A Machine Intelligence based Approach for Analysis and Prediction of Rice Yield in Semi-Arid Regions in India

Ratnakirti Roy[1], Ramesh Adiga K.[2]

[1, 2]Department of MCA, Acharya Institute of Technology, Bangalore, India.

E-mail: [1]ratnakirtiroy@ieee.org, [2]km.23.mcav@acharya.ac.in

Abstract

With global food security under increasing threats from population expansion and climatic uncertainty, innovative techniques are key to revolutionizing agriculture. The current project employs the Random Forest method to make sound predictions of rice crop yields and environmental conditions under which rice can grow based on a historical record of agricultural and the environmental data from 1986 to2017. There are two primary objectives in this study. The first is to forecast rice yield, from cultivated area and production statistics, to assist in planning and resource allocation. The second is to forecast the soil nutrient levels Nitrogen, Phosphorus, Potassium, and pH and analyze environmental factors like humidity, rain, rainfall, and temperature that play a significant role in crop growth and soil health. Overall, the Random Forest model performed well across a range of evaluation metrics, demonstrating a clear ability to capture sophisticated relationships s within the data. In both instances, by setting target yield levels identifying the best soil and climate conditions, this research has the potential to offer relatively straightforward-to-implement data-based advice for maximizing agricultural yield, enabling farmers to cope with the pressures of an evolving environment. The research highlights key trends, including ideal humidity, rainfall, and temperature ranges required to achieve greater yields and preserve nutrient-laden soils. Ultimately, this research merges advanced machine learning with actual agricultural needs, creating a cost-effective and sustainable means to secure food systems and adopt climate-resilient farming practices.

Keywords: Random Forest, rice yield prediction, crop yield optimization, machine learning in agriculture, Precision Agriculture, Climate Data, Machine Learning.

1. Introduction

Farmers would be able to forecast the harvest and maintain crops to address climate change and the loss of arable land. Global food systems are facing this issue, where increased population, climate change, and the destruction of arable land threaten food systems for billions of individuals. Rice is a source of food for half of the world's population, and it is crucial that farmers be empowered to enhance their harvests while simultaneously coping with environmental changes. This is where machine learning (ML) as a decision-making mechanism for farmers comes in, improving their decision-making using data. In this article, we utilized the capabilities of the Random Forest algorithm to build rice crop yield models and determine the best nutrient levels (Nitrogen, Phosphorus, Potassium) and pH level as a function of environmental and weather conditions (Humidity, Rainfall, and Temperature). We used historical records from 1986 to 2017 and concentrated on two main outcomes: (1) to precisely forecast rice yields as a function of area and production information; and (2) to determine the best soil conditions to promote healthy crop growth. We compared the algorithms' performance based on Mean Absolute Error (MAE), R-squared (R^2), and Precision [12]. We seek to present evidence that the model offers interpretable and stable estimates. This contribution is a first step towards precision agriculture, embedding cutting-edge ML with agricultural, real-world solutions to enable farmers to be assured that they can consistently produce food in situations of uncertainty and to enhance productivity and resilience to ensure a more sustainable future for food security.

2. Related Work

ML has transformed agricultural research through the ability to make virtually accurate predictions of crop yields, thereby addressing developing global food security concerns that are becoming more urgent with climate uncertainty and continuous population expansion. In the review [2], the authors touch on theoretical and applied ML techniques employed in forecasting crop yields using palm oil but offer extensive ML methods, such as regression techniques, decision trees, and ensemble techniques, applicable to rice. These techniques make use of historical data and/or environmental data to estimate complex relationships that serve to explain yield forecasting. Along with this, [4] considers both regression and deep learning techniques and demonstrates the capability of both model types to understand and analyze high-dimensional datasets in agriculture using various factors such as soil nutrients, weather, and crop management practices in relation to rice yield determination.

Several studies highlight the strengths of Random Forest (RF) and other ML techniques in their application to crop yield prediction. Reviews [3, 7] demonstrate how RF can effectively model rice yield considering non-linearity and different types of data (e.g., weather and soil measures). These reviews propose multiple advantages of employing RF in crop yield modeling, such as its proven capacity to handle large datasets without heavy overfitting, a strong point for current agricultural applications. In particular, [6] explains ML techniques to enhance rice yield estimation, such as hyperparameter-tuned weighted regularized extreme learning machine and hyperparameter-tuned stacked autoencoders, which consider temporal and environmental information and produce high accuracy in predicting yields. Generally, the papers encompassed in [6] suggest that the more the model parameters were optimized (hyperparameters), the higher the predictive accuracy.

IoT-integrated smart agriculture systems, along with some level of ML [10], can enhance the decision-making of rice farmers through real-time data on soil moisture, temperature, and nutrient levels, thereby enabling the prediction of rice yield with high accuracy. There have been studies specific to rice [11, 12, 13, 16] that utilized various ML models to predict rice yield, employing models such as vision transformers, support vector machines, and regression techniques, all of which have predicted yield with a high degree of accuracy. In general, these studies employed datasets that included a mixture of past yield data, agronomic human expertise, weather conditions, and soil types; they were specifically interested in our own research and randomly predicting rice yield with Random Forest. For instance, [12] compares various ML models to determine the best models for rice yield prediction in particular regions, highlighting the importance of regional data.

Besides yield forecasting, applications of ML in agriculture are transferable to related problems. For example, [14] applied ML to forecast diseases in rice crops with characteristics, including leaf color and leaf texture, which influence yield models since they are related to conditions impacting rice crop productivity. [17] formulates a bivariate joint model of rice yields and methane emissions with different environmental and agri-data features, which is applicable to sustainability. Sustainability is highly crucial in rice production since it has adverse impacts on the environment. Other research [15, 20] highlights the extensive use of ML algorithms (e.g., RF) for numerous crops and datasets, further demonstrating their usability in predicting rice crop yield.

3. Proposed Work

The suggested system illustrates a methodological pipeline to forecast rice yields and determine maximum soil conditions using the Random Forest algorithm with a comprehensive historical database from 1986 to 2017 [2, 12]. The dataset dimensions have proven Random Forest to predict more accurately in contrast with other traditional methods such as Logistic Regression and SVM. The approach has engaged a sequence of processes, including such as data retrieval, exploratory analysis, pre-processing, model learning, assessment, and comparison to provide meaningful rice yield forecasting. The approach is indicative of trustful and proven machine learning practices with a capture of intricate and non-linear relationships between agricultural and climatic variables, which comply with precision agriculture standards [1]. In general, we constructed the pipeline for robustness and valid usage in real-world applications towards farm resilience under the constraints of variability in climate and best supply of soil nutrients.

3.1 Data Understanding

The data set used in this study is crucial, as it carries significant historical agricultural and environmental information necessary for rice yield prediction and soil condition estimation for rice cultivation [2]. The dataset contains observations of rice yield (tons per hectare), land area (hectares), total production (tons), environmental measurements such as humidity (percentage), rainfall amount (millimetres), temperature (degrees Celsius), and soil data involving Nitrogen (N), Phosphorus (P), Potassium (K) (kg / hectare), and pH. The dataset recorded past observations between 1986 and 2017, an aggregate of 31 years of seasonal and climatic fluctuations in various rice-producing areas in the basins 12. According to data stored in an ICRISAT repository, the dataset contains numerical values for all variables measured and some metadata concerning soil classification and irrigation practice if available. Please note that the information from a variety of instruments and other surveys was fused as defined in [1], to make the dataset richer, and permit machine learning to provide more integrity. This rich dataset not only advanced the two objectives for predicting rice yield and soil nutrient levels and pH, but it also provides a go-to reference point for using machine learning.

3.2 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) herein aimed to apply nine varied visualization methods to discover patterns and relationships in the historical dataset (1986 to 2017) utilized in rice yield prediction and estimation of soil nutrient levels as guided in depth in [3]. Histograms and kernel density estimation plots revealed a right skewed distribution of the rice yield, area and production variables, which may reflect the nature of smaller farm sizes. The kernel density estimation (KDE) plots also revealed peaks at moderate levels for humidity and rainfall, less than 80 and 1500 respectively, which reflects normal environmental conditions for a rice crop growth. The scatter plots also implied a nonlinear correlation of rice yield with environmental factors like humidity and rainfall. Feature selection for the tree-based models and the Random Forest model started with feature importance analysis for the initial models that implied that humidity, rainfall, temperature, and nitrogen were the first level predictors [7]. The cyclical trends of yield that we observed through temporal analysis informed by seasonality cues gave directions for feature engineering specifically. The same cyclical temporal analysis informed our preprocessing and modeling by validating what variables could possibly have an effect on yield and what variables could be utilized to generate a reliable prediction.

3.3 Pre-Processing

Preprocessing was necessary to guarantee the quality and uniformity of the data that we sought to employ for modeling [10]. As it was, there were around 5% missing records in total (excluding those records that were found irrelevant for yield), with missing records appearing more frequently in environmental variables like: rainfall and humidity. These were imputed with median values to retain the datasets as a single set and not render the set biased. Outliers were also identified during exploratory data analysis, which were capped to reduce their effect on the model's performance using the interquartile range (IQR). Any values that were discovered to be greater than 1.5 times the IQR from the lower and upper quartiles of data were capped with intelligent values. For example, if one data value was a boundary value for yield or rainfall but happened only from sporadic weather occurrences, we corrected it so it would not take away from the other values estimating yield. Feature scaling was performed for all environmental and soil features with StandardScaler and was performed to achieve a mean of 0 and variance of 1 with StandardScaler so that they all had the same contribution and degree of influence towards our Random Forest model [12].

3.4 Model Training & Testing

According to the modeling methodology we employed, Random Forest (RF) was outlined in [3, 7], and was chosen due to its ability to model non-linear relations, high-dimensional data, and the intricate interactions between multiple agricultural, and environmental variables. The robust, ensemble nature of RF, combined with SVM, enables the RF method to counteract over-fitting, and therefore made it the obvious choice for the purposes of this related research.

3.4.1 Random Forest

The random forest model is strong and can incorporate all the features like 'area', 'year', 'N', 'P', 'K', 'PH', 'temperature', 'humidity', and 'rainfall' to predict 'yield' and all possible combinations of features. Merging the predictions from all the models will provide the best estimate and precision for yield. [12] The data set was divided into a training (80%) set and a test (20%) set, and there was a stratification function to ensure the same distribution of yield values. Random Forest's bagging and random selection of features from the overall number of features enhance the strength of the model. For the test set, prediction accuracy was determined by comparing the predicted values with actual values as a way of estimating predictive accuracy and generalizing to real agricultural systems in any relevant form.

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

$\hat{f}(x)$: is the predicted value obtained by averaging the predictions of all decision trees in the forest.

T_b (x): is the prediction of the bth decision tree

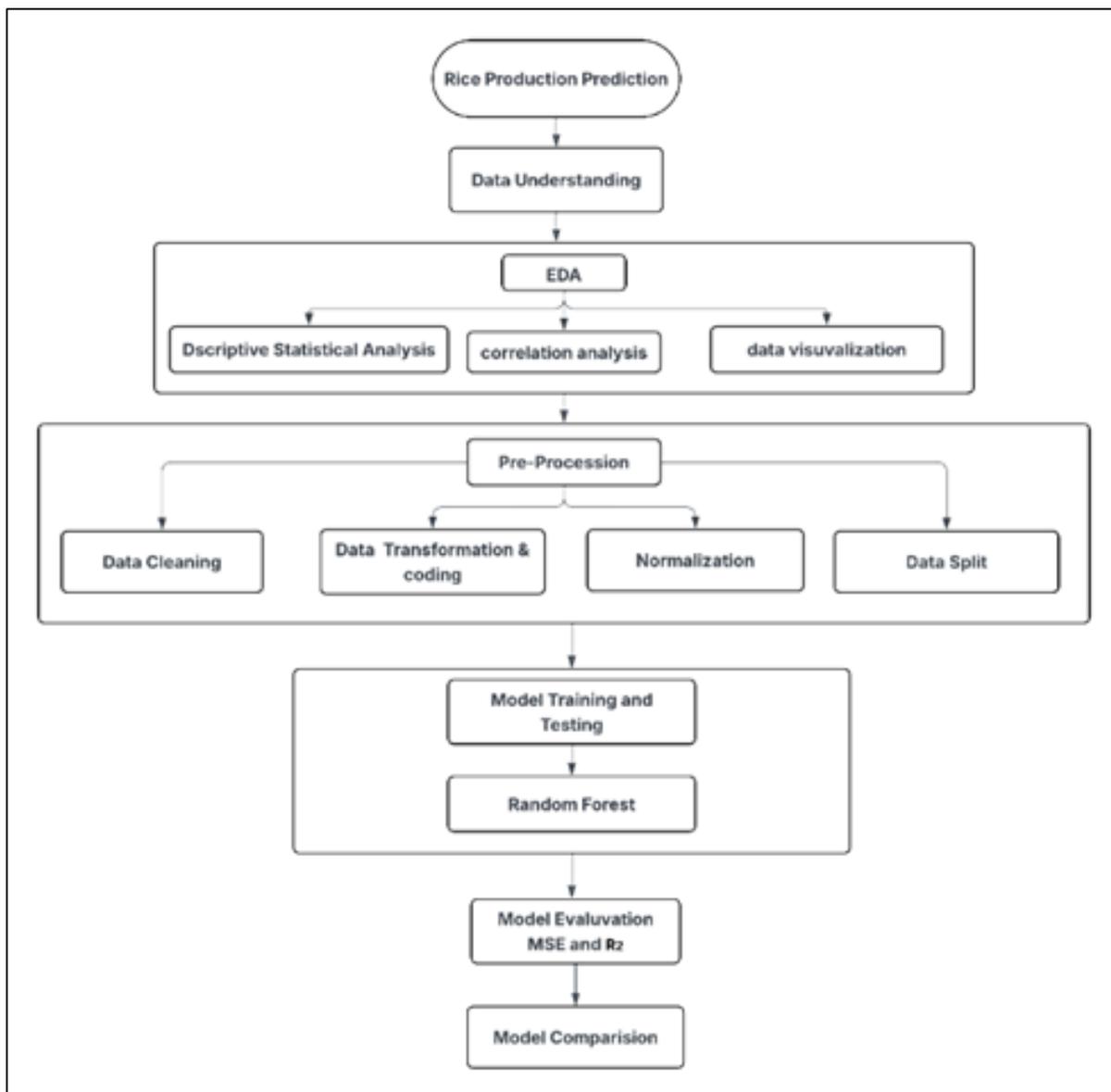B: is the total number of trees in the Random Forest

Figure 1. Proposed Analysis and Prediction Model

3.5 Model Evaluation

The model was tested on a number of metrics to verify its validity and reliability under the assessment methods discussed above. The R^2 Score, which describesd the proportion of variance in the independent variables (yield and nutrient levels) explained by the model and can be used to approximate its explanatory power, was one such metric. We also employed the notion of Mean Absolute Error (MAE) to measure the average absolute differences that came from model predictions with respect to the actual values (it is supposed to measure the level of precision of the numerical predictions). We also employed MSE, which displays the average

of squared differences between predicted values and the actual values (it can also be used for categorical outputs). This is particularly relevant as it puts more weight on greater penalties for larger errors and makes apparent instances of the model's poor performance for robustness of performance evaluation. For cases where (potentially categorical) predictions were needed (e.g. predicting yields as high, medium, or low), we were also able to evaluate Precision, or the accuracy of the model when presented with a classification task in order to inform farmers' decisions [12].

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

y_i: Actual value, $\hat{y}_i$: Predicted value, $\bar{y}$: Mean of actual values, n: Number of data points, R^2: Proportion of variance in the dependent variable that is predictable from the independent variable.

$$\text{Mean Absolute Error(MAE)} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

n: Total number of data points, y_i: Actual value, $\hat{y}_i$: Predicted value for the ith sample, $|y_i - \hat{y}_i|$: Absolute error for the ith prediction, MAE: The average of these absolute errors over all predictions.

$$\text{Mean Squared Error(MSE)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

n: is the number of observations, y_i: is the actual value, $\hat{y}_i$: is the predicted value

3.6 Model Comparison

In order to determine the appropriateness of Random Forest, its performance was contrasted with other machine learning models to ensure an appropriate comparison for rice yield prediction and estimation of soil condition [6,12,16,20]. For the purposes of resource constrained agricultural systems, this data is valuable to farmers for taking corrective measures on a given field, such as adjusting the quantity of irrigation water or altering the application of fertilizers. The limitation of Random Forest for very large datasets is typically computational intensity, however this was remedied through hyper parameters optimization [16], Deep learning models, such as those researched in [5, 20].

4. Results and Discussion

4.1 Dataset

ICRISAT data (http://data.icrisat.org/dld/src/about-dld.html) from 1986 to 2017 existed and provided a strong foundation to model rice yield and soil health in various rice-growing zones [2, 16]. It was in CSV format and included rice yield (tonnes/hectare), planted area (hectares), total yield (tonnes), humidity (%), rainfall (millimeters), temperature (degrees Celsius), and environmental conditions (soil nitrogen, phosphorus, potassium (kg/hectare), and pH) [16]. It contained agricultural and climatic data for 31 years and experienced a broad spectrum of climatic and seasonal impacts, including monsoons, droughts, and temperature variations, which contributed to the robustness of the predictions [2]. The Exploratory Data Analysis revealed strong trends and patterns in the data and facilitated the modeling process by providing qualitative insights into the plots and statistical summaries [3, 7].

4.2 Exploratory Data Analysis

Histograms and KDE plots were created for 1978 to 2017 to describe the distributions of variables of interest. The distribution of rice by area was found to be right-skewed in the histograms as well as in the KDE plots, indicating that the majority of the farms in the dataset had small areas, with an extremely smooth peak observed in the KDE curve, irrespective of the very large areas recorded in the histogram. Rice agriculture also shifted to the right, with the KDE curve showing that small-scale farmers constituted the most significant portion of rice producers. The distribution of temperature was nearly normal, with the KDE curve indicating central tendency around nearly normal tropical weather, which is ideal for rice farming. Rainfall and humidity distributions were also right-skewed, with KDE plots peaking at moderate values, consistent with the typical environmental conditions for rice-growing areas [9].
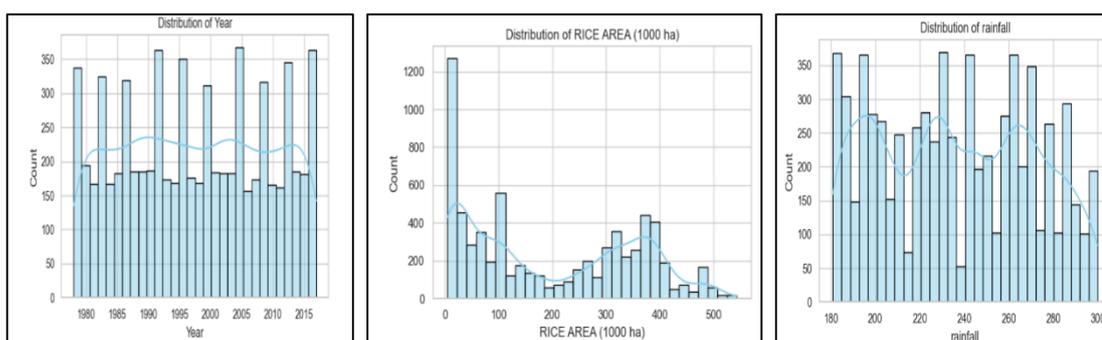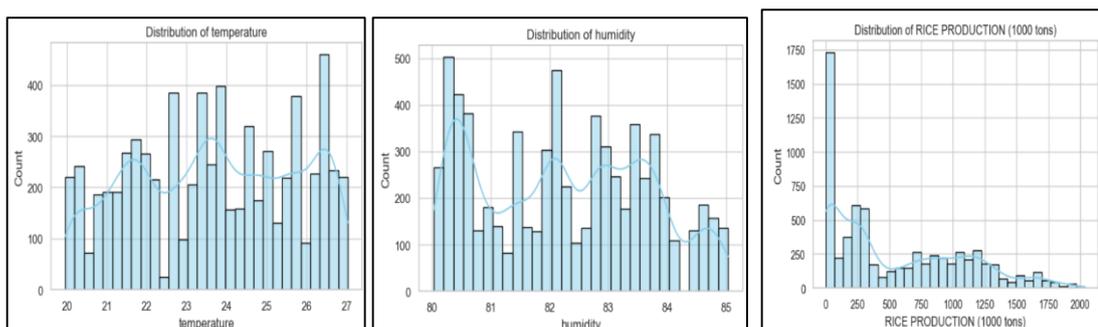
Figure 2. Rice Dataset: Distribution of Year, Area, Production, and Soil Nutrients

Similarly, in Table 1, from 1978 to 2017, observations based on the evidence itself indicate discernible changes in rice farming practices based on environmental as well as agronomic conditions. Rice production and yield varied geographically, while harvested acreages showed some intensive and small-scale farming practices. Temperature was, by and large, not extreme, whereas humidity and rainfall patterns were also, by and large, uniform, which are agro-climatically stable environments favorable to rice agronomy. Soil nutrients such as nitrogen, phosphorus, potassium, and pH all fell between comparable numbers over the years, providing rice farmers with good conditions to grow under that were, by and large, constant. This consistency is a solid basis on which to build predictive modeling and farm planning in the long run.

Table 1. Statistical Summary of the Dataset (1978-2017)

	Rice Yield	Production	Year	Rice Area	Temperature	Humidity	Rainfall	N	P	K	ph
Count	6922	6922	6922	6922	6922	6922	6922	6922	6922	6922	6922
Mean	2427.3777	560.7493	1997.548	196.2092	23.67142	82.26	236.38	79.82	47.49	39.86	6.42
Std	1063.4082	519.4002	11.55853	155.8593	2.01340 4	1.409	34.15	11.84	7.866	2.935	0.76
Min	197.13	1.7	1978	4.9	19.96175	80.04	180.88	59	34	35	4.97
25%	1650.37	80.01	1988	41.8	21.89185	80.91	204.72	69	41	38	5.86

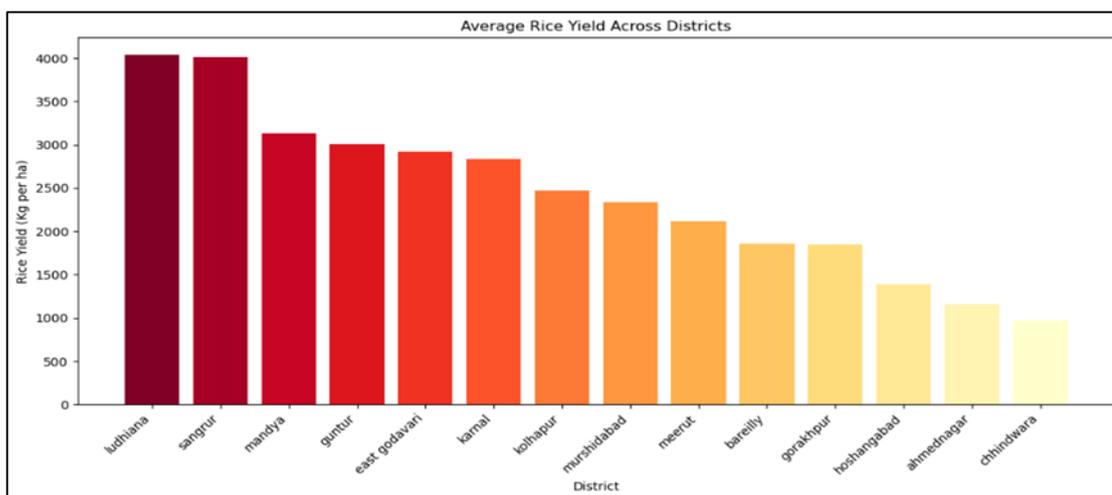| 50% | 2405.06 | 337.8 | 1997 | 162 | 23.69943 | 82.19 | 233.26 | 79 | 47 | 40 | 6.36 |
| 75% | 3135.81 | 977.23 | 2007 | 346 | 25.4906 | 83.46 | 264.56 | 91 | 54 | 42 | 7.04 |
| max | 5047.54 | 2038.6 | 2017 | 543.8 | 27.03149 | 85.04 | 300.01 | 100 | 61 | 45 | 7.92 |

Figure 3. Average Rice Yield Across Districts

Figure 3 depicts the average rice yield by district; it can be used to visualize the geographical spread of yield distribution. It also helps distinguish areas that experience continuous high (or low) productivity. In predicting farm yields using machine-learning methods, such spatial representation of analysis enables one to identify places with geographical variation (farmers experience varying environmental and agronomic conditions). Locating districts with high yield enables the Random Forest model to learn from the environmental conditions associated with high yield. Low-yield fields can be indicative of adverse environmental factors or circumstances of environmental failure, such as limited nutrients or unfavorable climate. Having the capability to geographically survey the fields of response helps improve the model and make it more sensitive to certain fields that need farm direction on a location-based level.
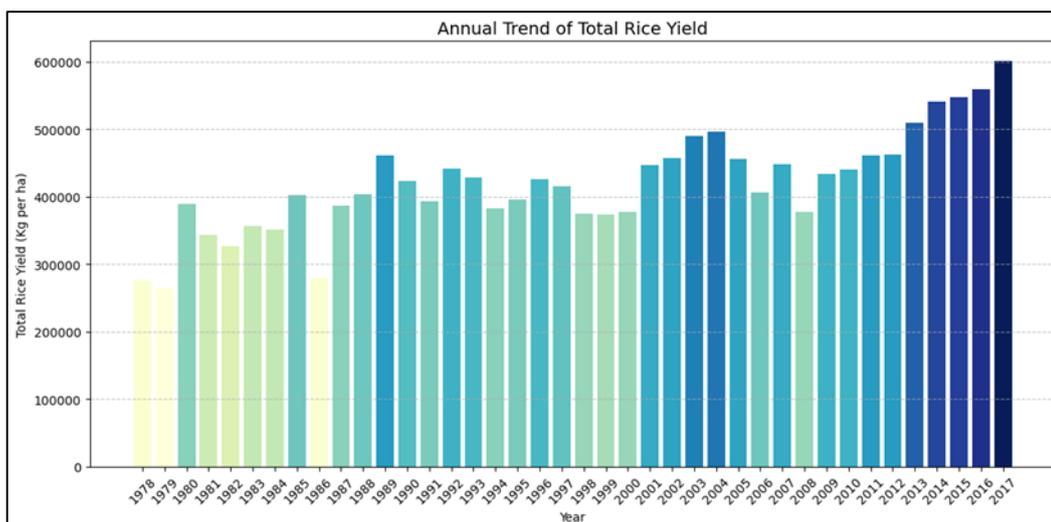
Figure 4. Annual Trend of Total Rice Yield

Figure 4 graphs the long-run trend for overall rice yield by year from the data set. The graph, being a time function, is highly useful in establishing long-run trends and changes in yield, possibly due to climate, farm policy, or technological change. Up or down sloping lines on the graph provide excellent context for the model because we're learning from the past. Trends over a period of time, i.e., whether there is activity at certain times or recurring activity (or patterns), will enable the machine learning model to forecast next year's rice crop estimates. Above all, though, the trend is an indicator of how accurately this model will forecast in the future and how accurately the model is adjusting for changing climatic trends and agricultural practices. Figure 5 presents a yearly mean of N, P, K, and pH, logging years of nutrient concentration and acidity.

Annual patterns in NPK and pH are therefore indicators of the extent to which nutrient management practices may have impacted yield. Some soil nutrient inclinations provide information regarding dominant fertility factors and yields. These factors, as input variables employed by the Random Forest model, enhance the model's potential to make recommendations regarding nutrient levels coupled with high-yielding outputs, thereby becoming highly appropriate for precision agriculture.
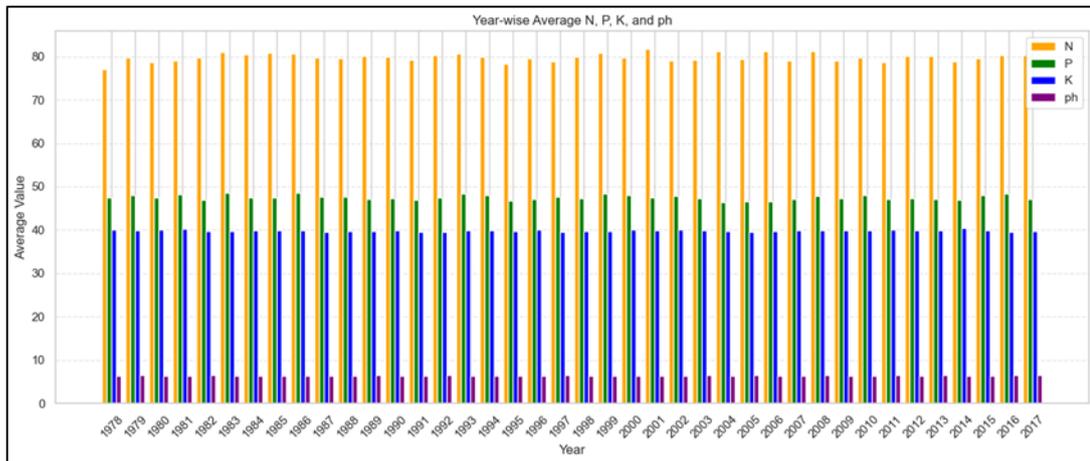
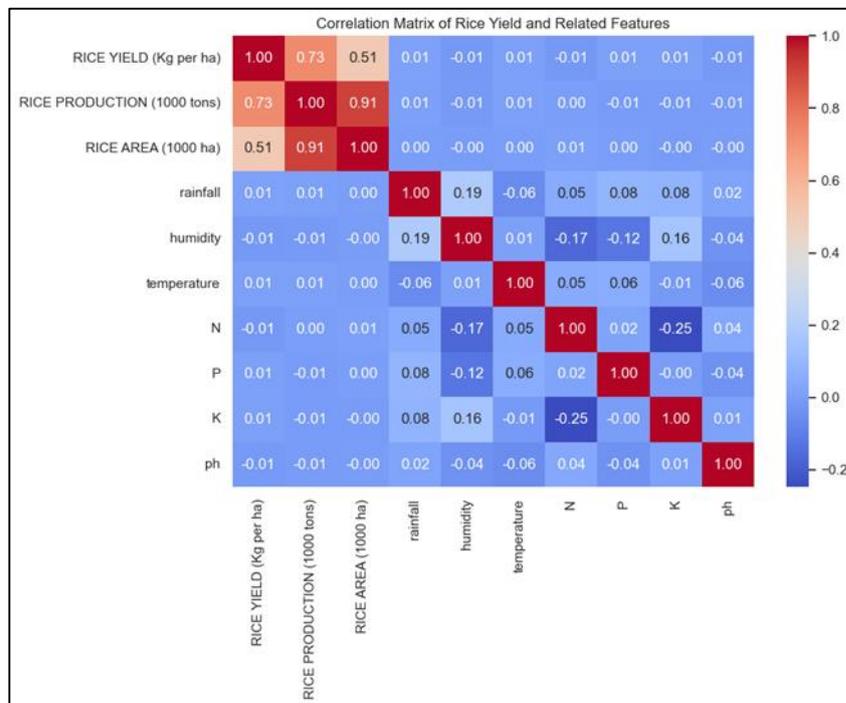**Figure 5.** Yearwise Average N, P, K, and pH

Figure 6. Heatmap of a Correlation Matrix

Heatmap of Correlation Matrix, as it is evident from Figure 6, is a correlation matrix plot of every variable in the dataset. Correlation is needed to identify highly correlated variables with yield. The humidity, rain, and nitrogen variables are likely to have a high positive correlation with yield. Such correlations will therefore inform model training and predictor selection so that the model gives more weight to the most predictive predictors. From a machine learning perspective, having such a correlation provides us with a greater ability to restrict

overfitting, ensure more generalizability to new data, and thus result in more accuracy in the prediction of yield.

4.3 Pre-Processing

Smoothened preprocessed data into a pure, clean, and ready-to-model raw dataset with complete sureness [10]. Missing values, which were partially (only rain and humidity) affected, were resolved by replacing missing variables by estimating the median in order to conserve observations, by looking at all the data and not introducing bias to the dataset. Outliers (an exceptionally high yield or precipitation measurement due to a very rare exceptional occurrence, though improbable) were limited to range values to improve the model by setting the boundary as the interquartile range (IQR) and proceeding further to try and adjust the outlier values in a way that would enhance prediction accuracy. Other soil and environmental parameters were also normalized using Standard Scaler before training the Random Forest model [13] so that no feature dominated the others. The train and test utilized a Random Forest model to predict rice yield and accessible soil nutrients (Nitrogen, Phosphorus, Potassium) and pH based on area, production, humidity, rainfall, and temperature features [3, 7].

4.4 Model Training & Testing

Data were split into a training set and a test set, with 80% for training and 20% for testing, using a stratified split on yield values. The test set evaluated the predictive accuracy by comparing the model's predictions to known/predicted yield values and nutrient values in the test set. The test set also confirmed the validity, as well as the ability of the models to generalize; hence, enabling it to be used in real-time (i.e., determining how much to use and when to sow based on a set of conditions) [6]. The Random Forest model was tested with respect to the extent to which it estimated the rice yield and the properties of Nitrogen, Phosphorus, Potassium, and soil pH.

4.5 Model Evaluation

Future parameters are of extremely significant use in crop yield estimation and farm practice management development. For the model performance evaluation under performance measures, some performance measures were suggested. Accuracy was left to the classes, or the ratio in which the model exactly predicted the data points, i.e., the classes of yields [12].

Accuracy added another aspect of goodness to classification, i.e., out of how many of the positive predictions (i.e, number of classes) they were correct, and it resulted in reducing the false positives [12]. In forecasting quantitative values, Mean Squared Error (MSE) was used to specify the mean square of error; the lower the MSE value, the greater the likelihood of prediction [11]. On the other hand, Mean Absolute Error (MAE) was the average prediction error in units and thus straightforwardly interpretable and had a point of standard deviation from which the predictions were far from observed facts [11]. Finally, the R² Score (Coefficient of Determination) gauged the fit of the model to observation data as it rose to indicate that the model had fitted more variance in the data [11, 13]. All the predicted predictions i.e rice yield predicted, quantity of nutrients found, and pH of the soil predicted are inputted into the Table: Evaluation Metrics for Rice Yield and Soil Nutrient Predictions Using Random Forest, and are measures of how consistent and reliable the model is by variable. All such values combined gave collective data regarding the precision, reliability, and appropriateness of the model for implementation in real agriculture [6, 11, 12, 13]. In overall-system-level performance analysis as proposed, results are compared against results of other methods reported in the literature.

Table 2. Evaluation Metrics for Rice Yield and Soil Nutrient Predictions Using Random Forest

Predicted Parameters	Accuracy	Precision	MSE	MAE	R^2
Rice Yield	0.9833	0.9825	0.6882	0.1746	1.0000
Nitrogen(N)	0.9986	0.9964	0.6776	0.1683	0.9953
Phosphorus(P)	0.9957	0.9906	0.6949	0.1102	0.9887
Potassium(K)	1.0000	0.9978	0.0359	0.0318	0.9959
Soil pH value	0.9986	0.9986	0.0022	0.0188	0.9960

4.6 Comparative Analysis

Since the data utilized in the study, Random Forest provides comparatively improved prediction accuracy and a balance of the performance measures. This is because the Random Forest model can be employed for predictive modeling of rice crop farm parameters in agricultural practice. Comparative results are given in Table 3.

Table 3. Comparative Analysis of the Proposed Method vs. Other Methods.

Algorithm	Accuracy (%)	Sensitivity	Specificity	F1 Score	MCC
SVM	80–85	0.78	0.80	0.79	0.68
Multilinear Regression	82–87	0.80	0.82	0.81	0.65
Random Forest	90-98	0.9	0.87	0.9	0.81
Gradient Boosting	90–93	0.83	0.89	0.88	0.77
Deep Learning	85–90	0.83	0.85	0.84	0.68
SCA-WRELM	87–91	0.85	0.85	0.87	0.74

5. Conclusion and Future Work

From this, we have demonstrated how machine learning can be utilized to predict rice yield and topsoil nutrient content using our example of a Random Forest model. From centuries of agricultural and environmental records, accurate predictions were generated, in addition to the determination of coupled patterns between crop yield and drivers over a broad range of temperature, moisture, rain, and soil content. The study not only validated that our model was accurate but also realistic in the context of actual day-to-day farm-level decision-making under regular agricultural production. Farmers can make appropriate decisions regarding the utilization of fertilizers, planting seasons, and climate variability resilience using the application of these forecasts, with a view to achieving optimal crop outputs and sustainability. Secondly, space for practical estimation of the quantities of soil nutrients with respect to the nature of the soils was required in an attempt to initiate productive exploitation of the soils, which would not only make farmers effective in reducing wastage but also add significantly to the environment. Last but not least, in the future, the second part of the research will be dedicated to the expansion of the application and use of the model. Even enhanced performance can be achieved if live satellite and IoT sensor data are utilized together in real-time, as this will enable maximum prediction accuracy and dynamic recommendations spatially. We also recommend utilizing even more advanced models like XGBoost, deep learning network models, and ensemble hybrid models for enhanced performance. Finally, situating this system

as a widely shared web or mobile application places farmers in action-connected real-time data they need, bridging the gap between high-tech AI technology and the farming practices they enable. This project is the culmination of advances in precision agriculture, where smart data-driven technology can be part of the solution to global food security, climate resilience, and agricultural sustainability.

References

[1] Reyana, A., Sandeep Kautish, PM Sharan Karthik, Ibrahim Ahmed Al-Baltah, Muhammed Basheer Jasser, and Ali Wagdy Mohamed. "Accelerating crop yield: multisensor data fusion and machine learning for agriculture text classification." *IEEE Access* 11 (2023): 20795-20805. https://doi.org/10.1109/ACCESS.2023.3249205.

[2] Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction. IEEE Access, 9, 63406–63439. https://doi.org/10.1109/ACCESS.2021.3075159.

[3] G L, S., V, N., & U, S. (2022). A Review on Prediction of Crop Yield using Machine Learning Techniques. 2022 IEEE Region 10 Symposium (TENSYMP), 1–5. https://doi.org/10.1109/TENSYMP54529.2022.9864482.

[4] Sharma, P., Dadheech, P., Aneja, N., & Aneja, S. (2023). Predicting Agriculture Yields Basedon Machine Learning Using Regression and Deep Learning. IEEE Access, 11, 111255111264. https://doi.org/10.1109/ACCESS.2023.3321861.

[5] Bharathi, P. S., Amudha, V., Ramkumar, G., Nagalakshmi, T. J., Nalini, N., & Jagadeesh, P. (2022). An Experimental Analysis of Crop Yield Prediction using Modified Deep Learning Strategy. 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 1–6. https://doi.org/10.1109/ACCAI53970.2022.9752492.

[6] Thirumal, S., & Latha, R. (2023). Automated Rice Crop Yield Prediction using Sine CosineAlgorithm with Weighted Regularized Extreme Learning Machine. 2023 7th InternationalConference on Intelligent Computing and Control Systems (ICICCS), 35–40.https://doi.org/10.1109/ICICCS56967.2023.10142403.

[7] Kulyal, M., & Saxena, P. (2022). Machine Learning approaches for Crop Yield Prediction: A Review. 2022 7th International Conference on Computing, Communication and Security (ICCCS), 1–7. https://doi.org/10.1109/ICCCS55188.2022.10079240.

[8] Ashfaq, M., Khan, I., Alzahrani, A., Tariq, M. U., Khan, H., & Ghani, A. (2024). AccurateWheat Yield Prediction Using Machine Learning and Climate-NDVI Data Fusion. IEEEAccess, 12, 40947–40961. https://doi.org/10.1109/ACCESS.2024.337673.

[9] Badshah, A., Alkazemi, B. Y., Din, F., Zamli, K. Z., & Haris, M. (2024). Crop Classificationand Yield Prediction Using Robust Machine Learning Models for AgriculturalSustainability. IEEE Access, 12, 162799–162813.https://doi.org/10.1109/ACCESS.2024.3486653.

[10] Basavaraju, N. M., Mahadevaswamy, U. B., & Mallikarjunaswamy, S. (2024). Designand Implementation of Crop Yield Prediction and Fertilizer Utilization Using IoT andMachine Learning in Smart Agriculture Systems. 2024 Second International Conference onNetworks, Multimedia, and Information Technology (NMITCON), 1–6.https://doi.org/10.1109/NMITCON62075.2024.10699184.

[11] S. G. Akther, A. Islam, E. J. Muna and J. J. Tasmi, "AI Meets Agriculture: Advanced Rice Yield Forecasting Based on Historical Data using Machine Learning Approaches," *2024 27th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, 2024, pp. 2570-2575, https://doi.org/10.1109/ICCIT64611.2024.11021921.

[12] S. S and K. Raja, "Comparison of Machine Learning Algorithms for the prediction of Rice Crop Yield in Karnataka," *2023 Fourth International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, Bengaluru, India, 2023, pp. 1-8. https://doi.org/10.1109/ICSTCEE60504.2023.10585142.

[13] Erlin, A. Yunianta, L. A. Wulandhari, Y. Desnelita, N. Nasution and Junadhi, "Enhancing Rice Production Prediction in Indonesia Using Advanced Machine Learning Models," in *IEEE Access*, vol. 12, pp. 151161-151177, 2024, https://doi.org/10.1109/ACCESS.2024.3478738.

[14] N. Madan and U. Soni, "Machine Learning Techniques for Disease Prediction in Rice Plant," *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, Erode, India, 2024, pp. 525-531, https://doi.org/10.1109/ICSSAS64001.2024.10760293.

[15] R. Verma, "Crop Analysis and Prediction," *2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, Aligarh, India, 2022, pp. 1-5. https://doi.org/10.1109/IMPACT55510.2022.10029218.

[16] T. -V. T. Ha, B. -A. Nguyen and T. -N. Do, "Enhancing Rice Yield Prediction using Vision Transformer and Regression Models," *2024 IEEE International Conference on Computing (ICOCO)*, Kuala Lumpur, Malaysia, 2024, pp. 273-278. https://doi.org/10.1109/ICOCO62848.2024.10928258.

[17] J. Augustin, M. Gopinath, B. Karali and Y. Rao, "Joint Prediction of U.S. Rice Yields and Methane Emissions: A Machine Learning Approach," in *IEEE Access*, vol. 13, pp. 70018-70043, 2025, https://doi.org/10.1109/ACCESS.2025.3562397.

[18] S. Durai, C. Mahesh, T. Sujithra and C. Shyamalakumari, "Germination Prediction System for Rice seed using CNN Pre-trained models," *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, Chennai, India, 2022, pp. 1-9, https://doi.org/10.1109/ACCAI53970.2022.9752611.

[19] S. Thirumal and R. Latha, "Automated H yperparameter Tuned Stacked Autoencoder based Rice Crop Yield Prediction Model," *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2023, pp. 14-18, https://doi.org/10.1109/ICOEI56765.2023.10126037.

[20] A. P, D. J, F. D and D. R, "Crop Yield Prediction Using Machine Learning Algorithms," *2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, Chennai, India, 2025, pp. 1-6, https://doi.org/10.1109/RMKMATE64874.2025.11042701