The Bioscan
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 21(1) S (1) 585-600, 2026 www.thebioscan.com

EFFECTIVE PREDICTION USING ENSEMBLE MACHINE LEARNING TECHNIQUES FOR THE TOMATO CROP RECOMMENDATION SYSTEM

A.SAI SRI, Dr. D. MURUGAN

1. Research scholar (Reg no: 23114012282031), Department of Computer Science and Engineering

Manonmaniam Sundaranar University, Tirunelveli-12, Tamil Nadu, India Email: ssaisri1993@gmail.com

2*. Professor, Department of Computer Science and Engineering, Manonmaniam Sundaranar University

Tirunelveli-12, Tamil Nadu, India

DOI: https://doi.org/10.63001/tbs.2026.v21.i01.S.I(1).pp585-600

KEYWORDS

Agricultural production,
Tomato Crop,
Multi-layer
Perception Random
Forest Regression (MLPRFR),
Mean Absolute Error (MAE),
Mean Squared Error (MSE),
Root Mean Square Error (RMSE)

Received on:

08-12-2025

Accepted on:

19-01-2026

Published on:

28-02-2026

Abstract

Agriculture is the fundamental industry responsible for food production and supplying the necessary raw materials for other industrial activities. Agricultural production growth isn't keeping pace with population expansion, which might lead to a global food shortage. Thus, developing nations with limited resources and territories must produce more food. Choosing a regionally appropriate agricultural product boosts productivity. Past data on environmental conditions, cultivation areas, and tomato crop output quantities are needed to anticipate agricultural production in a region. The data used to make these forecasts is private. Since India is a growing country with an agrarian economy, this dissertation focuses on it. We start by acquiring and preparing relevant Indian Agriculture and Welfare Department data. Subsequently, we introduce a sophisticated ensemble machine learning approach known as Multi-layer Perception Random Forest Regression (MLPRFR) to precisely forecast the yield of the primary crops, namely wheat, tomato, banana, and rice.

Following a comprehensive examination of three existing machine learning algorithms, Decision Tree (DT), Gradient Boosting (GB), and Gaussian Naïve Bayes (GNB). Four common assessment measures are used to assess the effectiveness of the suggested MLPRFR compared with existing machine learning models: mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and R².

The actual results unequivocally show that MLPRFR performs better. The model achieves a very low Mean Squared Error (MSE) score of 0.9412, indicating highly accurate and precise crop production estimates. Moreover, MLPRFR's reliability is further supported by the fact that its Mean Absolute Error (MAE) of 0.9874 indicates that it consistently obtains errors below the crucial threshold of 1. The high R² score of 0.94152 and RMSE of 0.9416 highlight the MLPRFR's explanatory power and clarity by indicating that it can explain 94.15% of the variability in the data.

1. Introduction:

In terms of both biology and culture, India is quite diverse. Nearly every Indian earns a living via farming or a closely connected profession. Agriculture greatly benefits from the Internet of Things (IoT). Farmers face a variety of issues, and IoT helps them to concentrate on other relevant occupations [1]. Many people are worried about the foreseeable future of agriculture because of the diversity of crops, but also because of the high standards of crops, profitability, and yield projection [2]. There have been few innovations as groundbreaking as precision farming. In addition to providing weather data and agricultural advice, it is doing everything in its power to prevent sickness and find infections early, preserving the field while selling his wares. Things like spraying plants and insecticides need accurate scalability since their actions are controlled by sensors and actuators. He has a chance to boost his income by using all of these tactics. Take note of his area of competence in particular [3]. Perhaps the landowner foreclosed on my property because the previous owners were unprepared for such a tough decision. The region of interest (ROI) could drop if crops aren't picked correctly [4]. Making ends meet becomes difficult if you rely on this money.

In addition to being publicly available, accurate and up-to-date information is also quite easy to come by. From my observations of case studies involving developing nations, I have concluded that academics are discouraged from pursuing such projects due to an absence of up-to-date data [5][6]. Based on the available resources, we have developed a system. To give a way out of this jam that would provide light on crop sustainability in the long run, machine learning algorithms are programmed to take monetary, social, and environmental considerations into account when making recommendations.

The agricultural industry suffers in the long run because this worry demotivates farmers. One solution to this problem is to help farmers plan their crops so that they can accurately anticipate their yields for the next year. For farms to make sensible decisions, the capacity to precisely forecast crop yields is now crucial [7]. To determine how many crops may be cultivated in a given area, several factors are considered, including soil type, weather, and crop management procedures. To reduce food shortages and agricultural-based import-export strategies, accurate crop forecasting is essential for both the short- and long-term planning of government programs [8].

For a developing country whose economy depends on it, agriculture serves as the main economic sector due to the ideal agricultural environment and an abundance of land. Production of food crops and raw resources is linked to agriculture. One important aspect of a nation's growth cycle is its ability to produce food, even in harsh environments and with little land [9]. Soil fertility has reportedly been decreasing over the years, which has impacted agricultural yields [10]. An agriculturally-based country undergoing transition is the focus of this article's case study. The agricultural land in rising countries is decreasing at a rate of 1% per year, even as their populations are expanding at a rate of 1.2% per year [11]. Furthermore, farmers are not paid a fair price for their goods since they do not have a clear picture of the expected harvest. This issue has a detrimental effect on the agricultural economy in the long run since it demotivates farmers.

Users may be able to make better crop selections with the aid of a proposed technique. All users' private information is stored in a system that requires their membership. The registration process for farmers is complete. With the use of historical crop data, one of the method's modules may recommend a crop to plant based on current conditions. An artificial neural network helps us complete the process. Finally, the developer has included a feedback mechanism for the farmer to help address any operational issues with the system promptly.

This study evaluates the accuracy of several agricultural algorithms and determines that they achieve a combined accuracy of 88%. The results have shown that Multiple Linear Regression achieved a reliability of 90–95% in estimating rice production. The ID3 algorithm was used to analyze and provide recommendations on the decision-making process for the soybean crop. The Support Vector Machines algorithm achieved a ranking of third place. Accurate prediction of all crops was achieved with minimal use of computer resources. The neural network we used, trained on wheat data, attained an accuracy rate of 95%. The data was obtained from Kaggle.com, as per the article *"Use of Data Mining in Crop Yield Prediction"*. The author's software, WEKA, was used to analyze the data.

Using sensitivity, specificity, accuracy, and root-mean-square error, we evaluated precision. A confusion matrix was used to measure how well each classifier performed. Evidence suggests that pruning might improve precision.

1.1 Motivation:

One of the most important parts of India's economy is agriculture. The soil's stability has been deteriorating for some time now, due to pesticide use and industrialization. When it comes to improving productivity, many farming methods fall short. The lack of knowledge about which crops are most suited to different regions is a common problem for Indian farmers. The particular needs of the soil affect production.

Several obstacles stand in the way of Indian farmers using climatic data to choose the best agricultural technology and crops [12][13]. The inability of Indian farmers to maximize agricultural production by selecting the right crops is a persistent problem. Economic and topographical factors must be considered to maximize agricultural yields. The goal is to maximize agricultural production yields by optimizing production [14]. Minimizing expenditures is the goal of manufacturing.

The Indian Agriculture and Welfare Department subsequently examined traditional machine learning methods such as Decision Trees (DT), Gradient Boosting (GB), Gaussian Naïve Bayes (GNB), and Multi-Layer Perception Random Forest Regression (MLPRFR) to improve the accuracy of agricultural output prediction. Ultimately, we created a recommender system that proposes crops appropriate for a specific plot of land and the ensuing growing season. Below is a concise summary of the key findings of the paper:

- Gathering, organizing, and analyzing data on environmental conditions, cultivated area, and earlier output is the goal of this machine learning dataset, which will be used to predict important crops.
- To better anticipate tomato yields, research and analysis were conducted to build a basic machine learning system using effective ML and ensemble ML models.
- Research and development of a novel ensemble ML approach for accurate tomato yield prediction, along with a comparison against standard ML and ensemble ML methods, including testing of the constructed ensemble ML model to prove its significance and excellence.

- Using a set of existing crops as input, an algorithm of recommendations was created to determine which tomato variety would be most suited for cultivation at a certain location in the following season.

2. Related Works:

Consequently, forecasting crop yield has become a challenging subject in the field of precision agriculture. Several factors are involved, with the most significant being the impact of rising temperatures on agricultural output. To forecast the impact of climate change on agricultural productivity, it is crucial to develop accurate prediction models. The future viability of agriculture is jeopardized by environmental changes, namely the escalation of temperatures and erratic weather patterns.

Utilizing multiple datasets is crucial to tackling this intricate topic. The characteristics addressed in this list, while not exhaustive, consist of soil composition, seed variety, fertilizer application, climatic conditions, and weather patterns [15]. Utilizing statistical models to predict agricultural productivity may be arduous and time-consuming. Conversely, the rise of big data has opened up possibilities for more sophisticated analytical methods, such as machine learning [16].

Descriptive and predictive machine learning models may be developed based on specific research topics and problems. Predictive models are used to forecast future outcomes, whilst descriptive models analyze data and elucidate previous occurrences [17–20]. Soil type, weather patterns, and previous harvest yield are just a few of the factors that crop recommendation systems (CRS) use to help farmers make educated decisions about which crops to plant [21–23]. Using CRS, farmers may get the most out of their water, fertilizer, and pesticide budgets.

In CRS, machine learning models often use neural networks, decision trees, and support vector machines. On the other hand, clients may lose faith in the system due to the intricacy and opacity of these models. In order to improve early sickness diagnosis, machine learning (ML) models are developed in several steps, such as preprocessing, feature extraction, and classification [24]. For predicting cardiac problems, the ML model employed methods for ensemble learning and hyperparameter tuning [25].

Changes in the global, regional, or local weather patterns that have an impact over an extended period of time are called climate change. The already difficult effort of combating climate change and cutting emissions of greenhouse gases is made much more difficult by the myriad of legal and administrative hurdles that must be surmounted [26]. The impacts of climate change are anticipated to worsen food insecurity, malnutrition, and hunger for a large portion of the world's population, especially in South Asia, Sub-Saharan Africa, and small islands [27].

The spread of agriculture in Africa is gravely threatened by climate change [28]. The nutrient density of farm-raised goods is very sensitive to environmental variables like temperature and air quality, as they influence soil composition. As a result, the next generation must figure out how to lessen the impact of environmental factors on crop yields. Agricultural output projection is still being actively monitored by scientists worldwide [29].

With the use of remote sensing data, state-of-the-art deep learning systems have been developed that can accurately forecast harvest yields. Researchers in developing nations employed a Convolutional Neural Network (CNN) with a Gaussian process component and a dimensional reduction approach to predict agricultural production every year. Gradient boosting, support vector regression (SVR), and k-nearest neighbours were used to forecast the yields of sunflower, sugar beet, potato, spring barley, and wheat crops in France, Germany, and the Netherlands [30–31]. Integrating convolutional and recurrent neural networks (CNNs and RNNs, respectively), researchers built a multilayer deep learning architecture that can grasp spatial and temporal features. Their goals were to determine how different databases affected the prediction task and how well the proposed method predicted US Corn Belt production. Conducting their trials in the US Corn Belt states, they used time-series satellite images and data on soil parameters as inputs. From 2013 to 2016, they predicted the quantity of corn grown at the county level [32].

3. Proposed Methodology:

As a nation that relies mostly on agriculture, India must precisely anticipate its agricultural output, which is particularly challenging given the wide variety of crops that might be grown in a given growing season. For the time being, farmers must rely on their expertise and experience when deciding which crops to grow, which does not always lead to accurate predictions.

A country's economy would take a major hit if its agricultural production, which is so important to the economy, went through phases of prosperity and decline. Estimates of agricultural productivity are also necessary for the government to anticipate how much food will be harvested in the next year.

We have developed an ensemble machine learning model, MLPRFR, to forecast agricultural yields in India's varied landscapes based on our thorough research of the most popular classical machine learning models. The current climatic conditions, the size of the farmed area, and past production records are all taken into account by this model.

Farmers may use MLPRFR to help them choose the most productive crops for cultivation, and policymakers can use it to generate more accurate production predictions for the next year. To emphasize, no prior study has attempted to forecast agricultural output within the specific context of India.

Identifying and addressing environmental concerns that affect agricultural output in India requires extensive engagement with farmers. To enhance the comprehensibility and availability of the proposed methodology, this experiment provides a comprehensive description of the concepts and materials used.

The proposed system employs a feature selection method that occurs after dataset processing. The machine learning models were provided with selected relevant attributes. The variables of the model were optimized to enhance accuracy. Several machine learning classification approaches, such as MLPRFR algorithms, were used to build the prediction model.

By fine-tuning hyperparameters, valuable features were extracted. After constructing the models, performance metrics were used to assess their effectiveness. Figure 1 depicts the basic architecture of the proposed system.

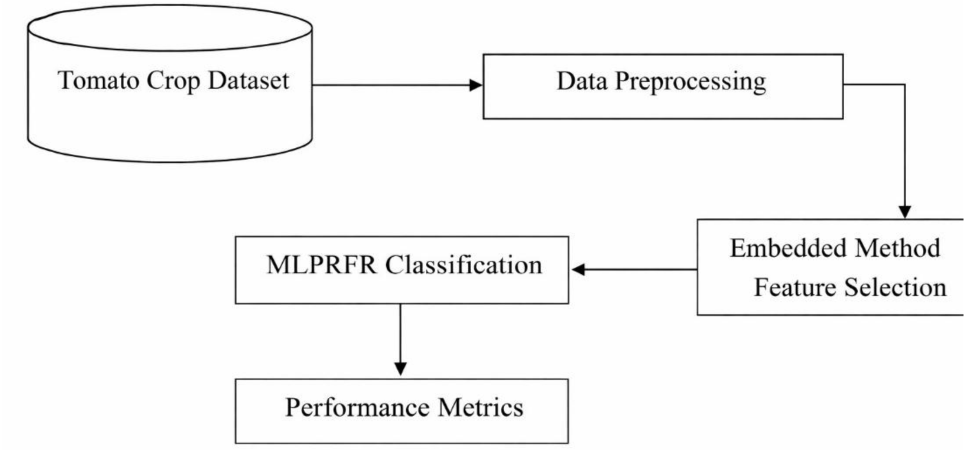

Figure 1: System Architecture

3.1 Dataset:

Temperature, humidity, pH, precipitation, nitrogen (N), phosphorus (P), and potassium (K) were among the seven variables included in a 2,200-record crop recommendation dataset. To further understand the dataset's features, the soil content necessary for each crop was established.

The dataset was split into a training set and a validation set before performing cross-validation. Kaggle was used to extract the data [33]. Factors essential for crop selection, such as humidity, temperature, rainfall, pH, and requirement ratios, were included in this dataset, which is why the algorithm was trained on it.

Nitrogen, potassium, phosphorus, humidity, and rainfall requirements vary per crop. There are no blanks or missing characteristics in the crop recommendation dataset.

3.2 Data Preprocessing:

To read the initial dataset for artificial intelligence, we engage in a procedure known as preprocessing. Omitting the procedure of labeling data during initial processing may be appropriate because our dataset exclusively comprises numerical values.

Furthermore, to enhance the trainability of the data, normalization is performed as the final step. The following are the preprocessing stages:

3.2.1 Data Cleaning:

This stage includes missing value management, data formatting, and error correction. It replaces missing data with the average of a characteristic. Environmental variable values are constant since a country produces agriculture. We may impute missing values by averaging neighbouring values. values in the data set we have. We restructure all the data through a specified structure for certain attributes to improve ML model performance [34].

3.2.2 Data Reduction:

A detrimental effect on ML models can be the presence of superfluous, redundant, or trash data. We remove unnecessary, redundant, and trash variables to build a robust ML dataset.

3.2.3 Data Normalization:

The dataset is then normalized to values in integers so that ML techniques may use it. For this reason, we normalize the dataset using the min-max method.

3.3 Feature Selection:

To address the issues arising from high dimensionality in many machine learning applications, selecting features is an essential preprocessing step. Before using a learning approach, it is necessary to choose a subset of the available data attributes. Feature selection is a process that aims to limit the characteristics of features used in a machine learning model. It involves evaluating the initial collection of features and eliminating those that are unneeded, redundant, or noisy, depending on a specific assessment criterion [29]. In this experiment, the names of several agricultural products serve as the variable of dependence. Our proposed method accounts for some exogenous factors, including humidity, precipitation, temperature, pH, nitrogen, phosphorus, and potassium. Feature selection approaches, including filter and embedded methods, principal component analysis (PCA), and correlation analysis, are used for the dataset at this step. The best indicators for the agriculture industry were identified using both methods. Independent of any learning method, filter techniques evaluate and choose features based on dataset properties to establish their worth [31]. The embedded approach finds the best potential feature combination for a particular machine-learning algorithm by methodically evaluating all feasible combinations. There is a numerical representation of how well the output labels match up with each attribute. This study finds the best environmental indicators by using an embedded selection method. Comparing the embedded feature selection approach to various feature selection strategies yielded the results shown in Table 1.

Table 1: Feature Selection for Different Methods

Algorithm	Count of Features	Feature Selection
PCA	4	Temperature, pH, humidity
Embedded	7	Temperature, humidity, pH, precipitation, nitrogen, phosphorus, and potassium

3.4 MLPRFR Classification:

The objective of this research was to determine the most effective classifier for predicting crop yield by evaluating the performance of several machine learning-based classifiers on the crop recommendation dataset. Given that our major emphasis is on MLPRFR was on multiclass classifiers, the subsequent classifiers were chosen randomly. Reducing data variability during model training is the main goal of adopting ensemble techniques. One way to make better use of models is to fit them into the data. The MLPRFR ensemble uses RF, an ensemble based on trees, and Multi-layer Perception, a method based on distance. The main objective of these penalized terms is to ensure regularization, which reduces the weights for the model to zero or near zero, hence preventing the model from excessively fitting the data. Contrary to curve-based approaches, the performance of an RF is not influenced by nonlinear circumstances. Hence, when there is a significant amount of non-linearity between the independent variables, the Random Forest (RF) may outperform alternative methods that rely on curves. It can often automatically manage irregularities and sustain durability. Instead of calculating distance, it utilizes a rule-based approach, eliminating the need for feature scaling (standardization and normalization). A novel ensemble model was created, using techniques that can independently regulate overfitting. This eliminates the need for further preprocessing during training and testing. This MLPRFR regression solution is created using a second-order ensemble technique that incorporates blending. Compared to the three pillars of ensemble methodology, MLP, RR, and RF, MLPRFR has a different operating mechanism. Ensemble approaches are both used in both the MLPRFR and RF methods that have been detailed. The individual components of MLPRFR work together as a blended ensemble to produce the finished product. To train individual trees and get the ultimate output, RF uses a sweeping ensemble approach that separates data into several packs. Because the data points are so variable, the benchmark models might experience overfitting and so perform worse than expected. However, these difficulties are overcome by the proposed method, which shows more tolerance and flexibility while understanding the dataset. We were able to solve this particular regression issue using our MLPRFR system, which resulted in better performance and better dataset fitting. The prerequisites for extending the proposed MLPRFR system to a different dataset are simple. The training data must be compatible with the MLPRFR infrastructure, much like the traditional training and testing technique in machine learning. Then, using the remaining testing data, the model is evaluated. We have already covered the MLPRFR architecture up top. On the other hand, MLPRFR is a foundational machine learning technique, and incorporating hyperparameter modification into it could increase its performance on various datasets. MLPRFR_Tomato_Recommendation Pseudocode below listed a Table 1.

Table 1: MLPRFR_Tomato_Recommendation Pseudocode

Step	Description
Step 1	Begin
Step 2	Load tomato crop dataset D
Step 3	Handle missing values and normalize numerical features
Step 4	Extract feature set F = {Soil, Climate, Environmental parameters}
Step 5	Extract target labels Y (Crop suitability classes)
Step 6	Split dataset D into Training set (Dtrain) and Testing set (Dtest)
Step 7	Initialize Multi-Layer Perceptron (MLP) with: • Input layer = |F| • Hidden layers = h neurons • Output layer = k classes
Step 8	Train MLP using backpropagation on Dtrain
Step 9	Generate intermediate feature representations FM from trained MLP
Step 10	Initialize Random Forest classifier with n trees
Step 11	Train Random Forest using FM as input features and Y as labels
Step 12	Aggregate predictions from all decision trees using majority voting
Step 13	For each test sample x in Dtest:
Step 14	Extract feature vector fi
Step 15	Pass fi through trained MLP to obtain transformed features fmi
Step 16	Classify fmi using trained Random Forest
Step 17	Assign final tomato recommendation label
Step 18	Evaluate model performance using Accuracy, Precision, Recall, and F1-score
Step 19	End

3.5 Model Evaluation:

For crop production prediction, we use both traditional and novel ensemble ML techniques, as well as the ensemble ML scheme we have developed. These methods are trained using the training data, and the algorithm learns the data sequences before making a prediction. The four assessment measures used to determine the ML algorithms' effectiveness are MAE, MSE, RMSE, and R2. The mean squared error (MSE) is the sum of the squares by which the actual value differs from the projected value. If we assume that the goal follows a normal distribution, then using MSE in regression will punish big mistakes more than small ones. With the MSE calculated as shown in Equation 1:

$$MSE = \frac{1}{S}\sum_{i=1}^{S}(\, D_i - \widehat{D_\iota})^2$$

On average, the MAE tells us how much of a mistake the forecast may make. Measured relative to a collection of points, MSE shows how near it is. This is achieved by squaring the errors, which are the variations in distance between the points and the line used for regression. To get rid of any negative signals, you have to square it. Here is how the MAE is determined, as shown in Equation 2:

$$MAE = \frac{1}{S}\sum_{i=1}^{S}(\, D_i - \widehat{D_\iota})$$

According to the RMSE is defined as the standard deviation of the residuals, that are also referred to as prediction errors. The root-mean-squared error (RMSE) measures the spread of the residuals, reflecting the distance between the data points and the regression line. In other words, it demonstrates the degree of compactness of the data points around the line that best represents them. The determination of the root mean square error (RMSE) is as follows, as shown in Equation 3:

$$RMSE = \sqrt{\frac{1}{S}\sum_{k=1}^{S}(\, D_i - \widehat{D_\iota})^2}$$

An important statistic in statistics is the coefficient of determination (R2), which measures how much variation in one variable can be explained by changes in another. Predicting future events based on current data is the main objective of this score. This statistic measures how well the model can reproduce the actual outcomes. It relies on the model's ability to account for a certain percentage of the total variation in results. Here is the formula for R2 as shown in Equation 4:

$$R2 = \frac{\sum(D_i - \widehat{D_\iota})^2}{\sum(D_i - \bar{D_\iota})^2}$$

3.6 Metrics for Evaluation:

Accuracy, precision, recall, and the F1 score were all used to evaluate the effectiveness of our proposed method. The four measures used to calculate these metrics are true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

TP: The ratio of samples properly identified by the method of detection model to the total number of samples.

TN: The fraction of samples for which the detection model's classification of their true type is accurate.

FP: The actual sample type is normal; however, the detection model incorrectly identified a large number of samples as coming from a crop recommendation.

FN: The number of crop recommendation samples that were incorrectly classified as "normal" samples.

Accuracy: This represents the proportion of input samples for which the detection model reached a positive verdict, as shown in Equation 5.

$$AC = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall: It is the proportion of crop recommendation samples accurately identified by an identification model out of the total number of attack samples, as shown in Equation 6.

$$R = \frac{TP}{TP + FN}$$

Precision: It represents the proportion of samples that the detection model has identified as being subject to a crop recommendation, as shown in Equation 7.

F1-Score: It is an aggregate measure of accuracy that takes into both recall and precision, as shown in Equation 8.

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

4. Results and Discussion:

Python's sci-kit-learn module was used to build all of the traditional and ensemble machine-learning frameworks that were looked at, particularly the one we recommend, MLPRFR. We employ error metrics like RMSE, R2, MAE, and MSE, and take into account the actual vs. projected curve while evaluating the trained ML models. We concentrate on the quantity of crop production in the goal area and utilize an eight-feature dataset to train supervised algorithms to forecast crop output. We calculate the average of the outcomes from three test sets, each with ten trials. Three ratios are used to separate the test and training data: 70:30.

Embedded, descriptive, and exploratory approaches, as well as principal component analysis, are used in the study. After running the dataset through these filters, the most important feature combination for classification models was identified, as shown in Figures 2a and 2b. We updated the dataset and deleted the seven most significant characteristics (temperature, humidity, pH, precipitation, nitrogen (N), phosphorus (P), and potassium (K)) to maximize the models' performance, following the suggestions of the integrated ranking filter. Using the embedded method, we were able to choose just the most essential and practical features.

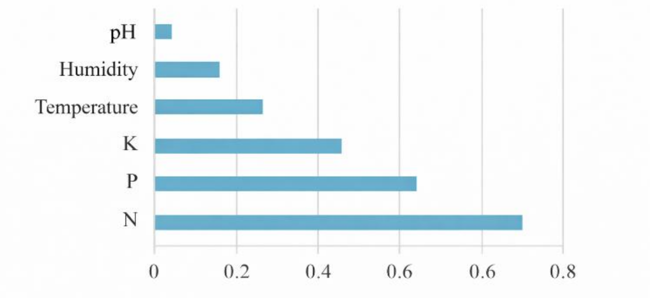

Figure 2a: Feature Selection for PCA Methods

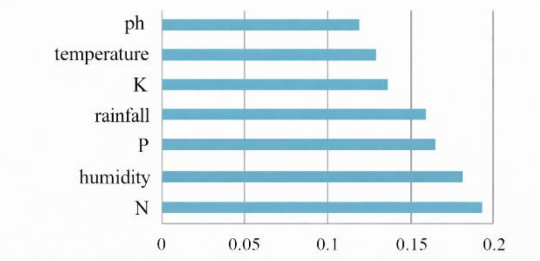

Figure 2b: Feature Selection for Embedded Methods

To assess the machine learning approaches that were studied and proposed for banana production forecasting, the R2 score, along with additional measures were used shown in Table 2 and Figure 3.

Table 2: Comparison for tomato production forecasting

Models	MSE	MAE	RMSE	R2
DT	0.486	0.502	0.518	0.607
GB	0.458	0.357	0.485	0.705
GNB	0.745	0.658	0.749	0.806
MLPRFR	0.901	0.924	0.907	0.950

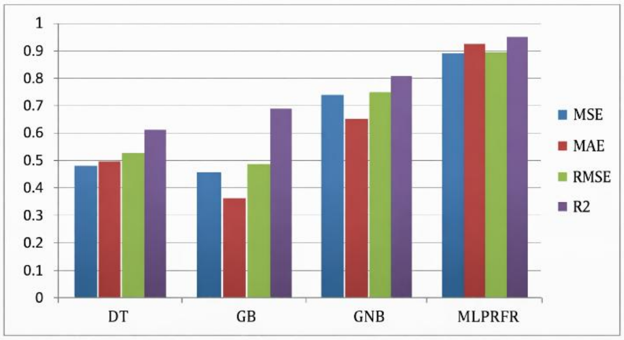

Figure 3: Comparison of tomato production forecasting

To assess the machine learning approaches that were studied and proposed for rice production forecasting, the R2 score, along with additional measures were used, as shown in Table 3 and Figure 4.

Table 3: Comparison for rice production forecasting

Models	Recall	Precision	F1-Score	Accuracy
DT	0.089	0.020	0.098	0.974
GB	0.105	0.026	0.142	0.980
GNB	0.084	0.030	0.050	0.924
MLPRFR	0.095	0.028	0.090	0.989

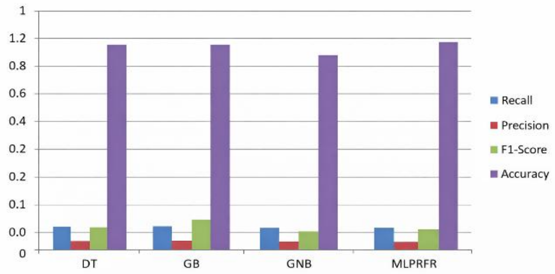

Figure 4: Comparison of rice production forecasting

5. Conclusion:

Gathering data on agricultural crop output predictions from different meteorological agencies and agricultural research institutions in India and assembling it into a usable dataset has been the primary goal of this project. The research starts by using an ensemble of three popular classical machine learning methods. We suggested using an ensemble method called MLPRFR to improve the precision of agricultural output forecasts. We discovered that our suggested ensemble method, MLPRFR, beats both the standard ML algorithms and the ensemble ML algorithms after doing extensive testing on various ML techniques. With the best R2 score and the lowest error rate of all the machine learning algorithms tested, MLPRFR comes out as the clear winner. In addition, to show that our MLPRFR method is better than other ML methods, we used an MSE, RMSE, R2, and MAE test. In addition, the final tally shows that yearly wheat output is down, while monthly yields of rice, tomatoes, bananas, and wheat are all up significantly. On top of that, we've included a crop recommender system that tells you which crops will do best on a certain plot of land in the next growing season.

References:

1. 2017 International Conference on I2C2, “Agriculture decision support system using data mining”, Prof. Rakesh Shirsath; Neha Khadke, Divya More.
2. Rezk NG, Hemdan EE, Attia AF, El-Sayed A, ElRashidy MA. An efficient IoT-based smart farming system using machine learning algorithms. Multimedia Tools and Applications. 2021; 80:773-97.
3. 2017 IEEE Region 10 Humanitarian Technology Conference, “RSF: A Recommendation System for Farmers”, Miftahul Jannat Mokarrama; Mohammad Shamsul Arefin.
4. IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO) 2015, “XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction”, Aakunuri Manjula, Dr. G Narsimha.
5. Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018), “Use of Data Mining in Crop Yield Prediction”, Shruti Mishra, Priyanka Paygude, SnehalChaudhary, Sonali Idate.
6. 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, T.N., India. 6 - 8 May 2015. pp.138-145, “Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique”, Rakesh Kumar, M.P. Singh, Prabhat Kumar, and J.P. Singh.
7. Jansson, C., Faiola, C., Wingler, A., Zhu, X. G., Kravchenko, A., De Graaff, M. A., et al. (2021). Crops for carbon farming. Front. Plant Science. 12, 636709. doi: 10.3389/fpls.2021.636709
8. Zhang, Z., Jin, Y., Chen, B., and Brown, P. (2019). California almond yield prediction at the orchard level with a machine learning approach. Front. Plant science. 10, 809. doi:10.3389/fpls.2019.00809

9. Goldstein, B., Moses, R., Sammons, N., and Birkved, M. (2017). Potential to curb the environmental burdens of American beef consumption using a novel plant-based beef substitute. PloS One 12 (12), e0189029. doi: 10.1371/journal.pone.0189029
10. Van Klompenburg, T., Kassahun, A., and Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agriculture. 177, 105709. doi: 10.1016/j.compag.2020.105709
11. Das, A. K., Rahman, M. A., Rahman, M. M., Saha, S. R., Keya, S. S., Suvoni, S. S., et al. (2022). Scaling up of jujube-based agroforestry practice and management innovations for improving efficiency and profitability of land uses in Bangladesh. Agroforest. Syst. 96 (2), 249–263.
12. Jain A. "Analysis of growth and instability in the area, production, yield, and price of rice in India", Journal of Social Change and Development, 2018;2:46-66
13. Manjula E, Djodiltachoumy S, "A model for prediction of crop yield" International Journal of Computational Intelligence and Informatics, 2017 Mar;6(4):2349-6363.
14. Sagar BM, Cauvery NK. "Agriculture Data Analytics in Crop Yield Estimation: A Critical Review", Indonesian Journal of Electrical Engineering and Computer Science, 2018 Dec;12(3):1087-93.
15. Bali N, Singla A (2021) Deep learning based wheat crop yield prediction model in the Punjab region of north India. Appl Artif Intell 35(15):1304–1328
16. Van Klompenburg T, Kassahun A, Catal C (2020) Crop yield prediction using machine learning: A systematic literature review. Comput Electron Agric 177:105709
17. Alpaydin E (2020). Introduction to machine learning. MIT Press.
18. Tarek Z et al (2023). Soil erosion status prediction using a novel random forest model optimized by the random search method. Sustainability 15(9):9. https://doi.org/10.3390/su15097114
19. Shams MY, Tarek Z, Elshewey AM, Hany M, Darwish A, Hassanien AE (2023) A machine learning-based model for predicting temperature under the effects of climate change. In: The Power of Data: Driving Climate Change with Data Science and Artificial Intelligence Innovations, A. E. Hassanien and A. Darwish, Eds., in Studies in Big Data. Cham: Springer Nature Switzerland, 2023: 61–81. https://doi.org/10.1007/978-3-031-22456-0_4
20. Elshewey AM et al (2023) A novel WD-SARIMAX model for temperature forecasting using the daily Delhi climate dataset. Sustainability 15(1):1. https://doi.org/10.3390/su15010757
21. Patel K, Patel HB (2023). Multi-criteria agriculture recommendation system using machine learning for crop and fertilizer prediction. Curr Agricult Res J 11(1), 2023.
22. Mittal N, Bhanja A (2023). Implementation and identification of crops based on soil texture using AI. In: 2023 4th International Conference

on Electronics and Sustainable Communication Systems (ICESCS), IEEE. 1467–1471.

23. Fenz S, Neubauer T, Heurix J, Friedel JK, Wohlmuth M-L (2023) AI- and data-driven pre-crop values and crop rotation matrices. Eur J Agron 150:126949. https://doi.org/10.1016/j.eja.2023.126949
24. Arif MS, Mukheimer A, Asif D (2023). Enhancing the early detection of chronic kidney disease: a robust machine learning model. Big Data Cognit Comput 7(3):3. https://doi.org/10.3390/bdcc7030144
25. Asif D, Bibi M, Arif MS, Mukheimer A (2023). Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. Algorithms 16(6):6. https://doi.org/10.3390/a16060308
26. McEldowney JF (2021) Climate change and the law. In: the impacts of climate change, Elsevier. 503–519.
27. de Oliveira AC, Marini N, Farias DR (2014) Climate change: New breeding pressures and goals. Encyclopedia Agricult Food Syst 2014:284–293
28. Williams TO, et al. (2015) Climate-smart agriculture in the African context. Unlocking Africa's Agricultural Potentials for Transformation to Scale, FAO and UNEP, Abdou Diouf International Conference, Dakar, Senegal, pp. 1–26, 2015.
29. Reddy PS, Amarnath B, Sankari M (2023). Study on machine learning and backpropagation for a crop recommendation system. In: 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), IEEE. 1533–1537.
30. J, Li X, Low M, Lobell D, Ermon S (2017) Deep Gaussian process for crop yield prediction based on remote sensing data. In: Thirty-First AAAI conference on artificial intelligence.
31. Paudel D et al (2021) Machine learning for large-scale crop yield forecasting. Agric Syst 187:103016
32. Sun J, Lai Z, Di L, Sun Z, Tao J, Shen Y (2020) Multilevel deep learning network for county-level corn yield estimation in the US corn belt. IEEE J Selected Top Appl Earth Obs.
33. https://www.kaggle.com/datasets/atharvaingle/croprecommendation-dataset
34. Stekhoven, D. J., and Buhlmann, P. (2012). MissForest non-parametric missing value imputation for mixed-type data. Bioinformatics. 28 (1), 112–118. doi: 10.1093/bioinformatics/btr597