrid machine learning framework combining Gradient Boosting Regression, Neural Networks, averaging, stacking, and feature augmentation for predicting total agricultural yield. The study utilized 57 years of Indian agricultural data containing 14 variables collected from RBI and the Ministry of Agriculture. The stacking ensemble achieved the best performance, with an R^2 of 0.9729, RMSE of 0.9890, and MAE of 0.7553 [3]. Sahu et al. developed a Random Forest regression framework using climate reanalysis and satellite-derived vegetation data to predict wheat and barley yields under drought conditions. The study examined agricultural data from Bundelkhand, India, and Germany, covering 1990–2021. Hydrologically balanced training improved prediction performance in several experiments, particularly in Bundelkhand, while NDVI emerged as the most influential predictor across regions and spatial scales [4].

Wang Y. et al. [6] provided a comprehensive overview of the recent advancements in crop yield prediction using deep learning algorithms and compared them with conventional machine learning approaches. The study covered various prediction models and their possible contribution to more accurate estimation of the agricultural yield. It also emphasised the need for crop yield prediction for providing support to farmers' decisions and agricultural policy making. In addition, the authors highlighted challenges in existing systems and recommended future research to improve the efficacy of the crop yield prediction models based on deep learning. Nigam A. et al. [7] wrote about using machine learning in the precision agriculture to predict the yield of crops. They noted that historical data from agricultural records, real-time readings from sensors and satellite imagery are crucial sources of information. Regression models and deep learning techniques can be used to detect complex relationships between environmental and agronomic variables and estimate crops yields. The incorporation of sensor and satellite-based information can also help in the management of resources, especially the use of water and fertilizers. But there still remain key challenges with data quality and model interpretability. This study highlighted the role of machine learning in mitigating climate variability and in ensuring food security. Savaliya et al. [8] suggested crop prediction method based on Ensemble of K-Means clustering, Random Forest, Linear regression and Support vector Machine (SVM). They used a majority-voting strategy to fuse the predictions of the various algorithms, for which their model got about 99.77% accuracy. The method used took into consideration the environment and the place data to suggest appropriate crops. The results indicated that ensemble learning can enhance the prediction accuracy and aid in agricultural decision-making. The study also suggested that the potential for synergies of multiple algorithms helps to increase the productivity of crops and the benefits of farmers. Sahoo et al. 2024 employed different machine learning algorithms, such as Cubist, Random Forest (RF), Gradient Boosting Machine (GBM), Multivariate Adaptive Regression Splines (MARS), and Extreme Gradient Boosting (XGBoost) to predict the yield of rice crops in Eastern India under changing climatic parameters [9]. Climatic and soil related factors were included in the study and the models were tested with 1889 field observations by cross validation. The results revealed that Cubist and Random Forest algorithms performed better than the

other algorithms that were tested. The results indicate the necessity to choose appropriate machine learning models for different environmental conditions to estimate the yield of rice crops. To evaluate the state of the art of machine learning and deep learning approaches in crop yield prediction, a systematic literature review of machine learning and deep learning approaches for crop yield prediction was carried out by Jabad and Murad (2024). They reviewed algorithms like Random Forest (RF), Artificial Neural Networks (ANN), Support Vector Machine (SVM), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks and Deep Neural Networks (DNN) [10]. The authors determined precipitation, soil properties, temperature, and vegetation indices (such as NDVI, EVI, and LAI) as essential input variables when predicting crop yield. The study highlighted the importance of choosing the right environmental features and the right learning algorithms to build up reliable prediction systems. In accordance with the fact that adoption of CSA practices depends on various factors, Noma and Babu (2024) created a machine learning model using the optimized Gradient Boosting algorithm to predict the adoption of CSA practices [11].

Data from the Rakai district, Uganda, were collected for this study from the Ugandan national, longitudinal household survey. The input features were socio-economic indicators, agroecological conditions and climatic variables. Model results showed an accuracy of about 60%, suggesting the potential of the model to capture patterns in the adoption of climate smart agricultural practices. The results indicate that machine learning can be used to aid the understanding of how factors affect agricultural decisions, but there is potential for better results if prediction performance is further improved. Talaat F. In 2023, (2023) proposed a Crop Yield Prediction Algorithm (CYPA) framework which integrates machine learning techniques with Internet of Things (IoT) technology [12].

The study was performed on five regression models, namely Decision Tree, Random Forest and Extra Trees regressors, with the hyperparameter optimization process. The performance of the three models mentioned is reported as 0.9814, 0.9933 and 0.9903 respectively. The framework also included features of active learning for the enhancement of computational efficiency and predictive performance. The results show that combining data collected using IoT with optimized machine learning models is a viable approach for crop yield prediction and monitoring. In 2022, Srivastava et al. developed a 1D-CNN model to forecast winter wheat yields in 271 counties in Germany from 1999 to 2019 [13].

The model inputs were weather, soil and crop phenological data. It was benchmarked against eight baseline models including Random Forest, XGBoost and Deep Neural Networks. The proposed model showed 7–14% less RMS error and 4–50% higher correlation than the baseline models. The SHAP analysis showed that some variables like DUL, wind speed during week 10 and solar radiation during week 7 are significant predictors. The study found that prediction errors varied across regions and prediction results were not very interpretable despite its high predictive ability. A Systematic Literature Review (SLR) was performed on the use of machine learning for crop yield prediction by Van Klompenburg et al. (2020) [14]. The authors searched six databases to find 567 studies, which were then screened based on the inclusion and exclusion criteria to yield 50 studies. They found that the input variables temperature, rainfall and soil type were important and pointed out the application of algorithms like Artificial Neural Networks. A further search revealed 30 studies related to deep learning techniques such as CNN, LSTM, and DNN architectures. The review offered a general overview of the prediction methods and input features and research trends in crop yield estimation. Pantazi et al. [15] explored the use of machine learning techniques and the latest remote sensing technologies for wheat yield prediction. They examined how a wheat production prediction can be generated by using a predictive algorithm and spatial sensing information. The results showed that the fusion of remote sensing information and machine learning techniques has the potential to enhance wheat yield estimation and optimization of resource management. The use of a combination of in-field information and observations from remotely sensed data for precision agriculture is emphasized.

Holzman et al. [16] studied crop productivity estimation using the remotely sensed water stress and solar radiation data prior to the crop establishment. The study demonstrated that the environmental conditions which are monitored with remote sensing can influence the crop performance before harvest. This data can be used to make a prediction of productivity and to understand potential agricultural risks. The results show that remote sensing-based indicators can be valuable inputs to crop yield prediction and agricultural management. Singh et al. [17] summarized the application of machine learning to high throughput plant stress phenotyping. The study reviewed novel phenotyping technologies to identify and analyze responses of crops to stress. Machine learning methods can analyze vast amounts of plant data and assist in pattern recognition that is indicative of environmental stresses. This data can be used to inform approaches to enhance the resilience and productivity of crops. The researchers pointed out that advanced data analysis combined with plant phenotyping could be a powerful tool for better agricultural monitoring. Nonlinear parametric modelling was used to investigate soil properties' effect on crop yield and vegetation indicators in

another study [18] that investigated the effect of soil properties on yield and vegetation indicators. This method explored the connections between soil properties and crop results. The study of these relationships is significant in the context of determining the interactions between the environment and productivity in agriculture. The results further indicate that the soil properties and vegetation data can be used to describe complex interactions between soil physical, chemical, and biological properties and crop production, using nonlinear modelling techniques.

Kuradusenge et al. [19] proposed an IoT-based system with machine learning algorithms for crop yield prediction in sub-Saharan Africa, called SMART-CYPS. The system was validated for potato and maize production based on weather and soil data in the Musanze District in Rwanda. Mean Absolute Percentage Error (MAPE) ranged from 0.177 to 0.339. The results showed the potential of combining sensor-based environmental data with machine learning for agricultural forecasting. The authors, however, suggested that more crops and areas should be included in the system to enhance its usability and its value to food security. Khanagoudar et al. [20] have done a systematic review of the machine learning approaches applied to the prediction of crop yields in relation to climatic factors. They studied various regression algorithms, such as Decision Trees, K-Nearest Neighbours, Random Forest and Linear Regression. The review focused on the effect of meteorological factors and plant factors on yield estimation. The authors highlighted the crucial point of selecting relevant predictive variables and the right algorithms in order to have a good forecasting system. The study also gave a brief overview of various machine learning methods for agricultural prediction problems.

Bloh, Lobell and Asseng [21] proposed a machine learning approach to wheat yield prediction, combining DSSAT crop simulations with neural networks and Random Forest models. They create synthetic training data with crop and climate models to assist with learning, for their hybrid approach. The proposed method resulted in around 8% reduction in prediction error over traditional methods. The results showed the advantages of the integration of process-based crop models and machine learning techniques. This could provide better yield estimation and inform agriculture practices with greater climate change resilience.

Proposed Methodology: Dataset and Feature Engineering

Dataset Description

In this study an agricultural dataset has been prepared by combining the data from three agricultural available sources namely Open Government Data (OGD) Platform of India, Food and Agriculture Organization (FAO) of the United Nations and India Meteorological Department (IMD). The integration and cleaning of the data were carried out in Google Colab. The final data set has 19,689 observations and 10 variables.

The records are for cereals, pulses, oilseeds, fibre crops, spices and horticultural crops in different agro-climatic zones of various Indian states during 1997-2020. They represent a single crop in a particular state, season and year.The variables included are Crop, Crop_Year, Season, State, Area, Production, Annual_Rainfall, Fertilizer, Pesticide and Yield. The target variable to predict is yield and the rest are input variables. The data are complete but the numeric variables are highly skewed (e.g. the median area is ~9,317 vs. the mean area of ~179,927). This skewness gave rise to the following transformations.

Data Pre-processing

Names of columns and text entries were trimmed to accommodate for extra space. Any repeats of spaces in Season labels were combined in order to not split a season into two categories. The pipeline included a procedure to fill missing values (median for numeric columns and the most frequent value for text columns) but it is not needed for this dataset.

Feature Engineering

Features were created in two stages expanding the table from 10 to 28 columns.

- **Ratio features:** Productivity (Production / Area), Fertilizer_Efficiency (Production / Fertilizer), Rainfall_per_Area and Pesticide_per_Area. A tiny constant (10^-10) was added to each denominator to avoid division by zero.
- **Log transformation:** log (1 + x) was applied to Area, Production, Fertilizer and Pesticide to reduce skew.
- **Interaction and polynomial terms:** Rainfall × Area and Fertilizer × Pesticide capture joint effects, while squared and cubed rainfall allow a non-linear response.
- **Seasonal and temporal encoding:** Seasons were mapped to integer codes and converted to sine and cosine values. The crop year was scaled to the range 0 to 1 and encoded the same way (Year_Sin, Year_Cos).
- **State yield cluster:** States were placed in four quartile-based groups (Low, Medium, High, Very High) according to their average yield.

Then 114 input features were obtained by one-hot encoding categorical columns with the first one removed. A fixed seed of 42 was used to train-test the data in an 80/20 ratio (15,751/3,938). The features were standardized using a scaler fitted to the training set, but none of the test-set information was used in the training.

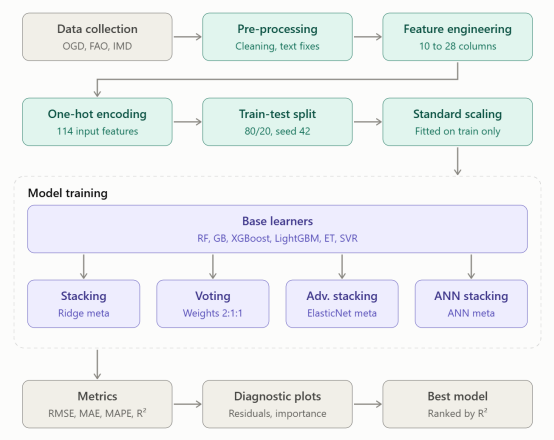


Figure 1: Architecture of Proposed Methodology

This proposed methodology is comprised of four stages: Data Preparation, Feature Engineering, Model Training, and Performance Evaluation as shown in Figure 1.

In phase 1 data from the OGD Platform of India, FAO and IMD were merged to create a dataset with 10 original variables and 19,689 records. The data cleaning consisted of the removal of extra spaces, the uniformization of the seasons and the establishment of a strategy for treating missing values.

For Phase 2: Feature Engineering, new features were created by creating ratios, log transformations, interaction terms, polynomial features, encoding with seasons, encoding with time, and grouping states by yield. Following one-hot encoding, the data set had 114 input features. Data has been split into a training set (80%) and a test set (20%) with a random seed of 42. The standardized parameters were trained on the training set and tested on the test set in order to prevent any information leakage.

In Phase 3: 6 regression algorithms were trained to predict crop yield, they are Random Forest (RF), Gradient Boosting (GB), XGBoost, LightGBM, Extra Trees (ET) and Support Vector Regression (SVR). Four variants of ensemble methods were also created: stacking with Ridge regression, weighted voting with RF, GB, ET (weights 2:1:1), advanced stacking with all six models with ElasticNet and ANN-based stacking with out-of-fold predictions from RF, GB and ET.

To evaluate the model performance RMSE, MAE, MAPE, MSE, R^2 and adjusted R^2 were used in this phase. To supplement the evaluation residual plots, feature importance and learning curves were explored. The models were evaluated by test-set performance and the best performing model is decided by R^2 as primary criteria.

Flow of Algorithm

Algorithm 1: Hybrid ensemble framework for crop yield prediction

Input: Merged dataset D (19,689 records); test fraction $\tau = 0.2$; seed $s = 42$

Output: Best model M*

1: Clean text and column names; fill missing values (median for numeric, mode for categorical)
2: Engineer features:
(a) ratios: Production/Area, Production/Fertilizer, Rainfall/Area, Pesticide/Area
(b) ln(1 + x) of Area, Production, Fertilizer, Pesticide
(c) Interaction and polynomial terms of rainfall, area, fertilizer and pesticide
(d) Sine/cosine encoding of Season and Year
(e) State yield cluster (four quartile groups)
3: One-hot encode categorical columns
4: Split data into (X_train, y_train) and (X_test, y_test) using τ and s
5: Standardise features, fitting the scaler on X_train only
6: for each B in {RF, GB, XGBoost, LightGBM, ET, SVR} do
7: Train B on (X_train, y_train)
8: end for
9: Build hybrid models from the base learners:
H1: stacking of RF, GB, ET with a Ridge meta-learner
H2: weighted voting of RF, GB, ET with weights 2:1:1
H3: stacking of all six learners with an ElasticNet meta-learner
H4: neural network meta-learner trained on out-of-fold predictions of RF, GB, ET
10: for each model M_k do
11: Predict on X_test; compute RMSE, MAE, MAPE and R^2
12: end for
13: M* ← model with the highest R^2
14: return M*

Base Learners, Meta-Learners and Hybrid Models

The base learners used were six regression algorithms: Random Forest (RF), Extra Trees (ET), Gradient Boosting (GB), XGBoost, LightGBM, and Support Vector Regression (SVR). RF and ET fuse the predictions of several decision trees for enhancing model stability. GB, XGBoost and LightGBM train a tree after another, using previous trees' errors as the basis for minimizing errors. SVR is a radial basis function (RBF) kernel based model with the aim to control deviations within a specified tolerance, taking into account nonlinear relationships. The different model set-up are stated briefly in Table 1.

To enhance the performance of predictions, four ensembles were developed. For stacking, the best known cross-validation method is five-fold cross validation which makes out-of-fold predictions of the training samples. The predictions from these models are then fed into a meta-learner; the base

models are then refitted to the full training set. When it is time for testing, the meta-learner is used to merge the predictions from the base models. The meta-learners were ridge regression, ElasticNet and an artificial neural network (ANN). Weighted voting is a weighted average of the predictions of RF, GB and ET with weights 2, 1, 1, respectively and without the need of a dedicated meta-learner.

Table 1. Models and their hyperparameters

Model	Role	Hyperparameters
RF	Base learner	200 trees, max depth 20, min split 5, min leaf 2
ET	Base learner	200 trees, max depth 20, min split 5
GB	Base learner	200 trees, learning rate 0.05, max depth 5, min split 5
XGBoost	Base learner	200 trees, learning rate 0.05, max depth 6, row and column sampling 0.8
LightGBM	Base learner	200 trees, learning rate 0.05, max depth 6, 31 leaves
SVR	Base learner	RBF kernel, C = 10, epsilon = 0.1, gamma = scale
Ridge	Meta-learner (stacking)	alpha = 1.0; inputs from RF, GB, ET; 5-fold
Weighted voting	Fixed combiner	Weights 2:1:1 for RF, GB, ET
ElasticNet	Meta-learner (advanced stacking)	alpha = 0.1, L1 ratio = 0.5; inputs from RF, GB, ET, SVR, XGBoost, LightGBM; 5-fold
ANN	Meta-learner (ANN stacking)	Layers 256-128-64-32 (ReLU), batch normalisation, dropout 0.3, Adam (learning rate 0.001), MSE loss, batch size 32, up to 100 epochs, early stopping (patience 20)

The tree based models were re-built with their own parameters in the hybrid models. According to the stacking model, 100 trees were used, maximizing the depth to 15 for Random Forest (RF) and Extra Trees (ET). For the voting model it was 150 trees and a depth of 15. The advanced stacking model employed was with 200 trees and learning rate of 0.03 for the boosting models and C (cost parameter) of 5 for the Support Vector Regressor (SVR). The models trained in Table 1 are reused in the ANN stack.

Results and Discussion

All models were trained on 15,751 records and tested on the remaining 3,938 records. Performance is reported with RMSE, MAE, R^2 and adjusted R^2. MAPE was also computed but is left out of the tables because many records have a yield of zero or close to zero which makes the percentage error blow up (values above 10^13 %). The yield values are also highly skewed (median about 1.03, mean about 80, maximum 21,105). As a result, RMSE is driven by a few very large values, while MAE shows the typical error.

Table 2 ranks the ten models by R^2. Random Forest (RF) performed best on every metric, with an RMSE of 60.99, an MAE of 4.22 and an R^2 of 0.9954. Advanced stacking, weighted voting and Gradient Boosting (GB) followed closely and six of the ten models reached an R^2 above 0.989. For all eight models with R^2 above 0.98, adjusted R^2 stays within 0.0005 of R^2, so the large number of input features (114) does not inflate the scores. SVR and the ANN-based stack performed poorly.

Table 2. Test-set performance of all models, ranked by R^2

Rank	Model	Type	RMSE	MAE	R^2	Adjusted R^2
1	Random Forest	Base	60.99	4.22	0.9954	0.9952
2	Advanced stacking	Hybrid	65.17	6.66	0.9947	0.9945
3	Weighted voting	Hybrid	68.48	4.53	0.9941	0.9940
4	Gradient Boosting	Base	69.41	4.48	0.9940	0.9938
5	Extra Trees	Base	84.31	4.62	0.9911	0.9909
6	Stacking (Ridge)	Hybrid	93.74	5.39	0.9890	0.9887
7	Light GBM	Base	118.70	9.17	0.9824	0.9819
8	XGBoost	Base	120.10	7.69	0.9820	0.9815

Figure 1 compares the performance of ten regression and ensemble models using RMSE, MAE, MAPE, and R^2. Random Forest achieved the best overall performance, with an RMSE of 60.988, an MAE of 4.216, and an R^2 of approximately 0.995, indicating strong predictive accuracy. Advanced stacking, weighted voting, and Gradient Boosting also performed well, with R^2 values above 0.994. In contrast, SVR showed poor performance, while ANN-based stacking produced a high prediction error and a lower R^2 value. The MAPE values are extremely large for several models because zero or near-zero yield values make percentage errors unstable. Therefore, RMSE, MAE, and R^2 provide more useful comparisons for this dataset.

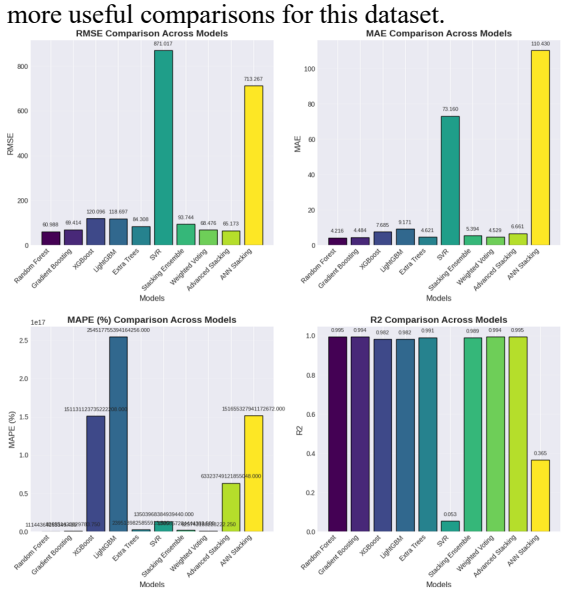


Figure 1: Comparison of RMSE, MAE, MAPE and R^2 across all models

Comparative analysis

Among the individual models, the best results were obtained by the two bagging based models (Random Forest (RF) and Extra Trees (ET) and Gradient Boosting (GB)) closely followed by the two boosting based models (Boosted XGB (XGB) and Boosted Decision Trees (BDT)). For comparison, the RF result had an RMSE of around 49% lower than the RMSE of the other methods, XGBoost or LightGBM. In terms of the solution, RF's prediction error was about half the prediction error made by these two boosting models. For instance, in the case of XGBoost and LightGBM errors around 120, the error for RF would be approximately 61 which is half of the error for XGBoost and LightGBM. The large gap is probably due to the fact that XGBoost and LightGBM were not extensively tuned (close to their default setting of around 120), while RF was able to be tuned to perform well without much tuning. The performance of SVR was poor because the algorithm is sensitive to the scale of the target variable, and the yield values of this dataset were quite large ranging from 0 to more than 21,000, which was difficult for this type of input.

Hybrid Models: Table 3 shows a comparison between the hybrid models and the standalone RF model which is used as the benchmark (RMSE = 60.99). The Change vs RF column shows the difference between each hybrid and this reference as a percentage with positive percentages indicating that the hybrid was worse than the RF. All four hybrid approaches performed worse than RF with all values shown in the column being negative. Advanced stacking was the next best to RF with an RMSE 6.9% higher than RF (65.17 vs. 60.99) but still the second lowest RMSE of all the models. However, its MAE (6.66) was higher than RF, GB, ET, and weighted voting meaning that it had less number of very large errors but a little larger number of typical (average) errors. Weighted voting did not even outperform GB by more than 1.3% in terms of RMSE. Ridge stacking was significantly less effective than GB, ET and RF alone (53.7% worse error rates than RF).

Table 3. Hybrid models compared with the standalone Random Forest

Hybrid model	**Combines**	**RMSE**	**Change vs RF**
Advanced stacking (ElasticNet)	RF, GB, ET, SVR, XGBoost, LightGBM	65.17	+6.9%
Weighted voting	RF, GB, ET	68.48	+12.3%
Stacking (Ridge)	RF, GB, ET	93.74	+53.7%
ANN stacking	RF, GB, ET	713.27	+1,070%

Discussion

The main reason for RF having such good performance is some of its properties: many deep decision trees are aggregated which means it has a low variance; it is resilient to skewed input distributions; it is not required to scale the target variable. The gains in the hybrid models, on the other hand, were negligible. One possible explanation is that the error made by the various members of the ensemble are likely to be correlated because they are likely to see the same feature set as targets and are likely to make similar errors; hence, the combiner has little independent information to utilize. The meta learner has to rely on out-of-fold predictions, which are the least reliable for rare, high-yield observations, in addition.

One of the highlights from all of the top-performing models is that there is a significant difference between the RMSE and the MAE. The MAE is approximately 1/14 of the RMSE for RF suggesting that most of the predictions have relatively small

errors with some extreme-yield values having relatively large errors. The best-performing model is shown in Figure 3 its prediction and error distribution and Figure 4 is the learning curve of the model.

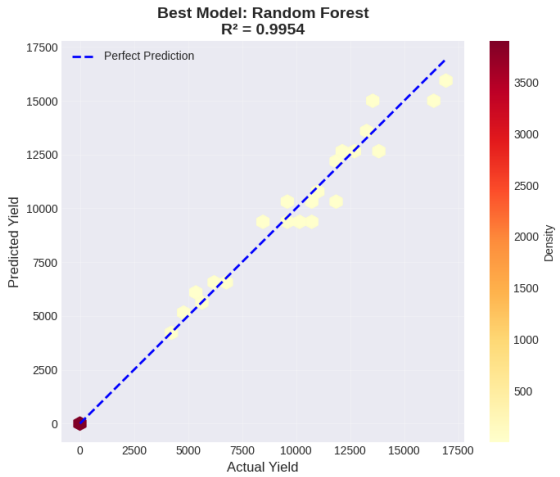


Figure 3: Actual versus predicted yield and error distribution of the best model

The observed and predicted yield values for the Random Forest model are presented and compared in figure 3. Which achieved R^2 of 0.9954. Most of the points lie close to the diagonal line, indicating a good agreement between the observed and predicted values. A pattern like this means that the model has high predictive accuracy for the entire range of yields. But a few observations are slightly off the mark from the reference line, suggesting there are small prediction discrepancies. All these results together show that the Random Forest model is reliable for yield prediction.

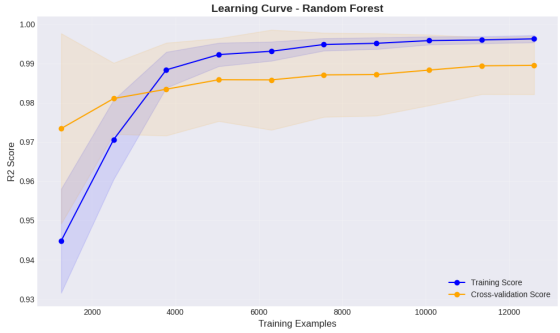


Figure 4: Learning curve of the Random Forest model.

The learning curve of the Random Forest model is shown in figure 4 and created in the same way as explained above for the training with training and cross validation. R^2 scores. The training score continues to improve from about 0.945 to 0.996 as the number of examples used to train the model increases; the cross-validation score increases from 0.973 to 0.990. As the number of training examples increases, the two curves get closer together, corresponding to improved generalization ability and reduced overfitting. The consistently high validation score highlights the model's strong predictive capabilities and consistency, making it suitable for yield prediction applications.

The reviewed literature shows that machine learning, deep learning, remote sensing, and the technique of IoT have the potential to predict agricultural issues. Their approaches vary, however, in the datasets and data variables used in the model, as well as their prediction goals and the methods used to evaluate the model [1]–[16]. The proposed methodology aims at predicting the crop yield with an integrated database of 19,689 records from Open Government Data Platform, India, FAO and IMD. This approach is different from the specific crop specific and area based studies like the prediction of rice yield in Eastern India [4] and wheat yield estimation [8, 10, 16] as it accounts for multiple crop categories, Indian states, seasons and years from 1997 to 2020. It also accounts for complex relationships in agricultural data through feature engineering (such as log transformations, ratios, interaction terms, polynomial features, and seasonal and temporal encoding).

Conclusion

The present study presented a machine learning model for crop yield prediction using an integrated dataset of a total of 19,689 records from OGD India, FAO and IMD, which included various crops, states, seasons and years. A wide range of input variables were created through a feature engineering pipeline, reflecting the complex interactions between agriculture.A thorough feature engineering pipeline resulted in 114 input variables, reflecting complex interactions within agriculture. The following methods were tested: the four hybrid ensemble models and Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM, and Support Vector Regression. Random Forest showed the highest R2 value of 0.9954 and RMSE value of 60.99, so it has the best performance in the results of all hybrid ensembles. The results show that the features and the proper use of the appropriate models is more important than ensemble complexity for this dataset. The framework is extensible to other crops, places and times for data-based decision making in agriculture.

References

1. Fares A., Veettil A. V., Rahman A., & Awal R. (2026). A review of artificial intelligence applications for predicting crop performance and enhancing food security under drought-induced water stress. *Agricultural Water Management, 325*, Article 110212.
2. Kshatri A. S., & Matlani P. (2026). Using machine learning technology crop yield prediction and optimization: A survey. In M. S. Kaiser, J. Xie, & A. Joshi (Eds.) Intelligent strategies for ICT: ICTCS 2025

(Lecture Notes in Networks and Systems, Vol. 1893, pp. 146–156). Springer.

3. Nirmaladevi S., & Jagatheswari S. (2026). A data-driven hybrid machine learning approach for total agricultural yield prediction with integrated resource efficiency analysis. Journal of Hazardous Materials Advances*, 22*, Article 101242.
4. Sahu H., Garg P. K., Vijay S. & Dasgupta A. (2026). Predictive drivers and transferability of multi-scale machine learning based crop yield prediction under drought across European and Asian climates. Agricultural Water Management*, 327*, Article 110254.
5. Subbulakshmi S., Vasudev P. C., Nair A. S., Deyal A. A., Arjun G. (2026). Crop and yield prediction using data analysis and machine learning algorithms. Computing and machine learning: CML 2025 Lecture Notes in Networks and Systems, Vol. 1614, pp. 179–194. Springer.
6. Wang Y., Zhang Q., Yu, F. Zhang, N., Zhang, X. (2024). Progress in Research on Deep Learning-Based Crop Yield Prediction. Agronomy, 14(10), 2264.
7. Nigam A., Garg S., Agrawal A., & Agrawal P. (2019, November) Crop yield prediction using machine learning algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 125-130). IEEE.
8. Savaliya L., Sapovadiya M., Garg D., Patel P., & Shah M. (2024, October) Agricultural Analysis of Machine Learning Algorithms for Crop Prediction. In 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (pp. 1196-1203). IEEE.
9. Sahoo S., Singha C., & Govind A. (2024). Advanced prediction of rice yield gaps under climate uncertainty using machine learning techniques in Eastern India. *Journal of Agriculture and Food Research, 18*, 101424
10. Jabeed M. A. and Murad M. A. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon, 10*, e04836.
11. Noma F. and Babu S. (2024). Predicting climate smart agriculture (CSA) practices using machine learning: A prime exploratory survey. *Climate Services, 34*, 100484
12. Talaat F. M. (2023). Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. *Neural Computing and Applications*, 35, 17281–17292
13. Srivastava A. K., Safaei N., Khaki S., Lopez G., Zeng W., Ewert F., Gaiser T., & Rahimi, J. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports, 12*(1), 32315.
14. van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture, 177*, 105709.
15. X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, ‘‘Wheat yield prediction using machine learning and advanced sensing techniques,’’ Comput. Electron. Agricult., vol. 121, pp. 57–65, Feb. 2016.
16. M. E. Holzman, F. Carmona, R. Rivas, and R. Niclòs, ‘‘Early assessment of crop yield from remotely sensed water stress and solar radiation data,’’ ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 297–308, Nov. 2018.
17. A. Singh, B. Ganapathysubramanian, A. K. Singh, and S. Sarkar, ‘‘Machine learning for high-throughput stress phenotyping in plants,’’ Trends Plant Sci., vol. 21, no. 2, pp. 110–124, Feb. 2016.
18. R. Whetton, Y. Zhao, S. Shaddad, and A. M. Mouazen, "Non-linear parametric modelling to study how soil properties affect crop yields and NDVI," Comput. Electron. Agricult., vol. 138, pp. 127–136, Jun. 2017
19. Kuradusenge, M., Hitimana, E., Mtonga, K., Gatera, A., Habiyaremye, J., Ngabonziza, J., Hanyurwimfura, D., Rukundo, P., & Mukasine, A. (2024). SMART-CYPS: An intelligent internet of things and machine learning powered crop yield prediction system for food security. *Discover Internet of Things*.
20. Khanagoudar, P. S., Sushma, B. S., Chandana, C. N., Manikanta, N., & Awatimath, S. S. (2023). Crop yield prediction using machine learning algorithm based on climate variables. *International Journal of Cheminformatics, 1*(2).
21. Von Bloh M., Lobell, D., & Asseng, S. (2024). Knowledge informed hybrid machine learning in agricultural yield prediction. *Computers and Electronics in Agriculture, 227*, 109606.