International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Resilient Forecasting Model for Sustainable Agriculture and Optimal Yield Production

[1]K. Suganthavalli, [2]C. Meenakshi

Submitted: 05/05/2024 **Revised**: 18/06/2024 **Accepted**: 25/06/2024

Abstract: Given India's enormous breadth and dense population, the forecast of agricultural output is vital for guaranteeing food security. The process, however, is hard due to the influence of a myriad of elements, such as farming techniques, environmental circumstances, and technology improvements. Existing machine learning (ML) models have issues due to the quality and variety of data, model overfitting, sophisticated model architectures, insufficient feature engineering, and temporal dependencies. Therefore, a robust and efficient model that addresses these difficulties is important. In this work, an analysis was undertaken employing five prominent ML algorithms — Random Forest (RF), XGBoost, Decision Tree (DT), Support Vector Machine (SVM), and Linear Regression (LR) — on a crop prediction dataset collected from Kaggle. Algorithms that displayed the highest coefficient of determination (R^2) were selected to develop a hybrid model for aggregate prediction. Results revealed that the suggested hybrid model, covering DT, XGBoost, and RF, surpassed individual classifiers in terms of R^2score and outperformed the existing models, obtaining an accuracy of 98.6%. This provides a robust and efficient paradigm for agricultural yield forecasting. Consequently, a user-friendly tool, 'Crop Yield Predictor', was developed, rendering the model accessible and practical for on-ground applications in agriculture. This technology effectively turns complicated data and algorithms into actionable insights, bridging the gap between advanced machine learning techniques and practical agricultural applications.

Keywords: *Machine learning, agricultural yield prediction, hybrid model, decision tree, support vector machine, random forest, gradient boosting and linear regression*

INTRODUCTION

Agriculture accounts for 60.45% of land use in India. Its agrarian economy stands as a cornerstone of its socioeconomic fabric, with agriculture firmly entwined in its history and present. The nation's reliance on the yearly monsoon and farming techniques determines the trajectory of its economy and the well- being of its inhabitants [1]. As one of the most populous countries internationally, India's ability to assure food security rests upon the efficacy of its agricultural yields. In this scenario, the correct prediction of crop yields becomes critical importance. On the one hand where farmers want timely guidance to anticipate crop output and establish effective methods to boost agricultural produce thereby earning better return on investments, Governments on the other hand must be able to accurately predict agricultural production to achieve national food security and make knowledgeable decisions regarding imports thereby saving crucial forex. To address the rising food demands of India's burgeoning population, advanced technology methods in agriculture are essential.

Previously, farmers depended on their own experiences and precise historical data to anticipate crop yields and make critical production decisions based on the prediction. However, in recent years, new advances such as crop model simulation, precision agriculture, and machine learning have surfaced to estimate yield more precisely, as can evaluate huge volumes of data using high-performance computers [2-5]. However, the process of projecting agricultural yields effectively is plagued with intricacy. It comprises the delicate interplay of a myriad of components, covering agricultural practices, environmental variables, and technological breakthroughs. The challenge is further compounded by the varied and frequently unpredictable nature of these variables, making typical forecasting approaches fall short in giving solid forecasts.

This strategy not only maximizes agricultural yields but also minimizes resource wastage and environmental effect.

[1]Department of Computer Science Vels Institute of Science, Technology and Advanced Studies (VISTAS) Pallavaram, Chennai, Tamil Nadu, India, suganthavallimca@gmail.com
[2]Associate Professor, Department of Computer Application Vels Institute of Science, Technology and Advanced Studies (VISTAS) Pallavaram, Chennai, Tamil Nadu, India,
meenasi.c@gmail.com

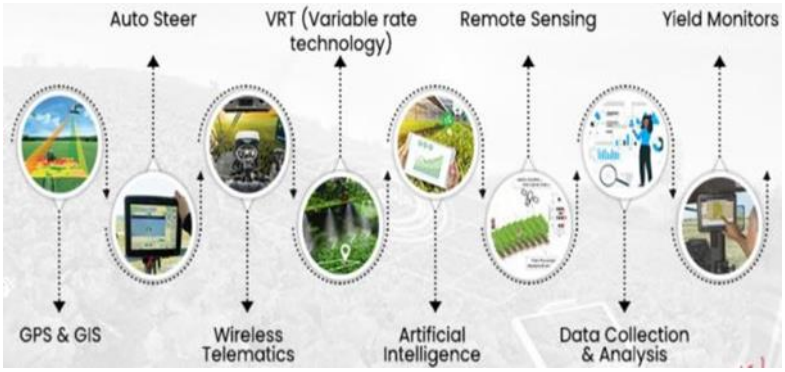

Fig 1. Components of precision agriculture

Statistical models, in comparison, provide an approach that predicts direct correlations between predictor variables and crop yield within a particular dataset without taking into consideration underlying mechanisms in crop ecology and physiology [7]. When given adequate and trustworthy data, statistical models can produce reasonable predictions, but they may be hampered by the boundaries of the training data. However, commonly used performance evaluation metrics are feasible to get on statistical models that are helpful for uncertainty investigations at regional scales and are less dependent on field calibration data. For estimating agricultural yield, statistical models like multiple linear regressions (MLR) and simple linear regressions are often applied [8].

There are a wide range of advantages to utilize machine learning (ML) models to forecast agricultural productivity. Machine learning lays an emphasis on detecting patterns and correlations in data settings to correctly anticipate yields depending on numerous variables. ML models nonetheless must train on datasets that reflect previous experiences and results to generate prediction models. During the training process, the parameters of the models are determined by using previous data. During the testing phase, the performance of the model is evaluated by making use of a portion of the historical data that was not utilized during the training phase [9]. As a result of their ability to adapt and learn from the nonlinear and dynamic processes of crop growth, machine learning algorithms are essential for the creation of accurate estimates for agricultural production. Due to learning capabilities from datasets, machine learning models are particularly suited for predicting agricultural outputs at big scale. Further, their adaptability to different parameters like crop varieties, temperature, rainfall, humidity, nutrient content (nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, Sulphur, manganese, and copper), pH levels and geographical area, make ML techniques them suitable to model and forecast crop yields [10].

While the potential of machine learning models to anticipate crop yields holds promise, existing approaches have encountered a range of challenges. Chief among these issues is the scarcity of comprehensive and high-quality data, the intrinsic variability within agricultural systems, the hazard of overfitting sophisticated models, and the intricacies of temporal connections. Additionally, the models' efficacy has been impeded by the intricacy of their structures and the insufficiency of thoughtful feature engineering.

Considering these challenges, there emerges a compelling need for the creation of more robust and efficient prediction models that can aid in the domain of sustainable crop production. Modern machine learning techniques such as hybrid or ensemble modeling, driven by their capacity to make several models work together and handle complex and heterogeneous datasets, present a transformative possibility. Leveraging these methods could enable more precise projections, benefiting farmers, policymakers, and stakeholders alike in making educated decisions. In this research, we harness the individual potential of machine learning methods such as Decision Tree, XGBoost, Random Forest, Support Vector Machine, and Linear Regression utilizing the agriculture dataset comprising of top 10 globally consumed crops from Kaggle repository. Choosing the best three algorithms in terms of performance, the study then exemplifies the fusion of these algorithms to present a hybrid ML framework for effective agricultural production prediction and offer this as a platform for usage by all stakeholders. Therefore, the important contributions from this study are:

- ✓ A hybrid ML framework leveraging top three independently performing algorithms for an efficient and robust crop yield prediction model.
- ✓ To ensure the accessibility of this model for farmers and policymakers, a user-friendly interface named ‘Crop Yield Predictor’ is developed, delivering efficient crop yield projections. This tool can assist in optimizing resources and boosting sustainable food production.

The manuscript is organized as below: Section 1 gives an overview of agricultural yield prediction using machine learning and statistical models. Existing methods for

agricultural production prediction are evaluated in Section 2 under related works. Section 3 outlines the resources and methods for agricultural production prediction that have been given. Section 4 covers the results and discussion. Section 5 closes with references and makes conclusions.

RELATED WORKS

In the subject of crop production prediction, various studies have been proposed largely employing machine learning algorithms compared to statistical or ensemble models.

Authors [11] focus on estimating wheat yields using a combination of machine learning algorithms and advanced sensing technologies. The researchers study how these strategies can boost the accuracy of projecting wheat yields, which is vital for agricultural planning and management. The research covers the implementation of machine learning techniques coupled with sophisticated sensing instruments to acquire data relevant to wheat growth and production. By examining this data, the study reveals how the combination of various technologies might lead to improved projections of wheat yields, contributing to more informed decision-making in agricultural practices.

Authors [12] address the present and potential applications of statistical machine learning methods in agricultural machine vision systems. The authors highlight how these algorithms are currently being applied in numerous parts of agriculture, notably focusing on machine vision systems. They also highlight the potential future uses and developments in this subject, emphasizing the importance of statistical machine learning approaches in enhancing agricultural processes and systems. The study gives insights into the growing environment of agricultural technology and its integration with machine learning for better efficiency and productivity.

Authors [13] give a detailed survey on the integration of agrarian elements and machine learning algorithms for yield forecasting. The authors dive into the numerous methodologies that integrate agricultural characteristics and machine learning techniques to estimate crop yields. They review the existing approaches, discuss the obstacles faced in yield forecasting, and estimate the potential benefits of merging agrarian elements with machine learning algorithms. The paper presents insights into the state-of-the-art in yield prediction approaches, emphasizing the significance of using both domain knowledge and advanced machine learning techniques to boost the accuracy of yield forecasts in agricultural contexts.

The authors [8] explore the implementation of Random Forests, a machine learning technology, for predicting crop yields at both global and regional scales. The authors study the potential of Random Forests in improving agricultural production estimates by combining several data sources, including climate, soil, and remote sensing data. They demonstrate that Random Forests may successfully capture complicated interactions between these parameters and agricultural yields, leading to more accurate predictions. The paper underscores the usefulness of machine learning approaches like Random Forests in boosting our capacity to anticipate crop yields, which is vital for global food security and agricultural planning.

The authors [14] present an ensemble method that combines different machine learning algorithms to generate a more robust prediction model for crop yield. They utilize many aspects and data linked to crop growth and climatic variables to train their model. By harnessing the strengths of many algorithms, the ensemble model tries to increase the accuracy of crop production projections. The research underlines the potential of ensemble approaches in boosting the trustworthiness of forecasts in the agricultural environment.

Authors [15] focus on estimating crop production with various machine learning systems. They examine the approach and results of utilizing these algorithms on datasets including relevant agricultural information. The paper's purpose is to highlight the potential of machine learning in forecasting crop yields, thereby contributing to enhance agricultural planning and decision- making.

Authors [16] focuses on estimating rice crop yields in India using support vector machines (SVM). They employed pertinent data relating to rice agriculture, environmental conditions, and historical yield records to train and validate their model. The research seeks to demonstrate the feasibility and accuracy of utilizing SVM for predicting rice crop yields, which has implications for optimizing agricultural practices and food security in India.

Authors [17] offer an ensemble technique for estimating agricultural yields. The authors offer a system that integrates numerous predictive models or algorithms to provide a more accurate and dependable prediction for crop yields. They detail the implementation of their ensemble approach, employing several data sources relating to crop growth, climate, and other pertinent parameters. The research intends to highlight the effectiveness of ensemble techniques in enhancing the precision of crop output projections, helping to better agricultural decision-making and planning.

Authors [18] examine various supervised machine learning strategies to construct predictive models for crop productivity. They discuss how they exploited tagged datasets including information about crop growth, ambient conditions, and other pertinent aspects to train

and evaluate their models. The study seeks to demonstrate the applicability of supervised learning systems in accurately predicting crop yields, which can help to better agricultural planning and management strategies.

The authors [19] give a comprehensive analysis of the many machine learning methods that have been utilized in precision agriculture for the purpose of predicting crop production and estimating nitrogen status. The authors likely evaluate numerous machine learning approaches used for predicting crop yields and evaluating nitrogen levels in crops. They might discuss the benefits, limits, and comparative effectiveness of certain strategies. The paper seeks to outline the state-of-the-art in employing machine learning for precision agriculture activities, emphasizing their potential in boosting crop production efficiency and sustainability. As noted, there are very few research exploiting hybrid ML models in literature.

MATERIALS AND METHODS

In agriculture, crop yield prediction uses a number of methodologies and instruments to anticipate how much yield can be produced in a particular region. The process comprises acquiring information about the individual crop being farmed as well as environmental parameters including weather patterns and resource availability. To anticipate agricultural yield, this data is subsequently reviewed using statistical models and machine learning approaches. The stages of data collection, pre- processing, feature extraction, model selection, training and prediction are generally included in this study for crop yield prediction in agriculture.

Following exploratory analysis, four input variables (features in ML language) were identified: item, pesticide, rainfall, and temperature. The average crop output was the variable to be predicted. The mean and standard deviations of each variable in the training subset were used to normalize the data (i.e., subtract mean and divide by normal deviation). There are five distinct features in the Kaggle dataset. Table 1 includes the feature descriptions for the dataset's various parameters. Figure 2 displays the general design and workflow of the proposed system.

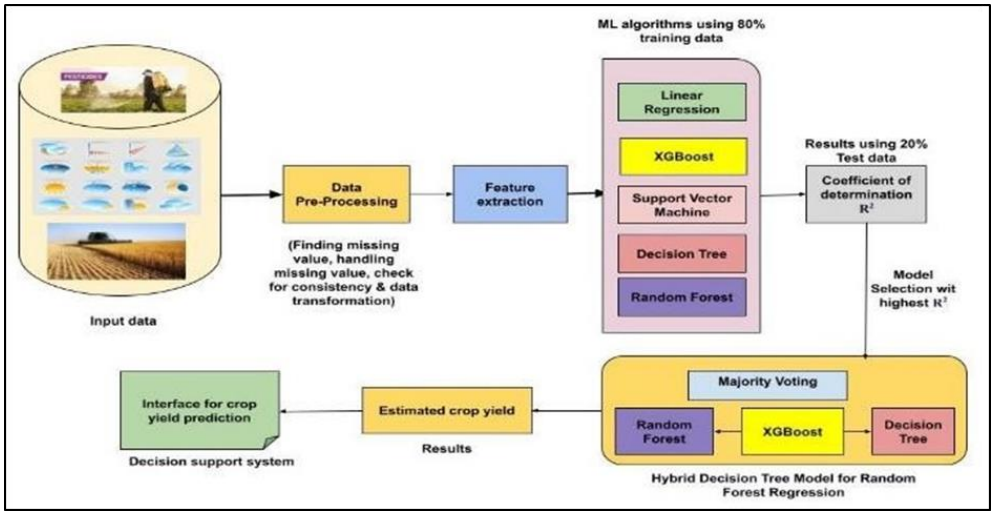

Fig 2. Overall architecture and workflow of the proposed framework

Feature	Feature ID	Description	Dependent / Independent
Item	Crop	A total of 10 distinct crops were employed in this dataset such rice, wheat, maize etc.	Independent
Pesticides (NPK)	T (tonne)	Data is taken from Food and Agricultural Organization (FAO) and active ingredients or pesticide consumption is calculated in tons	
Rainfall	AAR (average annual rainfall)	The average yearly rainfall of different season is measured in milli meters	Dependent
Temperature	AT (average temperature)	Average temperature of particular year in particular region is measured in degree Celsius	
Yield	Y (tonne)	The data is collected and measured in tonnes	

Table 1. Feature description in dataset

3.1 Machine learning models

The ML models utilized in this work were chosen after completing a literature review. We use Random Forest, XGboost, Decision Tree, Support Vector Machines, and Linear Regression models, which are briefly detailed on how they are leveraged.

Linear regression: Linear regression is a simple and commonly used machine learning technique for predicting a numerical value "(dependent variable) based on one or more input features (independent variables). In the context of agricultural yield prediction, we can use linear regression to model the link between many parameters (such weather conditions, soil qualities, etc.) and the crop yield.

For a single independent variable (feature), technically, the linear regression equation is:

$y = b_0 + b_1 * x + \varepsilon$ (1)

where:

y is the projected crop yield

b0 is the intercept term

b1 is the coefficient for the independent variable

x is the value of the independent variable (e.g., weather data)

ε represents the error term

For many independent variables (features), the equation becomes:

$y = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_n * x_n + \varepsilon$ (2)

where:

x1, x2, ..., xn are the values of the independent variables (e.g., weather, soil qualities, etc.)

b1, b2, ..., bn are the corresponding coefficients for the independent variables

For crop yield prediction, we collect historical data that contains both input features (e.g., temperature, rainfall, soil nutrients) and the matching crop yield values. The purpose of training the linear regression model is to determine the coefficients (b0, b1, ..., bn) that minimize the difference between the predicted crop yields and the actual crop yields in the training dataset [20]. The training phase comprises utilizing the least squares method to identify the ideal coefficients that best suit the data. Once the model is trained, we utilize it to forecast the crop yield for new sets of input features by putting them into the equation [21, 22].

The advantages of linear regression include that it is a straightforward model that is easy to learn and use. It provides a clear interpretation of the link between the independent variables and the dependent variable (crop yield). The coefficients in the linear regression equation provide insights into the direction and size of the effect of each independent variable on the dependent variable. This can aid in discovering which elements are most influential in predicting crop yield. These models are computationally efficient, making them useful for short analysis and prediction tasks, especially with a relatively limited number of features. Linear regression doesn't presume a particular distribution of the data, which might be useful when working with different types of agricultural data. Linear regression at best acts as a baseline model, helping to provide a foundation for more complex modeling techniques if needed.

The limitations of linear regression are that it implies a linear relationship between independent and dependent variables. If the underlying relationship is non-linear, linear regression might not represent the intricacy of the data adequately. Linear regression may struggle to capture subtle interactions and non-linear patterns evident in crop yield data. Other approaches, like polynomial regression or machine learning techniques, might be better suited for such scenarios. Linear regression is susceptible to outliers, which can disproportionately influence the model's coefficients and predictions. Outliers are not rare in agricultural statistics owing to numerous variables. When independent variables are correlated with each other, it can lead to multi collinearity concerns in linear regression. This can alter the interpretation of coefficients and the model's stability. Without suitable regularization techniques, linear regression can easily overfit (capture noise) or underfit (oversimplify) the data, resulting to poor predicted accuracy. Linear regression presupposes that the connection between each independent variable and the dependent variable is independent of other factors. It might not handle complex interactions between characteristics well. In circumstances when the relationship between variables is exceedingly complex, linear regression could not produce reliable predictions. More complex approaches may be required.

Decision Tree: This is a non-parametric supervised machine learning method that is extensively applied in classification and regression applications. Figure 3 displays a hierarchical structure with a root node, branches, internal nodes, and leaf nodes. The Gini impurity and information gain techniques are the two most utilized as splitting criterion in decision tree models, which aid in evaluating the effectiveness of each test condition and its capacity to categorize samples into a given group. Gini impurity and knowledge gain are provided by:

$H(S) = -\Sigma(P(c) * \log_2(P(c)))$ (3)

where, H(S) is the entropy of the dataset S, Σ is the sum over all possible classes c and P(c) is the fraction of occurrences in class c inside the dataset S.

$$IG(S, A) = H(S) - \Sigma((|S_v| / |S|) * H(S_v)) \quad (4)$$

where, IG(S, A) is the information gain of the dataset S by splitting on feature A, H(S) is the entropy of the original dataset S, Σ is the sum over all possible values v of feature A, |S_v| is the number of examples in the dataset S with feature A equal to v, |S| is the total number of instances in the dataset S and H(S_v) is the entropy of the subset S_v.

$$I(S) = 1 - \Sigma(P(c)^2) \quad (5)$$

Yield Y (tonne) The data is gathered from FAO, and it is measured in tonnes Dependent

where, I(S) is the Gini impurity of the dataset S, Σ is the sum over all possible classes c, P(c) is the proportion of occurrences in class c within the dataset S.

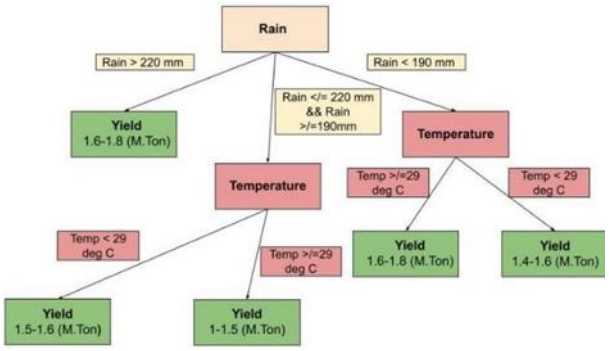

Fig 3. Representation of a decision tree implementation

The advantages of DT are that it provides a visual and intuitive picture of the decision-making process. The tree structure allows us to quickly comprehend the components and thresholds that contribute to forecasts. They can capture non-linear correlations and interactions between factors in the data, making them suited for complex agricultural systems where linear models can fall short. It can handle both categorical and numerical variables without the need for considerable preprocessing, which might be advantageous when dealing with various agricultural data. DTs are relatively robust to irrelevant features. Features that don't contribute significantly to prediction will tend to be trimmed off during the tree-building process. They can handle missing data well by placing them in a distinct branch during the tree-building process. DTs can be merged into ensemble approaches like Random Forests or Gradient Boosting, which often lead to enhanced predictive performance by reducing overfitting.

The downsides of DT include that they can grow too complex and fit noise in the training data, leading to poor generalization on fresh, unknown data. Small changes in the data can result in dramatically diverse tree architectures, rendering decision trees susceptible to variances and potentially leading to conflicting forecasts. In datasets with imbalanced classes, decision trees could have a bias towards predicting the majority class, especially if not corrected appropriately. DTs employ a greedy technique for creating the tree, which might not lead to the globally best answer in some instances. Single decision trees can have considerable variance, especially on tiny datasets, which might require ensemble methods to alleviate this issue. While DTs can handle non-linearity well, they might not be the ideal choice for applications where linear relationships dominate. Preventing DTs from developing too deep or becoming too complicated is critical to avoid overfitting and maintain interpretability.

Random Forest: A RF is an ensemble learning strategy that mixes numerous decision trees to increase predictive performance and prevent overfitting. In the context of crop yield prediction, a Random Forest model can be used to predict crop yields based on various input features. The working can be explained in below 3 steps:

Step 1. Collect a dataset that includes historical data of crop yields and corresponding input features (e.g., weather data, soil properties, etc.). Randomly sample the dataset multiple times with replacement (bootstrap samples). Each sample is used to train an individual decision tree.

Step 2. For each bootstrap sample, build a decision tree using a subset of the input features. At each split node, randomly select a subset of features to consider for splitting. Split the data based on the selected feature that maximizes a certain criterion (e.g., information gain or Gini impurity) at that node. Repeat this process recursively until the tree is fully grown or a stopping criterion is met (e.g., maximum depth, minimum samples per leaf).

Step 3. Once all decision trees are built, predictions are made by aggregating the predictions of individual trees. For regression tasks like crop yield prediction, this aggregation is usually done by calculating the average of the predictions from all decision trees.

The mathematical representation of the Random Forest model is the ensemble average of individual decision tree predictions:

RandomForest(x) = (1/N) * ∑(DecisionTree_i(x)) (6)

where:

RandomForest(x) is the predicted crop yield for input features x using the Random Forest model.

N is the number of decision trees in the ensemble.

DecisionTree_i(x) reflects the forecast of the i-th decision tree.

The advantages of RF for are that it reduces overfitting by integrating numerous decision trees, which together create more solid predictions on new, unseen data. It can capture complex non-linear correlations between input features and crop yields. The model can provide insights into feature relevance, helping to understand which characteristics contribute most to crop production estimates and it is less vulnerable to outliers and data noise due to the ensemble structure of the model.

The downsides of RF is that it can be computationally expensive and requires adjusting to maximize performance. Although RF might signal feature relevance, the ensemble structure can make the model less interpretable compared to a single decision tree. While less prone to overfitting than individual decision trees, Random Forests can nevertheless overfit if the number of trees is too great relative to the dataset size.

Support Vector Machine: SVMs are a strong machine learning method used for classification and regression tasks. For crop yield prediction, SVMs can be applied to develop a regression model that predicts crop yields based on input features. SVM regression works as below:

Step 1. We collect a dataset containing historical data of crop yields and relevant input variables (e.g., climate data, soil characteristics, etc.).

Step 2. Define the SVM regression issue. In SVM regression, the goal is to construct a hyperplane that best fits the data while minimizing the margin violations (deviations from the projected values). Choose a kernel function (linear, polynomial, radial basis function, etc.) that maps the input features into a higher-dimensional space, allowing for more intricate interactions between variables. Solve the optimization problem to discover the coefficients of the hyperplane (weights) and the bias term that minimize the error while maximizing the margin between the anticipated values and the actual crop yields.

The mathematical form of the SVM regression model can be expressed as:

y = w * x + b (7)

where:

y is the projected crop yield.

w represents the weights (coefficients) of the hyperplane.

x is the vector of input features.

b is the bias term.

Step 3. Once the SVM regression model is trained, we utilize it to estimate crop yields for new sets of input features.

The SVM regression model's purpose is to find a hyperplane that best depicts the relationship between the input features and crop yields. The forecast is created based on the distance of a fresh data point to the hyperplane.

The advantage of SVM Regression is that it can capture non-linear connections between input characteristics and crop yields using multiple kernel functions. SVMs often have high generalization properties and can handle overfitting successfully. SVMs focus on the support vectors, which are the data points that influence the margin of error, making the model less sensitive to irrelevant features.

The downside SVM Regression is that the training can be computationally costly, especially with large datasets or complicated kernels. SVMs have hyper parameters that require tuning, such as the choice of kernel, regularization parameter, etc. SVMs can be less interpretable compared to simpler models like linear regression and they might face issues in terms of efficiency and scalability when dealing with huge datasets.

XGBoost: XGBoost (Extreme Gradient Boosting) is a popular and powerful machine learning method that has shown good performance in different tasks, including regression problems like crop yield prediction. It is an ensemble learning method that combines the predictive potential of numerous weak learners (individual models) to generate a robust and accurate predictive model. The working of XGBoost can be explained as below:

Step 1. We collect a dataset that comprises historical data of crop yields and the relevant input variables (such as climate data, soil characteristics, etc.).

Step 2. Define the problem as a regression task, where the goal is to estimate crop yields (a numerical number) based on input features. Install the XGBoost library and prepare the dataset in the appropriate format.

Step 3. XGBoost features numerous hyperparameters that govern the model's behavior. These include learning rate, maximum depth of trees, number of boosting rounds, regularization terms, and more. We do hyperparameter

tweaking to discover the optimal combination of these parameters that results in the greatest predicted performance. Techniques like grid search or random search can be utilized for this purpose.

Step 4. Train the XGBoost regression model on the dataset. The model operates by sequentially adding decision trees to the ensemble, where each new tree seeks to rectify the errors committed by the preceding trees.

Step 5. Once the model is trained, we utilize it to estimate crop yields for new sets of input features.

XGBoost is noted for its accuracy and robustness, making it suited for complicated prediction problems like agricultural yield prediction. It provides insights about feature importance, helping identify which aspects are most relevant in affecting crop yields. It can capture complex non-linear correlations between input features and crop yields. Through hyperparameter tuning and regularization techniques, it can mitigate overfitting.

On the other side, building up and tuning a model will involve more effort compared to simpler methods. While it can give feature importance, the model's ensemble structure can make it less interpretable than linear models."

3.2 Methodology

The methodology utilized for agricultural production prediction involves gathering and loading the data into a .csv file, followed by data pre-processing to handle outliers and choosing key features. The performance of the model is evaluated using relevant metrics, and the expected crop yields are displayed as the final output. Figure 4 displays the flow chart of the methodology adopted which is briefly detailed below.

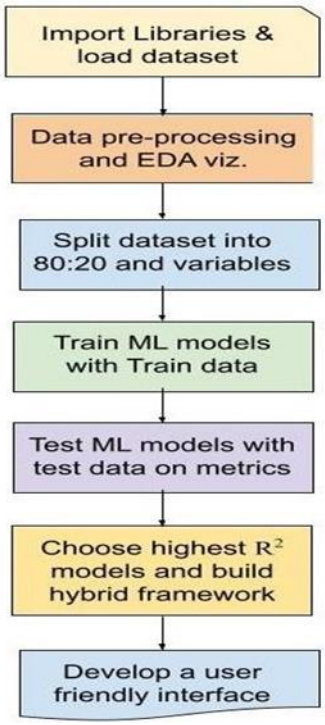

Fig 4. Flow-chart of the implementation

Data Collection and Loading. This entails importing all the needed libraries in Python and loading the dataset from kaggle.com. We merge two datasets acquired from http://www.fao.org/home/en/ and https://data.worldbank.org/. Feature description in the dataset is presented in Table 1. The collection offers information on top 10 crops that are cultivated in USA. There are features like rainfall, temperature and pesticides that are included to determine the yield of the crops. There are roughly 28K records in the collection.

Data Pre-processing. The collected data undergoes pre-processing to assure data quality and integrity. Exploratory data analysis (EDA) is done on the dataset to find and pick featured based on their relevance to the prediction goal. Furthermore, an outlier removal technique is conducted to find and eliminate any unusual data points that could negatively impact the model's performance.

Model Training and Testing. Here, the dataset is separated into a train set and a test set, as well as into independent and dependent variables, or X and Y. 80% is the training data and 20% is the test data set split. Next, all the five ML models are trained using the train dataset and then the algorithm performance is tested on the remaining test data on all metrics. Based on best of three model performance on coefficient of determination (R2) score, we construct a hybrid ML framework. It is thus an arithmetic average of the absolute errors $|ei| = |yi - xi|$, where yi is the

prediction and xi the true value. Both the MAE and RMSE can range from 0 to ∞. They are negatively oriented scores. Lower values are better.

4) **Mean Squared Error (MSE):** The MSE is calculated as the average of the squares of the errors, or the average squared difference between the estimated and real values.

If a vector of n predictions is generated from a sample of n data points on all variables, and Y is the vector of observed values of the variable being predicted, with Y being the predicted values (e.g., from a least-squares fit), then the predictor's within-sample MSE is calculated as Prediction and Evaluation. Once the hybrid ML models have been chosen and trained, they are ready to make

$$1\ MSE = n \sum(Y_i - Y_i)^2 \quad (10)$$

forecasts on new, unseen data. The input data is transmitted through all the trained models, and it generates projections for the related crop yield. Evaluation metrics are used to measure the performance of all the models and compare their predicted accuracy.

Output Display. The final part of the system design involves showing the results, which comprise the estimated crop yields generated by the hybrid ML model. These predictions can be displayed visually, such as in a graphical manner, or as a tabular output exhibiting the projected yields for each input instance. Finally, a user-friendly interface is constructed using the Gradio module in Python to view the estimated crop production as per the inputs submitted by any user.

3.3 Evaluation metrics

The proposed models are assessed using the following metrics.

Coefficient of determination (R2): When anticipating the result of a specific event, the coefficient of determination is a statistical measurement that analyzes how variations in one variable may be explained by differences in another one. In other words, this coefficient measures the strength of the linear relationship between two variables. This statistic is represented by a number between 0.0 and 1.0, with 1.0 signifying perfect correlation. As a result, it is a trustworthy model for forecasting the future.

Root Mean Square Error (RMSE): It is the residuals' standard deviation (prediction errors). Residuals are a measure of how far away data points are from the regression line; RMSE is a measure of how spread out these residuals are. It displays how concentrated the data is towards the line of best fit.

$$RMSE = \sqrt{(\overline{f-o})^2} \quad (8)$$

where, f= forecasts (anticipated values or unknown results), and o=observed values (known results).

Mean Average Error (MAE): The MAE estimates the average magnitude of the errors in a group of forecasts, without considering their direction. MAE is determined as the total of absolute errors divided by the sample size.

$$MAE = \frac{\sum_{i=1}^{n} |yi-xi|}{n} = \frac{\sum_{i=1}^{n} |ei|}{n} \quad (9)$$

4. RESULTS AND DISCUSSION

The results are split into three comparative studies, including exploratory data analysis, intra-model and inter-model comparisons as mentioned below.

4.1 Exploratory data analysis

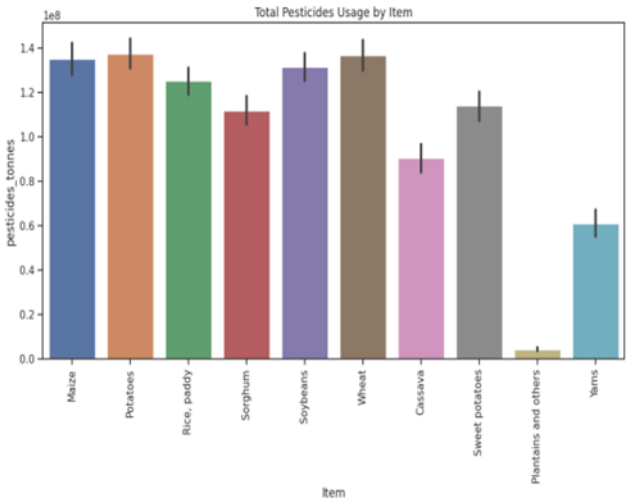

Fig 5. Total pesticide usage by crops

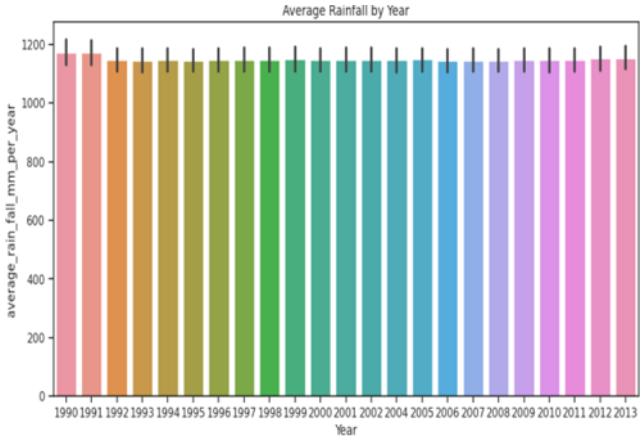

Fig 5. Average rainfall by year

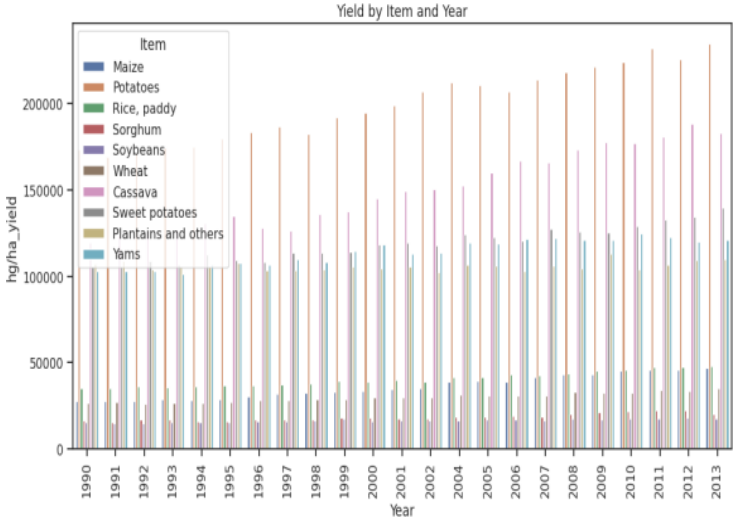

Fig 7. Yield by item and year

The results of EDA are first studied. Figure 5, Figure 6 and Figure 7 provide some significant insights into the dataset such total pesticide consumption by crops, average rainfall for the year and yield by item and year.

4.2 Intra-model comparison

Results of all the models implemented in a standalone mode are reported in Table 2. It is noticed that the R2 value of Random forest, Decision Tree and XGboost are greatest in comparison to SVM and Linear regression suggesting a perfect connection. This is owing to their capacity to capture complicated relationships, handle non-linearities, and address overfitting better compared to SVM and LR models. The SVM and LR models assume linear correlations between features and target variable yield. In the case of complicated interactions or non-linearities, they struggle to capture the underlying patterns in the data, leading to lower R2 scores. Hence, DT+XGBoost+RF models are used for implementing the hybrid ML model.

Hybrid Model

The first model in the hybrid framework is the Decision tree regressor, second model is Gradient boosting regressor and third model is the Random forest accordingly. We choose this sequence of models because of their functions and contribution to the overall model. Firstly, Decision trees give various essential contributions to the hybrid ML model as they offer interpretability, facilitate feature selection, capture nonlinear relationships, permit ensemble learning, and handle missing data and outliers. These properties enhance the accuracy, understanding, and usefulness of the model for crop production prediction. Secondly, Gradient Boosting Regression is a sophisticated technique which contributes by sequentially generating models, handling non-linear interactions, providing feature importance analysis, robustness to outliers, regularization to control overfitting and enabling some level of model interpretability. These contributions enhance the accuracy, reliability, and interpretability of the hybrid model. Random forest boosts the hybrid ML model by enhancing accuracy, robustness to noise and outliers, offering feature importance insights, managing high-dimensional data, enabling error estimation, and supporting parallel processing. These qualities help to more precise and dependable estimates for crop yields, assisting in agricultural decision-making and resource management.

For training and testing the hybrid ML model, we use specific parameters like n-estimators and max depth. We optimize the results by hyper tweaking certain settings. Table 3 illustrates the parameter values which yielded the best results.

Following the Decision Tree Regressor, the retrieved features and the dataset are transferred to the Gradient Boosting model. XGBoost is an implementation of Gradient Boosted decision trees. In this approach, decision trees are built in sequential manner. Weights play a vital part in XGBoost.

Weights are applied to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted erroneously by the tree is increased and these variables are then supplied to the second decision tree. These separate predictors then ensemble and for each candidate in the test set, it uses the class with the majority voting as the final prediction as shown in Figure 8. The result of the Gradient Boosting model is then input into the Random Forest regression model as depicted in Figure 9. Random Forest employs a decision tree as its basic classifier. An attribute split/evaluation measure is used in decision tree induction to identify the best split at each node of the decision tree. The generalisation error of a forest of tree classifiers is dictated by the strength of the individual trees in the forest as well as their correlation.

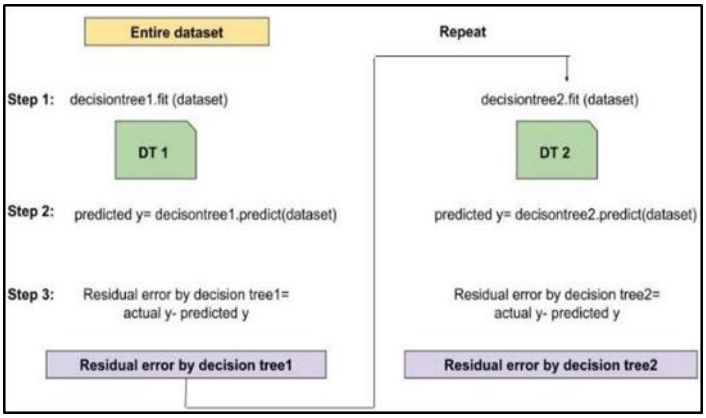

Fig 8. XGboost learning algorithm

First, a random forest is constructed by selecting one of the five split measures at a time. For example, random forest with information gain, random forest with gain ratio, etc. Following that, a unique hybrid decision tree model for random forest classifier is constructed. Individual decision trees in Random Forest are generated in this paradigm using various split measures. Weighted voting dependent on the strength of individual trees augments this paradigm. This hybrid model's assessment metrics and correctness are checked. Combining numerous decision trees enhances the accuracy and stability of forecasts.

From findings in Table 4, it is apparent that the hybrid ML model delivers the greatest R2 value of 0.9847 compared to all individual models. Since various factors like temperature, rainfall, soil composition, and agricultural practices interact in complex ways, the hybrid model is better equipped to handle these intricate relationships and spatiotemporal nonlinearities present in the data and uses them effectively in predicting the yield. Also, the ensemble nature of RF by merging several DTs, feature importance metrics, and XGBoost's boosting process adds to higher R2 score of the hybrid model.

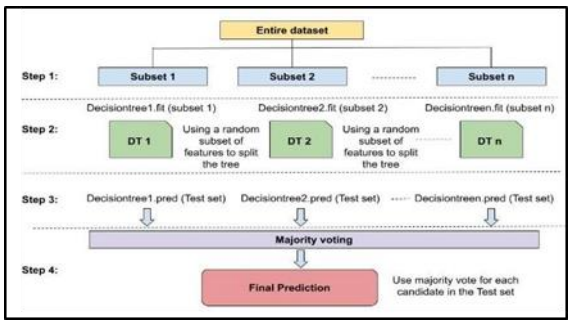

Fig 9. Hybrid decision tree model for random forest regression

Table 2. Intra-model experimental results

Metrics	Random Forest	Decision Tree	SVM	Linear Regression	XG Boost
R^2	**0.9837**	**0.9737**	-0.2040	0.08628	**0.9732**
RMSE	10645.42	13938.83	14083.62	94468.60	82295.73
MAE	3999.17	4191.63	7954.61	57669.24	62779.32
MSE	6772588013.10	194519546.09	198348525.23	8924316839.95	6772588013.10

Table 3. Hybrid model tuning parameters

Model	**N-Estimator**	**Max Depth**
Decision Tree	-	10
Gradient Boost	500	10
Random Forest	500	11

Figure 10 demonstrates the R2 scores comparison of all the techniques, including the hybrid ML model. On comparison of all other metrics, the proposed hybrid ML model returns the lowest MAE, MSE and RMSE values by leveraging the complementary strengths of different algorithms, reducing bias and variance, capturing complex relationships, and creating an ensemble effect that improves predictive accuracy. MAE scores provide information regarding the discrepancy between actual and projected values. A lower MAE value suggests a performance. By combining the correct base models and different hyperparameter settings can also help. Ensemble approaches like stacking can further optimize and improve accuracy. Applying regularization strategies to reduce overfitting in individual models by pruning decision trees or applying dropout in neural networks can aid enhance generalization and contribute to the hybrid model's accuracy.

Table 4. Hybrid model results

Metrics	**Hybrid Model**
R^2	0.9847
RMSE	9378.81
MAE	3829.6532
MSE	119182984

Computing the RMSE score helps measure the standard deviation of the residuals. The lower the MSE and RMSE values, the better the model for calculating returns.

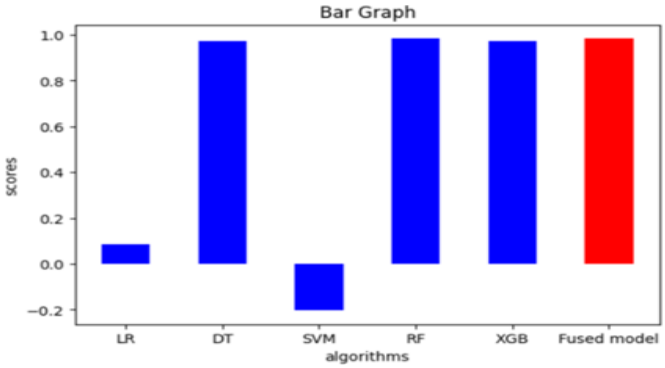

Fig 10. R2 scores comparison of all methods

4.3 Inter-model comparison

The suggested hybrid model is compared with current models using various methodologies on the same dataset. Table 5 displays the comparison in terms of model accuracy.

From inter model comparison results, the proposed hybrid model returns the maximum accuracy of 98.6% consequently proving that the DT+XGBoost+RF model surpasses other state-of- the-art models compared with 4.47 pp better accuracy than the next best model. There are various potential enhancements that might be made to the hybrid model to further enhance predicted accuracy. Feature engineering is one where quality and relevance of features may aid to enhance model.

Figure 11 and Figure 12 exhibit the Actual vs. Predicted yields for two crops, namely rice and wheat. The tight alignment between the true & anticipated values indicates the correctness & reliability of the suggested hybrid model.

On account of superior model performance, an Industry use- case produced is a tool named 'Crop Yield Predictor' implemented with an easy user friendly interface that can be used by farmers, policymakers and other stakeholders for informed decision making as illustrated in Figure 13.

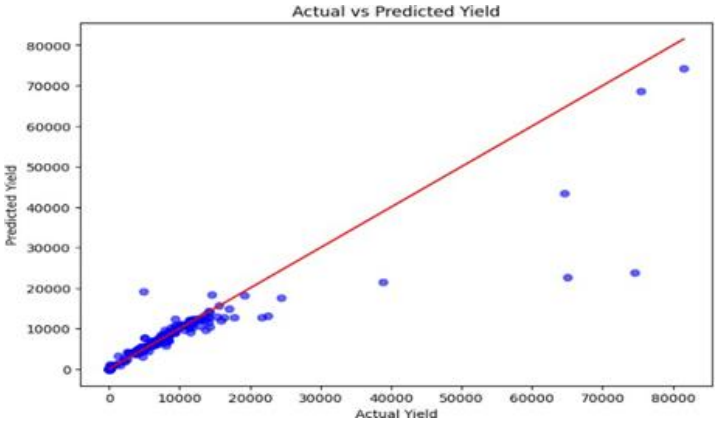

Fig 11. Actual vs. Predicted yield for rice

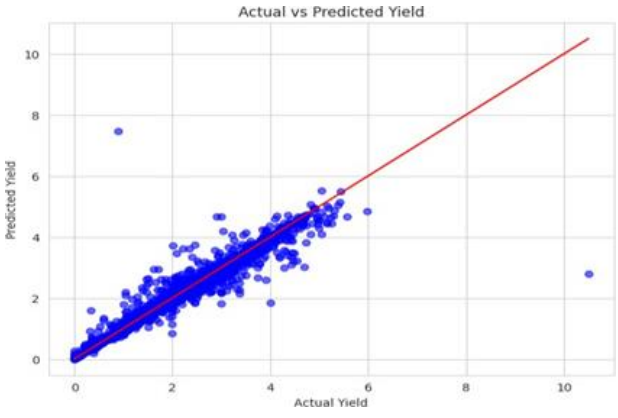

Fig 12: Actual vs. Predicted yield for wheat

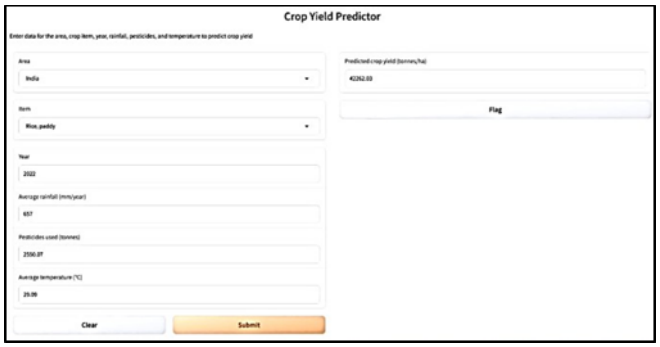

Fig 13. The 'Crop yield predictor' user interface created

Conclusions

A hybrid ML model is suggested for crop yiled prediction which incorporates the best three out of five ML models evaluated on their R2 scores, namely DT, XGBoost and RF. From results, the hybrid model exceeds the individual models applied with an R2 score of 0.9847 and all others metrics including RMSE, MAE and MSE. In intra-model comparison with current models, the suggested hybrid model surpasses them with an accuracy of 98.6%.

To enhance the accessibility of this model to policymakers and farmers at large, a user-friendly tool named 'crop yield predictor' is developed. Our findings contribute by proposing a novel technique to estimating crop production, expanding our understanding of hybrid modeling, providing insights into feature relevance, and addressing practical difficulties in agriculture and sustainability. The practical implications of the findings have considerable potential for improving day-to-day agricultural operations and directly influence numerous areas of agricultural decision-making for farmers and various stakeholders, benefiting productivity, sustainability, and economic outcomes for farmers. By turning complicated data and algorithms into useful insights, our discoveries bridge the gap between advanced machine learning techniques and practical on-ground applications in agriculture. As part of future study, we want to investigate how temporal dynamics affect predictions. Also, we want to incorporate crop disease, climate change and incorporate remote sensing data to gather spatial information about soil quality, vegetation health, and other elements that influence crop output.

References

[1] Rashid, M., Bari, B.S., Yusup, Y., Kamaruddin, M.A., Khan, N. (2021). A detailed evaluation of crop yield prediction using machine learning algorithms with special emphasis on palm oil production prediction. IEEE Access, 9: 63406-63439. https://doi.org/10.1109/ACCESS.2021.3075159

[2] Chlingaryan, A., Sukkarieh, S., Whelan, B. (2018). Machine learning algorithms for crop yield prediction and nitrogen status estimate in precision agriculture: A review. Computers and Electronics in Agriculture, 151: 61-69. https://doi.org/10.1016/j.compag.2018.05.012

[3] Basso, B., Liu, L. (2019). Seasonal agricultural yield forecast: Methods applications and accuracies. Advances in Agronomy, 154: 201-255.https://doi.org/10.1016/bs.agron.2018.11.002

[4] Shahhosseini, M., Martinez-Feria, R.A., Hu, G., Archontoulis, S.V. (2019). Maize yield and nitrate loss prediction with machine learning techniques. Environmental Research Letters, 14(12): 124026. https://doi.org/10.1088/1748-9326/ab5268

[5] Shahhosseini, M., Hu, G., Archontoulis, S.V. (2020). Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science, 11: 1120. https://doi.org/10.3389/fpls.2020.01120

[6] Khosla, E., Dharavath, R., Priya, R. (2020). Crop yield prediction utilizing aggregated rainfall-based modular artificial neural networks and support vector regression. Environmental Development and Sustainability, 22: 5687-5708. https://doi.org/10.1007/s10668-019-00445-x

[7] Van Klompenburg, T., Kassahun, A., Catal, C. (2020). Crop yield prediction using machine learning: A thorough literature review. Computers and Electronics in Agriculture, 177: 105709.

a. https://doi.org/10.1016/j.compag.2020.105709

[8] Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.S., Kim, S.H. (2016). Random forests for global and regional crop yield projections. PLoS ONE, 11(6): e0156571. https://doi.org/10.1371/journal.pone.0156571

[9] Walczak, S. (2016). Artificial neural networks and other AI applications for business management decision support. International Journal of Sociotechnology and Knowledge Development (IJSKD), 8(4): 1-20. https://doi.org/10.4018/IJSKD.2016100101

[10] Dahikar, S.S., Rode, S. (2014). Agricultural crop yield forecast utilizing artificial neural network technique. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, 2(1): 683-686.

[11] Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R. L., Mouazen, A.M. (2016). Wheat yield prediction utilizing machine learning and sophisticated sensing techniques. Computers and Electronics in Agriculture, 121: 57-65. https://doi.org/10.1016/j.compag.2015.11.018

[12] Rehman, T.U., Mahmud, M. S., Chang, Y.K., Jin, J., Shin, J. (2019). Current and future applications of statistical machine learning techniques for agricultural machine vision systems. Computers and Electronics in Agriculture, 156: 585-605. https://doi.org/10.1016/j.compag.2018.12.006

[13] Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y., Srinivasan, K. (2018). Forecasting yield by merging agrarian elements and machine learning models: A survey. Computers and Electronics in Agriculture, 155: 257-282. https://doi.org/10.1016/j.compag.2018.10.024

[14] Balakrishnan, N., Muthukumarasamy, G. (2016). Crop production-ensemble machine learning model for prediction. International Journal of Computer Science and Software Engineering, 5(7): 148-153.

[15] Medar, R., Rajpurohit, V.S., Shweta, S. (2019). Crop yield prediction using machine learning techniques. In Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, pp. 1-5. https://doi.org/10.1109/I2CT45611.2019.9033611

[16] Gandhi, N., Armstrong, L.J., Petkar, O., Tripathy, A.K. (2016). Rice crop yield prediction in India utilizing supportvector machines. In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, pp. 1-5. https://doi.org/10.1109/JCSSE.2016.7748856

[17] Keerthana, M., Meghana, K.J.M., Pravallika, S., Kavitha, M. (2021). An ensemble algorithm for crop yield prediction. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, pp. 963- 970. https://doi.org/10.1109/ICICV50876.2021.9388479

[18] Suganya, M. (2020). Crop yield prediction using supervised learning techniques. International Journal of Computer Engineering and Technology, 11(2): 9-20.

[19] Chlingaryan, A., Sukkarieh, S., Whelan, B. (2018). Machine learning algorithms for crop yield prediction and nitrogen status estimate in precision agriculture: A review. Computers and Electronics in Agriculture, 151: 61-69. https://doi.org/10.1016/j.compag.2018.05.012

[20] Freedman, D.A. (2009). Statistical Models: Theory and Practice. Cambridge University Press.

[21] Kavitha, S., Varuna, S., Ramya, R. (2016). A comparison analysis on linear regression and support vector regression. In Proceedings of IEEE International Conference on Green Engineering and Technologies (IC- GET).

[22] Huber, L.A., Xu, Q.B., Jürgens, G., Böck, G., Bühler, E., Gey, K.F., Schönitzer, D., Traill, K.N., Wick, G. (1991). Correlation of lymphocyte lipid composition membrane microviscosity and mitogen responsiveness in the aged. European Journal of Immunology, 21(11): 2761-2765. https://doi.org/10.1002/eji.1830211117

[23] Agarwal, S., Tarar, S. (2021). A hybrid technique for crop yield prediction utilizing machine learning and deep learning algorithms. Journal of Physics: Conference Series, 1714(1): 012012. https://doi.org/10.1088/1742- 6596/1714/1/012012

[24] Fayaz, S.A., Kaul, N., Kaul, S., Zaman, M., Baskhi, W.J. (2023). How machine learning is revolutionizing agricultural sciences: An technique to predict apple crop yield of Kashmir province. Revue d'Intelligence Artificielle, 37(2): 501-507.

[25] Elavarasan, D., Vincent, P.M.D. (2020). Crop yield prediction utilizing deep reinforcement learning model for sustainable agrarian applications. IEEE Access, 8: 86886-86901. https://doi.org/10.1109/ACCESS.2020.2992480

[26] Pandith, V., Kour, H., Singh, S., Manhas, J., Sharma, V. (2020). Performance study of machine learning algorithms for mustard crop production prediction from soil data. Journal of scientific research, 64(2): 394-398. https://doi.org/10.37398/JSR.2020.640254.